THE CONSTRUCTION OF UTILITARIAN MORAL BEHAVIOUR: EVIDENCE FROM PERSPECTIVE-TAKING ACCESSIBILITY

Rosanna Karina Martin

A thesis submitted to the University of Huddersfield in partial fulfilment of the requirements for the degree of Doctor of Philosophy

The University of Huddersfield

May 2020

Acknowledgements

I would like to express my gratitude to Professor Petko Kusev, a colleague, mentor and most importantly a true friend whom without his patience and belief in my abilities, this thesis would not be possible. I look forward to collaborating with him on many more exciting research projects to come.

Besides Professor Petko Kusev, I extend my gratitude to my dear friend, Hristina Kuseva for the support and kindness she has shown me both before and since we moved to West Yorkshire.

I further extend my appreciation, love and pure admiration to my wonderful Fiancé, Kathleen Lopez, who has loyally stuck by me and encouraged me throughout the course of this research project.

I would also like to thank Lisette Martin and Paolo Francis for their continued support and encouragement in my every endeavour.

Finally, I would like to acknowledge two strong women and role models, Doreen Stenson and Caterina Francis.

Copyright Statement

Abstract

For centuries, moral philosophers and decision-making theorists have been interested in whether people's preferences conform to utilitarian expectations, where it is acceptable to sacrifice one person in order to save many others (Bentham, 1789/1970). However, I argue that despite many moral dilemmas requiring participants to reason about moral perspectives, they do not contain full perspective-taking (PT) accessibility. For example, in the autonomous vehicle (AV) dilemma (Bonnefon et al., 2016), participants are asked to make judgements in response to scenarios offering partial PT accessibility. In particular, in the scenario, participants are presented with the perspective of a passenger travelling inside an AV that is about to crash into a group of 10 pedestrians but are not offered the perspective of the pedestrians. Accordingly, participants' judgements of moral appropriateness and subsequent purchasing behaviour (whether they would like to buy a utilitarian or non-utilitarian AV) are biased by scenarios offering partial PT accessibility. It is little wonder then why Bonnefon et al.'s (2016) findings reveal that people, despite their utilitarian moral judgements, do not express prosocial utilitarian purchasing behaviour; they do not want to buy the utilitarian AV that they judged to be the most moral. Accordingly, in 9 Experiments, I explored the influence of PT accessibility on participants' moral preferences. In Experiments 1-4, I demonstrate that when offered full PT accessibility to AV crash scenarios, participants' moral utilitarian judgements informed their utilitarian purchasing and usage behaviours (purchasing value, willingness to buy and ride AVs). In other words, people wanted to buy and ride the utilitarian AV that they judged to be moral. Moreover, utilitarian preferences were found to be consistent across type of involvement (stranger, participant and family member), judgement tasks (judgements of moral appropriateness, purchasing value, willingness to buy and willingness to ride AVs; Experiments 3, 4, 8 and 9), preference elicitation methods (judgements vs. choice; Experiments 5 and 6) and psychological processing employed by the participants (immediate, conscious and unconscious; Experiments 7-9). Thus, the present thesis provides theoretical and empirical evidence for PT accessibility and its importance in informing people's moral behaviour. Moreover, this thesis offers commercial and policymaking insights for promoting public acceptance of autonomous systems.

Table of Contents

List of Abbreviations

BPS	British Psychological Society
BSREC	Business School Research Ethics Committee
EUT	Expected Utility Theory
PT	Perspective-taking
SAE	Society of Automotive Engineers
SEU	Subjective Expected Utility Theory
UTT	Unconscious Thought Theory

List of Figures

Figure

List of Tables

Table

CHAPTER 1

Theoretical Background and Motivation for the Thesis

1.1 Overview of Chapter 1

The purpose of Chapter 1 is to provide a theoretical background to the broad and multidisciplinary moral decision-making literature, and to introduce my theoretical and experimental contribution. Accordingly, the theoretical background provides an extensive overview of the contributions that scholars from philosophy, behavioural economics, cognitive psychology and neuropsychology have made to the moral decision-making literature. In Section 1.2, I introduce what is arguably the most influential theory of morality and normative decision-making: Utilitarianism (Bentham, 1789/1970). The rationale behind focusing on utilitarianism is twofold: (i) utilitarian philosophy has had a timeless influence on both modern economics and moral psychology, which are important prerequisites to the present thesis and (ii) the main dependent measure across all 9 Experiments contained within this thesis is, *utilitarian behaviour*. Therefore, a critical exploration of utilitarianism provides an essential context for this thesis, including how utilitarianism contrasts with another ethical theory: deontology (Kant, 1785/2002). I end the section by exploring limitations of the methods used to measure utility in moral philosophy research.

In Section 1.3, alternative methods of measuring utility (introduced by behavioural economists) are discussed. Furthermore, I present normative theories of human decision-making that are driven by utilitarian logic and principles, where utilitarian maximisation is considered a rational normative rule that decision-makers are expected to adhere to. I then discuss the difference between expected normative behaviour (how rational humans are expected to behave; von Neumann & Morgenstern, 1944) and descriptive behaviour (how humans actually behave), as exemplified by Prospect Theory (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992).

In Section 1.4, psychological theories of moral decision-making are introduced. This section includes the opposing rationalist and intuitionist models of moral decision-making, and

their consolidation under the Dual Process Theory of Moral Decision-Making (Greene et al., 2001).

In Section 1.5, I discuss contemporary moral problems that arise from developments in artificial intelligence (autonomous vehicles). In particular, I discuss the potential ethical algorithms that can be programmed into autonomous vehicles in preparation for unavoidable collisions. The two algorithms discussed include prosocial utilitarian algorithms that aim to save the greater number of people, and passenger-protective (non-utilitarian) algorithms that aim to prioritise the lives of the passengers. I discuss an interesting moral hypocrisy exhibited in Bonnefon et al.'s (2016) experimental research - where people do not want to buy the utilitarian car that they judge to be the most moral for societal use. Accordingly, I introduce my novel contribution to the field of moral decision-making: Perspective-Taking Accessibility as well as my predictions of how Perspective-Taking Accessibility can influence people's moral judgements and purchasing behaviour for AVs and inform the moral hypocrisy exhibited in Bonnefon et al. (2016).

The chapter closes at Section 1.6, where I outline the experimental methods that I have applied to test my theory and predictions. Moreover, in this section I briefly overview the 4 Experimental Chapters of this thesis (Chapters 2-5).

1.2 The Moral Philosophy of Utilitarianism

1.2.1 The Greatest Happiness Principle

In the opening sentences of the *Introduction to the Principles of Morals and Legislation*, Jeremy Bentham describes two sensations, pleasure and pain, and points to their central role in guiding human moral behaviour (Bentham, 1789/1970). Although he was not the first philosopher to establish the importance of these two sensations in directing morality (such ideas date back to Epicurus [see, Annas, 1987] and Francis Hutcheson, 1728/1742), Bentham is credited with their connection and was the first to utilise the principle in moral legislation. Much like other enlightenment thinkers, Bentham argued that rather than relying on religion or intuition to determine the moral appropriateness of an action, one should instead employ reason. Bentham's notion of reason followed the pleasure-seeking philosophy of hedonism, where Bentham advised that that a morally permissible action is one that produces the greatest happiness; in that, reducing pain and increasing the pleasure of those affected. Therefore, Bentham (1789/1970) intended pleasure to be maximised and pain to be minimised where possible, and accordingly suggested that pleasure can be measured in terms of *utility*, and pain measured in terms of *disutility*. Bentham coined this moral doctrine *The Greatest Happiness Principle* (hereafter, Utilitarianism) and with this principle, he intended to reform the English judicial system.

Much like other consequentialist theories, utilitarianism follows the logic of utility maximisation: the ends (if, the best outcome) justify the means. In other words, as long as the end goal maximises utility, then any potentially egregious act required in order to deliver the goal can be justified. However, there is an important distinction between utilitarianism and other moral consequentialist theories. For example, egoistic consequentialism is characterised by the maximisation of utility for self-interest, in the absence of the interest of others (Brink, 2009). Alternatively, altruistic consequentialism aims to maximise utility for others, in the

absence of self-interest (Brink, 2009; Broad, 1971). Conversely, the goal of utilitarian consequentialism is to maximise utility for the greatest number of people possible, or according to Bentham's slogan: "it is the greatest happiness of the greatest number that is the measure of right and wrong" (Bentham, 1776/1988, p. 3).

Since Bentham's utilitarianism was originally a proposal for legislative purposes, Bentham believed that those with the authority to make a decision, should not also be one of the individuals affected by decision outcomes. This differs once again from egoistic and altruistic consequentialism, since egoistic decisions are intended to directly benefit the decision-maker, and altruistic decisions can in some cases cause the decision-maker harm (Huebner, & Hauser, 2011). Bentham reasons that if the decision-maker believes they will be affected by the decision outcomes then they are prone to a non-utilitarian bias (e.g., they may behave egoistically or altruistically). Utilitarianism is therefore a unique consequentialist theory in two ways: (i) the aim is to maximise utility for the greatest number of people, and (ii) when employing utilitarianism in its purist form, the decision-maker should not potentially be affected by any decision-making outcomes.

It is fair to say that Bentham's formation of utilitarianism was focused on pleasures and pains in terms of their *quantity*. However, given that pleasures and pains are not simply quantified, Bentham suggested a method (known as the felicific calculus) to measure them based on 7 dimensions. The 7 dimensions included a pleasure or pains intensity, duration, certainty/uncertainty, proximity/remoteness, fecundity[1], purity[2], and extent[3]. Therefore, each dimension should be assigned a subjective *pleasure* and *pain* value and then computed as follows:

[1] The probability that it will be followed by a sensation of the same kind.
[2] The probability that it will not be followed by a sensation of the opposite kind.
[3] The number of people the sensation extends to.

> Sum up all the values of all the *pleasures* on the one side, and those of all the *pains* on the other. The balance, if it be on the side of pleasure will give the *good* tendency of the act upon the whole … if on the side of pain, the *bad* tendency of it upon the whole … take an account of the number of persons whose interests appear to be concerned; and repeat the above process with respect to each. (Bentham, 1789/1970, IV: V-VI).

After employing the felicific calculus method to measure expected pleasure and pain outcomes, the decision-maker should be able to determine the most moral course of action. However, Bentham faced some criticism regarding his narrow quantitative view of pleasure and pain. In particular, Bentham was criticised for creating a 'doctrine worthy of a swine' (a criticism mentioned and later addressed by Mill, (1863/2014), since his notion of pleasure was no different from the pleasure animals could enjoy. Therefore, John Stuart Mill (1863/2014), a follower and student of Bentham extended measurements of pleasures and pains to incorporate not only their *quantity* but also their *quality*.

Although Bentham is considered the founding father of modern utilitarianism, it was Mill (1863/2014) who gave it its most widely used name. Mill (1863/2014) built on utilitarianism with 3 major claims: (i) pleasures can differ from one another based on qualitative distinctions, (ii) some pleasures are considered higher than other pleasures based on these qualitative distinctions and (iii) the qualitative distinction between pleasures concern whether pleasures require human or *limited* animal faculties in order to be experienced (West, 2004). Mill (1863/2014) believed that mental pleasures are superior to bodily pleasures since both can be experienced by humans, whereas only the latter can be experienced by animals. Mill therefore argued that higher pleasures were more worthwhile pursuing, particularly if a person is faced with a choice between experiencing the sensation of a lower or higher pleasure.

1.2.2 Opposition to Utilitarianism: Kantian Deontology

Another way to position utilitarian ethics, is to establish how it contrasts from other moral philosophies. A major contrasting moral philosophy of utilitarianism is deontology. Much like utilitarianism, deontologists such as Immanuel Kant (1785/2002) believe that moral appropriateness can be determined by reason. However, unlike utilitarian thinkers' whose notion of reason is based on whether an act will result in particular pleasures or pains, deontologists believe that behavioural acts themselves determine morality. Therefore, deontologists are concerned with behavioural acts of humans and whether such acts are consistent with a moral rule. Kant (1785/2002), in particular believed that in order to behave in a morally appropriate way, people ought to act according to a categorical imperative. A categorical imperative is an order of duty that can be applied in all circumstances and to all people. Accordingly, Kant (1785/2002) described the categorical imperative as follows: "Act only according to that maxim by which you can at the same time will that it should become a universal law" (p. 36). In other words, one should only engage in behaviours that they believe everyone else should also be allowed to engage in. For example, in order to establish whether it is morally appropriate to lie to someone, one should consider whether lying would be a suitable universally accepted behaviour. An individual considering this will most likely believe that it would not be suitable for everyone to be able to lie in all circumstances, therefore according to the categorical imperative, the individual should not lie (ever). This can be contrasted with a utilitarian who, in some cases would permit lying, if the lie resulted in the promotion of happiness or the prevention of unhappiness. For example, lying to protect a friend from a terrible truth would be permissible from a utilitarian standard, since the lie will prevent the friend from experiencing unhappiness.

With more egregious acts such as murder, a deontologist would, once again, argue that murder is an impermissible act that no one under any circumstance should engage in.

Utilitarian's may also agree but only if refraining from murder will result in the greatest happiness for the greatest number. In some circumstances this is not the case. For example, consider the following scenario (the trolley dilemma) originally assembled by Foot (1967) but adapted by Greene et al. (2001):

> A runaway trolley is headed for five people who will be killed if it proceeds on its present course. The only way to save them is to hit a switch that will turn the trolley onto an alternate set of tracks where it will kill one person instead of five. Ought you to turn the trolley in order to save five people at the expense of one? (p. 2105).

In response to this scenario and question, a deontologist would not permit the turning of the trolley as this act would involve murder. Accordingly, 5 people will die, and 1 will live. Alternatively, a utilitarian would permit the turning of the trolley since this act of murder will result in saving the lives of the greatest number of people. Therefore, 1 person will die and 5 will live. Unlike deontological ethics, utilitarian ethics rely entirely on the end result of a behaviour as opposed to the behavioural act itself. As with many philosophies, that require its followers to show consistency and loyalty to their doctrine, in order to be a true *utilitarian*, this logic must be strictly applied in all circumstance, even in cases such as the trolley dilemma where murder becomes permissible.

1.2.3 Selective Critiques of Benthamite Utility Measurements

Theories presented in any discipline are not accepted freely without its critics. Many of the critiques of utilitarianism (that are of particular concern to this thesis) are related to the way in which Bentham intends utility to be measured (see Edgeworth, 1877; Mill, 1863/2014; Plamenatz, 1966; Read, 2007; Sen, 1980-1981; Troyer, 2003). Criticisms of Benthamite utility measurements generally prompt the following question: how can sensations so subjective as pleasure and pain be observed or indeed empirically measured? Bentham suggested that we

should calculate the expected utility of a given action by assigning values to particular pleasures and pain outcomes. However, some authors have argued against the objectivity of assigning units to particular pleasures and pains, since such a subjective task would make unanimous agreement impossible (Marshall, 1920). This is particularly problematic for utilitarian theory, since a unanimous agreement on the value of particular pleasures and pains is a requirement if the goal is to promote the greatest happiness for the greatest number. Moreover, issues with Benthamite utility do not end with the subjective methods of utility valuation. Another concern with Bentham's utilitarianism is his attempt to combine incommensurable dimensions of pleasure/pain in order to obtain the expected utility of an outcome. For example, Bentham's calculation aims to establish the utility of a given outcome based on the values allocated to his 7 dimensions. However, Plamenatz (1966) pointed out a flaw in the proposed calculation; many of these dimensions are not commensurable and do not offer a maximising option. For instance, consider a situation where you are a nurse removing bandages from a burn patient's skin, and their experience of pain is a function of the intensity and duration of the procedure. You could either remove the bandages quickly where the patient experiences intense pain for a short period of time, or you could remove them slowly where the patient experiences a less intense pain for a long period of time. Imagine this problem translates to the following: choosing between (i) a pain with high intensity (at level 900) but short duration (for 60 seconds), or (ii) a pain with low intensity (at level 60) but long duration (for 900 seconds). Assuming, as Bentham believes, that intensity and duration are of equal significance in the calculation of utility, then there is no way to rationally maximise in response to this situation since both options will each independently result in a pain value of 54,000.[4]

[4] This example was inspired by research conducted by Ariely (1998, p. 43) and by his own experiences as a burn patient. He concluded that "bandages should be taken off slowly and steadily, which will cause a long duration for the treatment, but with a low intensity level…".

A similar problem has been explored by Troyer (2003) who argued that the dimensions' intensity and extent can sometimes be in direct conflict with one another. Troyer (2003) notes a problem with attempting to maximise two dimensions simultaneously when they vary independently. Consider the following example adapted from Troyer (2003, p. xiii):

> As a mayor of a small community, you must choose between introducing one of two policies that could result in the following outcomes for the people:
>
> | Outcome A | Outcome B |
> | --- | --- |
> | 800 people at happiness level 799 | 850 people at happiness level 750 |

Troyer (2003) argues that whilst the greatest happiness (intensity) occurs in Outcome A, the greatest number of people enjoying happiness (extent) occurs in Outcome B, thereby creating a disagreement between two values that characterise utilitarian maximisation. This example creates a particular problem for Bentham's, (1776/1988, p. 3) slogan "the greatest happiness of the greatest number", since the greatest happiness and the greatest number of people experiencing happiness cannot be maximised concurrently in this example.

Bentham's proposed methods for utilitarian maximisation are not logically or mathematically sound but they did serve as an integral starting point for many theories of behavioural economics and moral psychology to build upon.

1.3 Decision-Making: Normative and Descriptive Approaches

1.3.1 Normative Decision-Making

A major issue with Bentham's measurements of utility, was a result of his definition of utility itself. As outlined in the previous section, Bentham's used the term *utility* as a measure of pleasure and pain. However, this resulted in criticisms regarding how such subjective

experiences can be measured. Whilst still employing a utilitarian logic, modern economists on the other hand describe utility as a measure of an objects value. This has since led authors to distinguish between Benthamite *experienced utility* (the subjective feeling of pleasure and pain) and economists and psychologist's *decision utility* (a state obtained as a result of observable choice; see Kahneman et al., 1997).

Bentham's utilitarianism has had a lasting impact on behavioural economics and their approach to theorising about- and predicting human behaviour. Behavioural economists follow the approach of rational choice theory, which assumes that humans are rational agents who, when faced with economic decisions, aim to maximise utility (see Scott, 2000; Sudgen, 1991). Accordingly, behavioural theories that are normative to rational choice theory therefore outline how humans *should* behave if they were rational utility-maximising agents (Sugden, 1991). In order to maximise utility, normative theories assume that human agents make utility calculations based on objective known values (such as money and probability). Many normative theories offer such calculations, yet perhaps one of the most prevailing theories is Expected Utility Theory (EUT; von Neumann & Morgenstern, 1944).

EUT was designed as a prescriptive tool for improving human decision-making by offering a formulisation of Bentham's utilitarian approach of utility maximisation. Moreover, EUT has been particularly applied to economic decisions, where people are expected to make choices that maximise utility outcomes for themselves based on known monetary outcomes and the probability of obtaining the monetary outcomes. Employing EUT in decision-making involves a 3-step process: (i) calculating the expected value of all options, (ii) making a trade-off based on the expected values, and (iii) choosing the option with the highest expected value. Accordingly, calculating the expected value (EV) of an option can be achieved using the following formula:

$$EV = \sum P(X_i) * X_i$$

Where *X* refers to a particular outcome (in this case its outcome *i*), *P* denotes the probability of outcome X_i occurring (Von Neumann & Morgenstern, 1944). Therefore, the expected utility can be obtained by multiplying the value of the outcome (e.g., monetary value) by the probability of obtaining the outcome. This calculation can be repeated for multiple choice options until the decision-maker has obtained expected values for each option and from which, they can select the maximising option.

Whilst EUT provides a method for how to calculate known outcomes and probabilities, it does not account for what to do when these values are uncertain or dependent on personal preference. For example, the utility of attending particular events may be weather dependent (an uncertain variable) or dependent on whether or not you like the event altogether (personal preference). Therefore, in order to address this, Savage (1954) amended EUT to what is now known as Subjective Expected Utility theory (SEU), where individuals can include personal utilities in the calculation of expected utility.

Although introducing a subjective measure of utility sounds similar to Bentham's proposal, they are different, since Bentham assumes that the greatest number of people can benefit from a unanimously agreed value placed on a particular happiness, whereas SEU requires decision-makers provide personal subjective values for their own private utility calculations.

Conforming with the expectations of EUT and SEU relies on fulfilling axioms formulised by Von Neumann and Morgenstern (1944), and Savage (1954). According to these normative theorists, if a decision-maker violates any of the axioms of SEU, then they have

failed to behave according to normative assumptions. Six of the main axioms are described as follows:

1. Completeness: Decision-makers should be able to make comparisons between alternatives. For example, based on any critical attribute in question, they should be able to identify whether A > B, whether B < A or whether they are indifferent between A and B.
2. Dominance: The dominant option should always be preferred over the dominated option.
3. Transitivity: Decision-makers should follow a logical preference order. If A > B and B > C, then logically, A > C.
4. Continuity: If decision-makers possess a logical preference order: A > B, B > C, A > C, then the decision-maker should feel indifferent between the combination of A + C when compared to B.
5. Independence: Any outcome that is independent from the decision-makers choice should accordingly not affect the decision-makers choice.
6. Invariance: Decision-makers preferences should not be influenced or change according to how they are described.

Interestingly, the only axiom that will be indirectly violated if any of the other 5 axioms are violated first is the dominance axiom. This is because violation of any axiom will result in failing to choose the dominant option. This will be highlighted in some of the examples of axiomatic violations in section 1.3.2.

1.3.2 Descriptive Violations of Normative Decision-Making

Normative theories of decision-making assume that human decision-makers are rational, coherent and consistent in their choices (Sugden, 1991). This means that humans are expected to maximise utility by performing utility calculations, abide by axiomatic

assumptions of SEU, and do so consistently, regardless of the decision-making content or context. However, numerous examples from experimental research in psychology have demonstrated violations of EUT's and SEU's axiomatic assumptions, and thus violated normative behavioural expectations (e.g., Allais, 1953; Edwards, 1955; Slovic, 1995; Tversky, 1969).

One of the most well-known examples of an axiomatic violation is demonstrated in the Allais Paradox (Allais, 1953), where Allais' experimental findings reveal a violation of the independence axiom. In Allais (1953) experiment, participants are faced with two lottery situations (see Table 1). In situation 1, the experimenter asked participants to make a choice between 2 lotteries (Choice A or Choice B), where Choice A will result in 100% chance of winning £1,000,000, and Choice B will result in 1% chance of winning nothing, 10% chance of winning £5,000,000, and 89% chance of winning £1,000,000. Allais noted a tendency for participants to choose Choice A, the option that posed no risk and a certain gain of £1,000,000. However, when presented with Situation 2, where Choice C will result in 11% chance of winning £1,000,000, and 89% change of winning nothing, and where Choice D will result in 10% chance of winning £5,000,000 and 90% chance of winning nothing, participants preferred the riskier option, Choice D.

Table 1

The Allais Paradox as a Choice of Lotteries

		Lottery Ticket Numbers (1-100)		
		1 (1%)	2-11 (10%)	12-100 (89%)
Situation 1	Choice A	£1,000,000	£1,000,000	£1,000,000
	Choice B	£0	£5,000,000	£1,000,000
Situation 2	Choice C	£1,000,000	£1,000,000	£0
	Choice D	£0	£5,000,000	£0

Note. Adapted from "Judgement and Decision-Making" by P. Ayton. 2005, in N. Braisby & A. Gellatly (Eds.) *Cognitive psychology*. New York: Oxford University Press.

When these choices are represented in Table 1, it is clear to see that participants' choices were affected by information that is irrelevant to their decision. The change in value of lottery tickets 12-100 in Situation 2 is identical across both choice C and D, therefore, this information should not affect their initial strategy of choosing the option with the highest probability over the option with the highest monetary gain. Participants did however switch their preferences from risk-aversion in Situation 1 (Choice A), to risk-seeking behaviour in Situation 2 (Choice D).

It is also important to note that in addition to violating the independence axiom, participants also switched between choosing the option with the lowest expected value (violation of the dominance axiom) to the option with the highest expected value. The expected value for each choice are as follows:

Situation 1:

Choice A: (1% x £1,000,000) + (10% x £1,000,000) + (89% x £1,000,000) = EV of £1,000,000.

Choice B: (1% x £0) + (10% x £5,000,000) + (89% x £1,000,000) = EV of £1,390,000.

Situation 2:

Choice C: (1% x £1,000,000) + (10% x £1,000,000) + (89% x £0) = EV of £110,000.

Choice D: (1% x £0) + (10% x £5,000,000) + (89% x £0) = EV of £500,000.

In both Situations, the latter Choice provides the highest expected value, however participants chose the option with the lowest expected value in Situation 1 and then chose the option with highest expected value in Situation 2.

Another example of axiomatic violations (violation of the independence axiom) has been demonstrated in Lichtenstein and Slovic's (1971, 1973) preference reversal phenomenon. In one variation of their experimental paradigm, Lichtenstein and Slovic (1973) presented

participants visiting a Las Vegas Casino with the following gambles (as described in Slovic, 1995, p. 365):

P bet: 11/12 chance to win 12 chips.

1/12 chance to lose 24 chips.

$ bet: 2/12 chance to win 79 chips.

10/12 chance to lose 5 chips.

Participants were then asked to (i) choose a gamble to play and (ii) to indicate which gamble contained greater value (e.g., how much they would be willing pay for each gamble). Surprisingly, participants that chose the P bet to play also judged the $ bet to have greater value, demonstrating that people reject the option they perceive to be more valuable, and accept the gamble that they perceive to be less valuable. These experimental findings demonstrate a major violation of the invariance axiom, since different variations of identical information alter their preferences, indicating that humans possess separate functions for choices and for valuation judgements. Accordingly, in this example, the dominance axiom is violated which participants make choices but not when they make valuations.

Interestingly, violations of the invariance axiom in particular have been demonstrated widely, indicating how preferences are highly sensitive to variations in the decision-making context (e.g., Kusev et al., 2009; Tversky & Kahneman, 1981; Kahneman & Tversky, 1983). For example, experiments conducted by Tversky & Kahneman (1981; see also Kahneman & Tversky, 1979; Tversky & Kahneman, 1992) indicate that peoples risk preferences are dependent on whether hypothetical problems are framed in terms of loss or gain (Tversky & Kahneman, 1981). For instance, imagine a scenario in which the US are preparing for an outbreak of a disease that is expected to kill 600 people, yet one of two programs can be employed to prevent death (Tversky & Kahnemen, 1981, p. 453):

Program A	Program B
200 people will be saved.	1/3 probability that 600 people will be saved, and 2/3 probability that no one will be saved.

All options offer equal expected values and therefore choosing an option is a violation of the completeness axiom (not offering an opportunity for maximisation). However, when Tversky and Kahneman (1981) presented this problem to participants, they found that the majority (72%) chose program A, the risk-averse option that will result in 200 lives saved with certainty. However, these programs were framed in terms of gains or *saving people's lives* and not in terms of loss or *letting people die*. Therefore, Tversky and Kahneman (1981) also presented the same programs to participants but this time framed them in terms of loss (p. 453):

Program C	Program D
400 people will die.	1/3 probability that nobody will die and a 2/3 probability that 600 people will die.

In response to programs framed in terms of loss (C and D), 78% of people chose program D, the risk-seeking option that will result in an uncertain outcome. Therefore, despite pair A and B and pair C and D being identical in terms of expected value participants' risk preferences are heavily influenced by whether the programs are framed in terms of loss or gain which once again demonstrates a violation of the invariance axiom.

Presenting an exhaustive list of all axiomatic violations of normative decision-making is beyond the scope of this thesis. However, it is clear from the examples provided that humans

violate the axioms of EUT and SEU which means that these normative theories of decision-making do not account for how humans actually behave under risk and uncertainty.

1.3.3 Prospect Theory

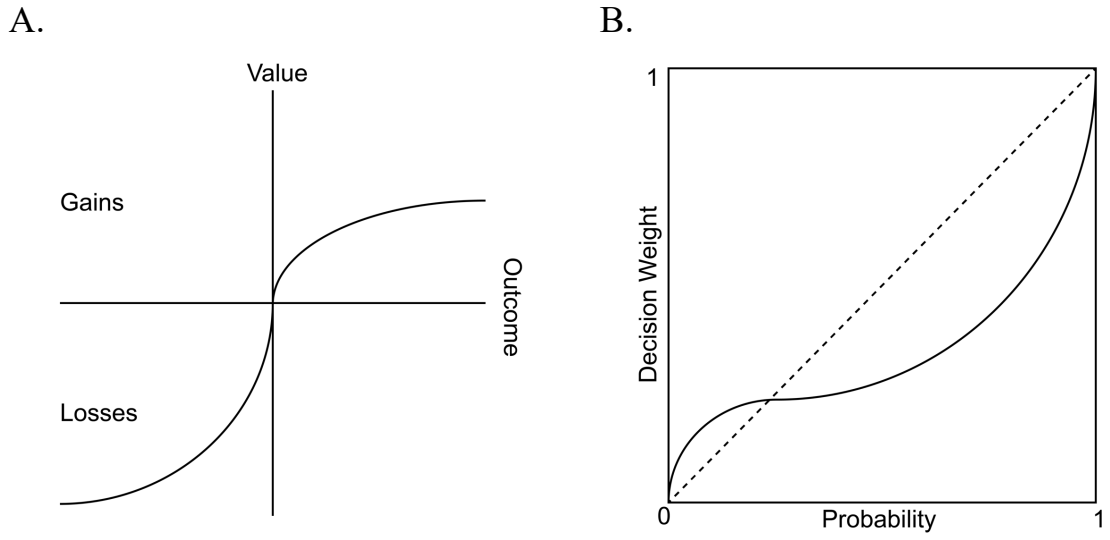
As a result of the many violations of EUT and SEU, psychologists Daniel Kahneman and Amos Tversky introduced a descriptive theory of decision-making under risk known as Prospect Theory (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992). Unlike EUT and SEU, Prospect Theory not only accounts for how humans actually behave but also predicts human behaviour based on variations in probability (high and low) and framing (loss and gain). In order to model human behaviour, prospect theory is based on experimental data from human participants, whose risk preferences are elicited from hypothetical gambles. Accordingly, Prospect Theory demonstrates many interesting phenomena about human behaviour. For example, decision-makers consider expected utility in terms of personal reference points (e.g., their current wealth) as opposed to the expected values (the utilitarian outcome). This has been modelled by the value function (see Figure 1A) where, for example, an increase in wealth from £0 to £5 is perceived as greater than an increase in wealth from £100 to £105, despite these monetary escalations being identical. Moreover, the value function also demonstrates how losses create a greater psychological impact than do equivalent gains; the loss aversion phenomenon (Figure 1A represents this, where a significantly steeper curve can be observed for losses than for gains). For example, losing £5 feels more psychologically impactful than winning £5.

Another important finding from Prospect Theory demonstrates the effect of high and low probabilities on respondents' risky decision-making. In particular, the probability weighting function (Figure 1B) illustrates the tendency for people to overweight small probabilities and underweight moderate-large probabilities of gains and losses. For example, people have the tendency to believe that large probabilities are smaller than they actually are,

and small probabilities are larger than they actually are. This explains why people buy lottery tickets (gains) and insurance policies (losses) overweighting the small chances of winning the lottery or losing their belongings.

Figure 1

Behavioural Predictions of Prospect Theory

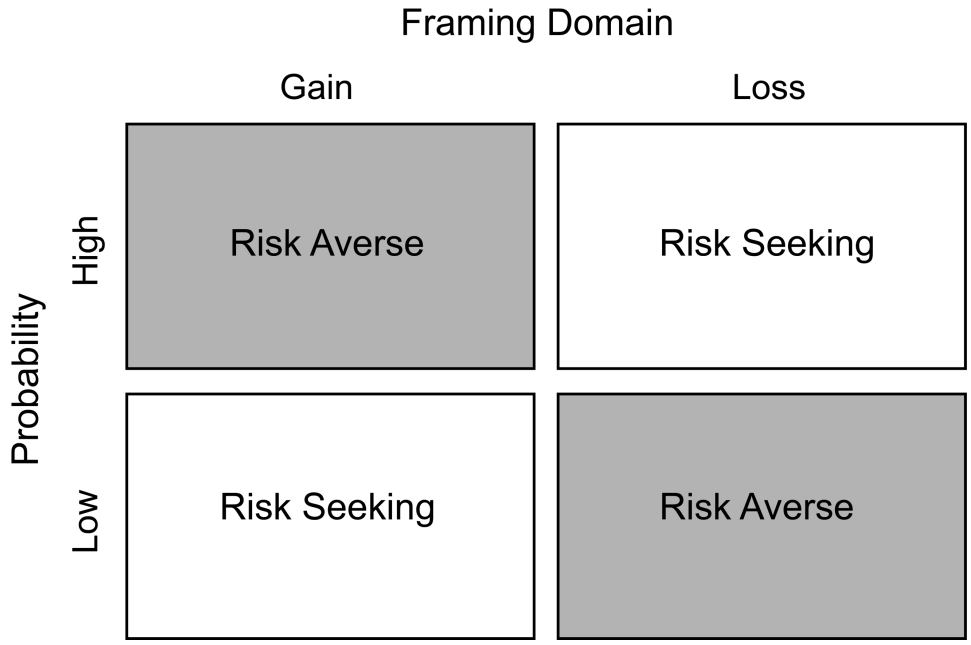


Note. Panel A: The Value Function. Panel B: Adapted from "Prospect Theory: An analysis of Decision under risk" by D. Kahneman and A. Tversky, 1979, *Econometrica, 47*(2), p. 279. Panel B: The Weighting Function. Adapted from "Advances in Prospect Theory: Cumulative representation of uncertainty" by A. Tversky and D. Kahneman, 1992, *Journal of Risk and Uncertainty, 5*(4), p. 313.

As a result of these findings, one of the major contributions that Prospect Theory has made to decision-making psychology and behavioural economics is the fourfold pattern of risk preferences (see Figure 2). The fourfold pattern incorporates the effect that probability (high and low) and domain of decision-making (gain and loss) has on respondents' risk preferences. In particular, behavioural findings indicate that people are risk-seeking for low probability gain and high probability loss, and risk averse for high probability gain and low probability loss (Tversky & Kahneman, 1992; see also Figure 2).

Figure 2

The Fourfold Pattern of Risk Preferences



Note. Adapted from "Preference reversals during risk elicitation" by P. Kusev, P. van Schaik, R. Martin, L. Hall, and P. Johansson, 2020, *Journal of Experimental Psychology: General*, *149*(3), 585-589.

Despite the success of the Nobel Prize winning research, Prospect Theory has faced criticism based on arguments that the behavioural patterns exemplified by experiments are merely an artefact of the risk elicitation method employed (e.g., Hertwig et al., 2004, Hertwig & Erev, 2009; Kusev et al., 2020; Pedroni et al., 2017). For example, Hertwig et al. (2004) argued that the descriptive nature of hypothetical gambles (e.g., their described probabilities) are not representative of the everyday *experience* of probability. For instance, reading descriptive accounts of the low percentage chance of winning the lottery and personally experiencing many unsuccessful lottery draws may elicit different risky behaviours. For example, descriptive methods of risk elicitation tend to lead to an over weighting of small probabilities, whilst experience-based methods of risk elicitation lead to an under weighting of small probabilities (when experience is under sampled). Despite this methodological discrepancy (e.g., Hertwig, et al., 2004; Hertwig & Erev, 2009; Kusev et al., 2009, 2020;

Pedroni et al., 2017), the effect of loss and gain framing on decision-making have remained relatively robust (Kusev et al., 2020; Marteau, 1989; McNeil et al., 1982).

Taken together, it is clear from research in experimental psychology that people do not behave according to normative utilitarian expectations; the typical decision-maker rarely abides by axiomatic assumptions, and when they do, this behaviour is restricted to specific contexts. These findings from behavioural economics and psychology are fundamental in informing research in moral decision-making, since many hypothetical moral problems have also been constructed in order to test normative utilitarian assumptions.

1.4 Moral Decision-Making: The Study of Descriptive Moral Behaviour

1.4.1 Rationalist and Intuitionist Approaches to Moral Decision-Making

Normative theories of decision-making (e.g., von Neumann & Morgenstern, 1944; Savage, 1954) assume that a rational agent uses reasoning to compute and determine utility-maximising options. Accordingly, in moral decision-making psychology utilitarian decisions are often classed as normative decisions that reflect reasoning. For example, when choosing between saving the lives of either 1 or 5 humans, utilitarian decision-makers should opt for saving the 5. However, this also means that decision-makers should also be willing to go as far as to *sacrifice* 1 human in order to save 5 others. Moral psychologists have studied responses to such problems as a method of scientific enquiry into human moral decision-making and the ethical principles individuals base their decisions on. However, some authors have argued against the implementation of hypothetical moral dilemmas in decision-making research, since they may not predict moral behaviour in response to real life moral dilemmas (Bauman et al., 2014; Bostyn et al., 2018; FeldmanHall et al., 2012; Patil, et al., 2014). For example, Bostyn et al. (2018) exposed their participants to both a hypothetical and real version of *the mouse dilemma* where the participants could save 5 mice from being electrocuted by redirecting the current to 1 mouse. Although the effect size for this finding was small, the results revealed that

participants were more utilitarian in response to the real dilemma than they were in the hypothetical version, indicating that hypothetical behaviour is not predictive of actual behaviour.

Despite the contribution Bosytn et al. (2018) makes to the field, their argument against hypothetical dilemmas misses the purpose of moral decision-making research (see, Greene, 2015; Plunkett & Greene, 2019) which intends to study the complex nature of human decision-making processes as opposed to predicting behaviour in real moral situations. Accordingly, research that employs hypothetical moral scenarios has informed many areas of study. For example, empirical research that reveals how moral systems develop in humans are based on children's answers to hypothetical moral problems (Colby et al., 1980; Kohlberg, 1973). Moreover, research employing hypothetical moral dilemmas have advanced our understanding of atypical moral behaviour exhibited by brain lesion patients (Koenigs et al., 2007) and people with personality traits associated with impaired emotional function such as trait alexithymia (Patil & Silani, 2014) and psychopathy (Koenigs et al., 2011). More recently, research into the complex nature of human moral decision-making has also contributed to proposed ethical programming in artificially intelligent machines (e.g., Awad et al., 2018; Bonnefon et al., 2016; Martin et al., 2017). Therefore, hypothetical scenarios have made a prominent contribution to our understanding of psychological processes that underlie moral decision-making.

Within the moral philosophy and moral psychology literature, there are two main approaches to understanding how humans process moral problems: the rationalist approach and the intuitionist approach. Moral rationalists believe that moral acts are the result of controlled moral reason. Alternatively, moral intuitionist believe that moral acts are the result of automatic moral intuitions. According to rationalist approaches to moral decision-making, controlled cognitive processing is a requirement when making moral decisions (e.g., Piaget 1932; Rest, 1986; Rest, Narvaez, Bebeau, & Thoma, 1999; Turiel, 1983). Kohlberg's (1969; 1981; 1973)

theory of moral development emphasises this rationalist approach. Accordingly, Kohlberg's paradigm was based on children's responses to hypothetical moral dilemmas, where children indicated whether a target behaviour is morally right or wrong and justified their reasoning (see for dilemma examples, Colby et al., 1980). Based on his experimental findings, Kohlberg (1973) suggested that children advance their moral reasoning capabilities through 3 developmental levels: (1) the pre-conventional, (2) the conventional and (3) the post-conventional. The pre-conventional level is characterised by punishment avoidance and egoistic self-interest. At this level, an individual can establish right and wrong based external punishments as opposed to an internal experience of guilt. Moreover, the individual behaves according to their own self-interest with little or no concern for others. The conventional level is defined by the ability of the individual to conform to social norms and adopt an internalised rule-driven understanding of moral conduct. Finally, at the post-conventional level, the individual develops their own moral principles and opinions, which may deviate from law. These principles could be based on philosophical viewpoints, engage perspective-taking (see Selman, 1971) and change on a case by case basis. Therefore, Kohlberg's theory of moral development suggests that we approach moral problems using reasoning; an ability that gradually matures over the course of development.

Similar to Kohlberg's theory of moral development, Social Domain Theorists (e.g., Nucci, 1981; Nucci & Turiel, 1978; Smetana, 1999, 2006; Turiel, 1983) postulate that children construct a moral understanding from their everyday social interactions and experiences with moral violations. Moreover, unlike Kohlberg (1973) who believed that children develop moral understanding in hierarchal steps (e.g., they must master one level in order to progress on to the next), Social Domain Theorists such as Turiel (1983) argue that different domains of moral understanding develop simultaneously. Accordingly, empirical findings based on Social Domain Theory, reveal that young children are able to categorise moral violations into three

domains: including the moral domain (issues with fairness and justice), the societal domain (issues with societal conventions) and the psychological domain (issues with individual discretion). Moreover, by employing reasoning, children as young as 3 are able to weight these domains differently. Such abilities are partially attributed to children's experiences with being a victim of a moral violations, which enables them to learn empathy and perspective-taking. Moreover, Smetana (1999) asserts that parents also play a critical role in the development of a child's moral reasoning by providing explanations of morally appropriate and inappropriate behaviour. Rationalist approaches to moral decision-making therefore argue that moral decision-making requires reason and base their claims on evidence of a developmental progression of moral reasoning capabilities in children.

In opposition to rationalist models of moral decision-making, moral intuitionists claim that human assessments of moral permissibility result from fast and emotionally driven intuition (e.g., Haidt, 2001; Hume 1739-1740/1969; Sunstein, 2005). In particular, Haidt (2001) argues that decision-makers initially react to moral problems with intuition and attempt to employ reason only after a judgement has been made (as a post-decisional justification). Haidt (2001) points to the example of a hypothetical scenario that describes an incestual encounter between a brother and sister whilst visiting France. In the scenario, both parties consented, each used a method of contraception and neither were negatively affected by the experience. When Haidt (2001) asked respondents whether, on this occasion, incest was permissible, the majority of respondents quickly protested that it was not. When asked why, respondents attempted to justify their beliefs, but soon realised there was no logical reason why they could not, on this occasional, engage in incest. For instance, it was not illegal (since they were in France), they had prevented an unwanted pregnancy, they had both consented and were not affected by the experience negatively. Eventually, respondents claimed that they cannot explain why but they just *know* the incestual encounter was wrong. This example serves as a

basis for Haidt's (2001) intuitionist model of moral decision-making where the initial reaction to a moral problem is quick and automatic but the post-decisional justification for their reaction requires reasoning. Moreover, as with the incest example, post-decision justifications do not always work, yet due to intuition (and not reasoning) most people still claim that the situation is morally wrong.

Such intuitive responses could be the result of innate evolutionary mechanisms (e.g., Lieberman et al., 2003; Mikhail, 2007), however some theoretical and empirical work suggest roots in reinforcement learning (e.g., Skinner, 1971; Crockett, 2013; Martin et al., under review). For example, Skinner (1971) provides a different account of moral development to that of Kohlberg (1973), contending that individuals learn moral rules through simple reinforcement learning processes, where learned rules become intuitive reactions/aversions to particular outcomes. Therefore, based on reinforcement histories, individuals learn to associate particular behaviours with pleasure or pain related outcomes. According to my recent experimental findings (Not part of this thesis: Martin & Kusev, 2017; Martin et al., under review), successful reinforcers that cause deviations from utilitarian choice can be as simple as verbal cues indicating whether a behaviour is *correct* or *incorrect*. As a result of learning moral rules though a reinforcement learning task, participants' moral choices resembled intuitive rule-governed responses as opposed to utilitarian maximisation strategies that would reflect controlled reasoning.

1.4.2 The Dual Process Theory of Moral Decision-Making

Both rationalist and intuitionist approaches to moral decision-making provide convincing arguments for their respective schools of thought. However, Greene et al. (2001; see also Greene & Haidt, 2002; Greene, 2015) proposed that decision-makers employ both controlled processing *and* emotional intuition when tasked with making moral choices. Greene et al.'s (2001) initial idea was based on a philosophical puzzle regarding how humans respond

inconsistently to different variations of the same hypothetical moral dilemmas (see Foot, 1967; Thomson, 1985). For example, consider once again the trolley dilemma but this time paired with the footbridge dilemma (Greene, 2001):

The Trolley Dilemma:

> A runaway trolley is headed for five people who will be killed if it proceeds on its present course. The only way to save them is to hit a switch that will turn the trolley onto an alternate set of tracks where it will kill one person instead of five. Ought you to turn the trolley in order to save five people at the expense of one? (p. 2105).

The Footbridge Dilemma:

> You are standing next to a large stranger on a footbridge that spans the tracks, in between the oncoming trolley and the five people. In this scenario, the only way to save the five people is to push this stranger off the bridge, onto the tracks below. He will die if you do this, but his body will stop the trolley from reaching the others. Ought you to save the five others by pushing this stranger to his death? (p. 2105).

Greene et al. (2001) noted that consistent with utilitarian moral principles, most people will agree that killing the 1 person in order to save 5 in the trolley dilemma is permissible (resembling controlled processing involving simple utility calculations). However, in the footbridge dilemma the majority of people often refrain from pushing the 1 man onto the tracks in order to save 5 (resembling harm-averse emotional intuitions). This inconsistency creates a puzzle for normative theorists since different representations of the identical (in terms of utility options) choice problems result in different behavioural outcomes. Some authors contend that these dilemmas elicit different moral choices because each dilemma differs in the level of personal involvement. For example, the footbridge dilemma is considered personal because it requires the decision-maker to actively *push* a man to his death, whereas the trolley dilemma

involves a mechanism that distances the decision-maker's actions from the harm outcome (Foot, 1967; Thomson, 1985). Greene et al. (2001) further hypothesised that personal and impersonal dilemmas elicit different psychological processes.

Suitably, Greene et al. (2001) investigated this hypothesis by employing fMRI technology to map activity of emotional and working memory related regions of the brain whilst participants read and answered personal and impersonal moral dilemmas. The authors found that decision-makers were indeed more utilitarian in their responses to impersonal dilemmas compared to their responses to personal dilemmas. Moreover, the neuroimaging findings revealed that areas of the cerebral cortex associated with emotion (Brodmann's Area 9, 10, 31 and 39) showed heightened activity when participants read personal moral dilemmas (e.g., the footbridge dilemma) compared to when they read impersonal moral dilemmas (e.g., the trolley dilemma). Moreover, areas of the cerebral cortex associated with working memory (Brodmann's Area 7, 40 and 46) showed heightened activity when participants read about impersonal moral dilemmas competed to when they read personal moral dilemmas. The authors therefore likened their findings to the dual process theory of decision-making (Stanovich & West, 2000); where one of 2 possible systems can be employed to process information, including system 1 (fast and intuitive, emotionally driven processing), and system 2 (slow and controlled cognitive processing). Accordingly, when considering an impersonal moral dilemma, system 2 processing is employed, which calculates the utility maximising option (utilitarian choice). However, when considering personal moral dilemmas, system 1 processing competes with and dominates system 2 processing which results in emotionally driven responses that avert choices that actively cause harm (a deontological choice).

Interestingly, the time taken to make utilitarian decisions also differed between dilemma types. For example, it took longer for people to make utilitarian decisions in response to personal dilemmas than it did to make utilitarian decisions in response to impersonal

dilemmas. This indicates an interruption in the processing of personal dilemmas which is most likely the result of emotional activations identified in the fMRI data.

Taken together, these findings indicate that utilitarian judgements are the result of slow and controlled processing whereas deontological judgements are the results of fast and intuitive emotional processing. The dual process theory of moral decision-making therefore provides evidence for both rationalist and intuitionist accounts of moral decision-making: moral decisions can be the result of either controlled cognitive or emotional intuitive processing and this is dependent on characteristics of the moral problem (e.g., whether the moral scenario requires personal or impersonal involvement; Greene et al., 2001; Greene & Haidt, 2002).

1.4.3 Contextual Accessibility in Moral Decision-Making Tasks

Characteristic of the cognitive approach of psychology, a large proportion of decision-making research has focused on how the human mind processes information. In particular, some decision-making theorists argue that our choices are the result of the methods used to process the choice options (e.g., Dijksterhuis & Nordgren, 2006; Greene et al., 2001; Stanovich & West, 2000). For example, dual process theorists claim that information can be processed via two competing systems, with the employment of each system resulting in different choices and judgements. Moreover, according to Unconscious Thought Theory (which will be discussed in Chapter 5 of this thesis) our choice preferences also depend on whether we have processed decision-making information at a conscious or unconscious level (Dijksterhuis & Nordgren, 2006). However, whilst the processing of information is important to the formation of judgements and decisions, the way that information is processed in the first place is highly dependent on the construction of the information itself (e.g., Kusev et al., 2009, 2016, 2018, 2020; Tversky & Kahneman, 1981, 1991). For instance, how descriptive information is presented to participants can greatly influence the individual's choice behaviour. As outlined in Chapter 1 of this thesis, Tversky and Kahneman (1981) demonstrated that the framing of

decision problems influences peoples risk preferences. In particular, decision-makers were found to make risk-averse choices when they read scenarios framed in terms of gain, and risk-seeking choices when they read scenarios framed in terms of loss. The framing effect offers an interesting example of how the formulation of information can impact risky decisions. More recently however, Kusev et al. (2016) has explored the influence of enhancing access to information on moral judgements. In their study, Kusev et al. (2016) argued that traditional moral dilemmas based on Thomson's (1985) trolley paradigm offer participants limited accessibility to dilemma details, rendering the scenarios cognitively challenging. Kusev et al. (2016) provide the following example of a typical moral scenario that offers only partial contextual accessibility:

> …The only way to save the lives of the five workmen is to hit a switch near the tracks that will cause the trolley to proceed to the right, where the lone workman's large body will stop the trolley. The lone workman will die if you do this, but the five workmen will be saved. Is it appropriate for you to hit the switch in order to avoid the deaths of the five workmen? Yes/No. (p. 1962).

The authors argued that although the scenario offers an account of what will happen if the decision-maker hits the switch, there is no corresponding account of what will happen if the decision-maker refrains from hitting the switch. Moreover, the moral question directs the decision-makers attention to whether they should turn the trolley and not to whether they should refrain from turning the trolley. Accordingly, the moral scenario and question produces a framing effect whereby the authors have emphasised how to save the five but have not offered the reverse. Kusev and colleagues therefore created moral dilemmas with full contextual accessibility (both moral actions and consequences made explicit):

> …The only way to save the lives of the five workmen is to hit a switch near the tracks that will cause the trolley to proceed to the right, where the lone workman's

> large body will stop the trolley. The lone workman will die if you do this, but the five workmen will be saved. The only way to save the life of the lone workman is not to hit the switch near the tracks. The five workmen will die if you do this, but the lone workman will be saved. Choose the option which is more appropriate for you: Sacrifice one workman in order to save five workmen or Sacrifice five workmen in order to save one workman. (pp. 1962-1963).

In this version of the trolley dilemma, Kusev et al. (2016) have enhanced the readers accessibility to dilemma information, detailing what will happen to all parties if the decision-maker chooses either action. Accordingly, Kusev et al. (2016) tested the influence of accessibility to dilemma information on moral decision-making by presenting both partial and full versions of moral dilemmas to participants and recording their moral judgements. The authors found that when presenting partially accessible dilemmas, participants were less utilitarian in response to personal dilemmas and more utilitarian in response to impersonal dilemmas (replicating Greene et al., 2001). However, when participants were presented with dilemma with full contextual accessibility, they were more utilitarian in their choices, regardless of the type of dilemma involvement (personal and impersonal). Kusev et al. (2016) also found that with full contextual accessibility respondents took less time to make utilitarian choices. Moreover, with partial contextual accessibility, the few participants that were utilitarian in response to personal dilemmas took significantly more time to reach a utilitarian decision than they did in response to impersonal dilemmas (this replicated Greene et al., 2001). However, with full contextual accessibility, the type of involvement (personal and impersonal) did not influence the length of time it took participants to make utilitarian decisions. In summary, Kusev et al. (2016) demonstrated that presenting full contextual accessibility eliminated behavioural differences (in utilitarian choice and response time) exhibited in response to personal and impersonal moral dilemmas. Kusev et al. (2016, p. 1966) concluded

that "…our results suggest that any emotional interference, with rational choices taking more time to make, is an artefact of presenting partial information and does not happen when full information is presented, with rational choices taking less time." Therefore, according to Kusev et al.'s (2016) novel findings, presenting full and unbiased contextual information (full contextual accessibility) about moral dilemmas to participants accordingly eliminates biases in their moral decisions. These finding are of particular importance to the current thesis and will be explored in the context of modern moral dilemmas in Experiment 1.

1.5 Autonomous Vehicles: A Very *Real* Moral Problem

1.5.1 Introduction to Autonomous Vehicle Ethics

So far in the moral decision-making research literature, hypothetical moral dilemmas have been implemented as tools to understand human cognition. However, as argued in an extensive review by Bauman et al. (2014), hypothetical moral dilemmas often detail considerably unlikely events. For instance – the trolley dilemma presents a highly unlikely combination of factors including (i) an out of control train, (ii) 5 workmen in direct danger of being hit by the train with no means of escape, (iii) a rail switch that happens to be available for public use, (iv) you also happen to be standing next to this switch, and finally, (v) another workman stands on a parallel track and will be killed if you chose to divert the train. Each factor alone is unlikely, and when combined, the moral hypothetical scenario seems highly improbable. The scenario itself therefore lacks mundane realism (how likely the scenario would occur in decision-makers daily life). However, whilst the trolley dilemma may lack mundane realism, recent advances in the development of artificially intelligent machines (e.g., autonomous vehicles) have resulted in trolley-like scenarios becoming a very real moral problem. For example, much like the decision the participant is expected to make in response to the trolley dilemma, autonomous vehicles can be pre-programmed to *choose* whose life to save in crash scenarios (and take into account all possible situational factors). Accordingly,

this initial moral decision must be programmed by a human (e.g., the car manufacturer or policy maker). This has created a new application of moral decision-making research, and accordingly new and realistic moral problems that must be addressed before autonomous vehicles become available to the public (Miller, 2016; Sharif et al., 2017). Therefore, in the following section, autonomous vehicles will be discussed along with the inevitable moral dilemmas that policy makers and car manufacturers face when implementing ethical algorithms.

Autonomous Vehicles (AVs) are cars that can take control of some or all aspects of driving including acceleration, deceleration, steering, and monitoring of the driving environment (Litman, 2018). Importantly, this means that AVs can replace human drivers in many or all of the highly demanding tasks required to operate a vehicle. Although, AVs may sound like science fiction, early conceptions and models have existed since the 1920's when the first radio-controlled *driverless* car was tested by the US military. While radio-controlled cars were not technically autonomous, they were constructed and tested with the intention of relieving drivers of driving tasks and promoting driving safety (Kröger, 2015). Since the 1930's the fictional concept of *actual* AVs that drive themselves and learn complex road networks have appeared in numerous sci-fi novels and films (see Kröger, 2015). AVs are therefore not a contemporary idea. However, due to the recent advances in the development of artificial intelligence and motion-sensing technology, AVs are expected to become commercially available as early as 2020 (Fagnant & Kockelman, 2015).

Whilst all AVs occupy autonomous driving features, not all AVs possess the same level of autonomy. The Society of Automotive Engineers (SAE) have defined 6 levels of automobile automation ranging from no automation (non-autonomous; level 0) which applies to cars without system operated driving assistance features (e.g., lane discipline) to full automation (level 5), where the AV occupies full autonomy over acceleration, deceleration, steering and monitoring of the driving environment in the complete absence of interference from human

passengers (See Litman, 2018 for all AV levels). Importantly, for the purpose of this thesis, the definition of AVs will be restricted to fully automated level 5 AVs.

AVs have received widespread multidisciplinary attention from researchers in artificial intelligence, engineering, transport, law, philosophy, psychology and business (e.g., Awad et al., 2018; Bonnefon et al., 2016; Fagnant & Kockelman, 2015; Faulhaber et al., 2019; Goodall, 2014; Martin et al., 2017; Maurer et al., 2015; Nyholm & Smids, 2016). Moreover, many car manufacturers such as Ford, Mercedes and Tesla, as well as technology companies including Google and Uber are currently involved in developing and testing autonomous driving technology (Fleetwood, 2017). The introduction of such vehicles presents many benefits, the most obvious being that AVs will relieve current drivers from highly demanding driving-related tasks. However, AVs can also be utilized by non-divers too since they will be capable of transporting disabled people, the elderly and children (Meyer et al., 2017). Moreover, AVs are also predicted to significantly reduce road traffic. According to Bose and Iannou (2013), only 10% of cars on a single highway segment need to be autonomous for there to be a significant improvement in highway congestion. Furthermore, despite the introduction of AVs potentially increasing the number of people on UK roads, AVs are also predicted to emit less greenhouse gasses into the atmosphere than non-autonomous cars due to the reduction in energy-wasting human driving errors and inefficiencies (Wadud et al., 2016).

Perhaps one of the most important implications of replacing human drivers with automated transport systems is the anticipated reduction in the number of road accidents often caused by human factors such as drink driving, fatigue, and human error (Fagnant & Kockelman, 2015; Goodall, 2014). Accordingly, Fagnant and Kockelman (2015) estimate that AVs will (at the very least) prevent 40% of road accidents. Nevertheless, given the unpredictability of other human drivers, cyclists, and pedestrians behaviour (as well as animals and debris), it is inevitable that AVs will still be involved in collisions. However, AVs can be

pre-programmed scan the environment and calculate the most *moral* course of action within seconds. Yet, what constitutes the most moral course of action must first be defined by humans.

Pre-programming AVs comes with many ethical, legal and safety implications (cf. Maurer et al., 2015). One concern relates to the possible ethical principles that will be embedded into AV algorithms. For instance, in preparation for potential unavoidable collisions, AVs could be programmed to make passenger-protective decisions (protecting the passenger at all costs), or to make decisions compatible with utilitarian ethical principles (protecting the greatest number of people). Of course, as pointed out by Nyholm and Smids (2016), AVs will never actually make a moral decision; instead, AVs will follow through pre-determined decisions that have been configured by humans. This therefore leads to the issue of *who* gets to decide how AVs should be programmed and how this may affect the AVs ethical behaviour. For example, the design of ethical algorithms might be informed by policymakers who tend to impose limits on individual freedom in order to benefit the overall community, which suggests that they might opt for utilitarian AVs (see Fleetwood, 2017). For example, in everyday driving it is illegal to drive over the designated speed limit. Whilst this restriction may limit the individuals' driving freedom, having speed regulations in place benefits the wider community as a whole. However, policymakers may not be tasked with regulating mandatory ethical standards for AVs. Alternatively, the car manufacturers themselves may have the power to choose how to program their vehicles and might opt for passenger-protective cars since they may be easier to market to consumers (Bonnefon et al., 2016). For example, in a 2016 interview, Mercedes Benz executive Christoph von Hugo assured his customers that in the event of a collision, future Mercedes AVs will prioritise the lives of its passengers (Miller, 2016).

Some authors have entertained the idea that the car buyers themselves should have a say in the ethical behaviour of their own cars (Contissa et al., 2017). Accordingly, AVs could

possess a personal ethics setting (PES), where AV owners can adjust the ethical setting in their car, selecting between protecting themselves or protecting the greatest number of people. However, the proposal of a PES has also been received with criticism, since according to game theory predictions, people will put their own personal safety over the social welfare for the community, ironically resulting in increased probability that the driver will die in an accident (see Gogoll & Müller, 2017).

According to utilitarian theory (e.g., Bentham, 1970/1789), the most moral course of action would be to programme AVs to minimise overall harm. Normative theorists would also argue that utilitarian AVs are the most rational cars since they maximise utility (von Neumann & Morgenstern, 1944; Savage, 1954). Likewise, in moral psychology, utilitarian choices are often considered the gold standard of moral decision-making and are associated with the best possible decision outcomes (Greene at al., 2001; Greene, 2015; Kusev et al., 2016). Utilitarian AVs have accordingly been perceived by many authors as the most prosocial vehicle and therefore the most morally appropriate vehicle for public use (Bonnefon et al., 2016; Gogoll & Müller, 2017; Sharif et al., 2017). For instance, existing driving laws are utilitarian in their very nature since they limit individuals' freedom in order to promote the greatest overall safety for the driver and other people. Moreover, one of the major goals of replacing human drivers with AVs is that they are expected to reduce the number of deaths and injury's caused by human driving errors (thus minimising harm). Therefore, utilitarian AVs fit in with current UK driving regulations and with the general harm minimising goals of autonomous driving technology. However, whilst introducing utilitarian AVs may be the most prosocial and beneficial to societal wellbeing, this does not necessarily mean they will be received well by the public.

1.5.2 Autonomous Vehicles and Moral Hypocrisy

The success of any business is highly dependent on consumers perception of its product. Accordingly, Gogoll and Müller (2017) argue that if the ethical standards of AVs do not match the moral preferences of the potential consumers, then marketing AVs will be a challenging feat. Given that it is highly unlikely that AVs will be embedded with a PES, it is important that the consumers moral preferences towards AV ethics are taken into account when AV ethical algorithms are developed. However, empirical research in psychology has revealed that peoples' moral preferences towards AVs are not straightforward. For instance, utilitarian moral preferences regarding the ethical programming of AVs have been found to vary as a function of gender, culture, and religious beliefs (Awad et al., 2018). Moreover, utilitarian moral preferences can be swayed by contextual factors such as how many people are involved in a collision and who is at fault (Awad et al., 2018; Faulhaber et al., 2019). However, perhaps one of the most intriguing empirical findings (see Bonnefon et al., 2016) reveals utilitarian preference inconsistencies within the same people, where people do not want to own the AV that they perceive to be the most morally appropriate. Accordingly, in study 3 of Bonnefon et al. (2016), the experimenters presented participants with a variation a scenario where they had to imagine themselves inside an AV that is about to crash into a group of 10 pedestrians in the road. The participants were told that the AV could be programmed to swerve off to the side of the road where it will impact a barrier and kill them (sparing the 10 pedestrians) or it could be programmed to stay on its current path and kill the 10 pedestrians (sparing themselves). Participants were then asked to rate which AV they perceived to be most moral, as well as to indicate their willingness to buy each AV. Bonnefon et al. (2016) found that participants judged utilitarian AVs as the most morally appropriate for societal use, yet they wanted to buy passenger-protective models themselves. Therefore, moral preferences towards AVs depend

upon the decision-maker's role: as citizens, people want prosocial utilitarian AVs but as consumers, people opt for passenger-protective models (Shariff et al., 2017).

This *moral hypocrisy*, where people want to appear moral without actually experiencing the cost of being so, has been demonstrated widely in social psychology research (Batson et al., 1997a, 1999, 2003; Batson & Thompson, 2001). For example, Batson et al. (1997a) demonstrates how participants will allocate themselves to a favourable task and a stranger to an unfavourable task, despite judging this allocation choice to be immoral. However, Bonnefon et al.'s (2016) findings apply this moral hypocrisy to a novel context, which renders the future of the AV market uncertain. Accordingly, as a result of their novel findings, Bonnefon et al. (2016) made the controversial claim that car manufacturers should give up on introducing *utilitarian* AVs (see also Greene, 2016). The authors reason that a lack of trust in utilitarian AVs might delay the adoption of AVs altogether. However, when referring to AV acceptance, Shariff et al. (2017) noted that "researchers need to investigate what information best fosters predictability, trust and comfort in this new and specific setting" (p. 3). One of Shariff et al.'s (2017) recommendations was to convey the message to consumers that utilitarian AVs will produce an absolute reduction in risk of harm to both passengers and pedestrians. Moreover, this absolute reduction in risk associated with AV technology should be emphasised over the small risk of harm associated with owning a utilitarian AV. In a commentary (Martin et al., 2017) I have argued that the information required to enable consumers to appreciate the benefit of utilitarian AVs is to access all perspectives in crash scenarios. For example, in all variations of Bonnefon et al.'s (2016) experiments, the participant was expected to take the perspective of the AV passenger and was not offered the alternative; the perspective of the pedestrians. I accordingly argued that limiting this perspective-taking information to only the perspective of the AV passenger may emphasise the risk of being in a utilitarian AV and neglects the risk that passenger-protective AVs pose on society (see Martin et al., 2017).

1.5.3 Moral Perspective-Taking Accessibility

Perspective-taking (PT) is the ability to mentally represent how another person is feeling by (i) imaging how another person feels in their situation or (ii) imaging how *you* would feel in another person's situation (Batson et al., 1997b, 2002, 2003; Ruby & Decety, 2004; Stotland, 1969). The ability to perspective-take has been linked to moral development (Walker, 1980; Kohlberg, 1973) and particularly prosocial behaviours such as social cooperation (Barnett & Thompson, 1985; Johnson, 1975). Moreover, experimentally induced PT has resulted in the reduction in the formation and expression of stereotypes (Galinsky & Moskowitz, 2000) and implicit racial bias (Groom et al., 2008), as well as an increase of helping behaviour toward outgroup members (Shih et al., 2009). Accordingly, a wide variety of methods have been employed to induce PT in experimental settings. For example, in some studies, participants are required to write a short story from the perspective of another person (e.g., Galinsky & Moskowitz, 2000). In other studies, participants are required to take on the perspective of another person by watching videos or reading vignettes from the perspective of a fictional character (e.g., Lamm et al., 2007; Negd et al., 2011). Modern approaches have involved the implementation of virtual reality technology, where participants can experience the perspective of another person by taking on the persona of an avatar (Groom, et al., 2008). Whilst these methods vary in their approach of inducing PT, they all achieve the same goal; they enable the participants to take the perspective of *one* other person. This is what I refer to as partial PT accessibility, where PT information limits the participant to taking the perspective of one agent in a scenario and disregards the perspective of other possible agents. This partial PT accessibility was also induced in Bonnefon et al. (2016), where in all variations of the experiments, participants were required to take the perspective of the passenger inside the AV, neglecting altogether the perspective of the pedestrians. However, in this thesis, I introduce a new definition and method of PT, full PT accessibility, where participants have access to the

perspective of all agents in a particular scenario. The purpose of taking multiple perspectives in moral dilemmas is to allow the decision-maker the opportunity to make informed and unbiased decisions. This is particularly important in research that investigates utilitarian choice since as Jeremy Bentham (1970/1789) warned, those with the power to make a moral decision should not also be directly affected by decision outcomes as this will lead to non-utilitarian behaviour. In particular, if people are aware that they will be affected by decision outcomes (such as in the AV dilemma), they may behave egoistically and make purchasing choices that seemingly benefit themselves. However, with PT accessibility, this egocentric behaviour could be directed at choosing a prosocial AV that will bring about not only the greatest safety for others but the greatest safety for the AV buyer.

I further argue that people PT is not an intuitive strategy; people do not always engage in PT when they do not have full accessibility to PT information. For instance, when presented with moral dilemmas that limit PT to only one agent (the passenger), people fail to take the perspective of other agents (the pedestrians). This leads to people having a limited understanding of AV safety, and overlooks the fact that all people (including AV buyers) will inevitably be pedestrians. In the current thesis, I therefore introduce full PT accessibility as method for enabling potential consumers to gain access to the benefits of utilitarian AVs. Accordingly, the following section outlines the experimental chapters that explore PT accessibility.

1.5.4 Summary of Chapter 1 and Outline of Chapters 2-6

In this chapter, I have discussed the broad multidisciplinary literature that informs the field of moral decision-making. In particular, I have explored the important and relevant existing theories from moral philosophy, behavioural economics, moral decision-making psychology and AV ethics. In Chapter 1, I have also introduced my contribution to these fields: *PT accessibility* and its proposed influence on moral decision-making behaviour in the context

of AV crashes. Suitably, in 9 Experiments reported in Chapters 2-5, I explore how variations in PT-accessibility (full and partial PT accessibility) influence utilitarian moral behaviour using a variety of preference elicitation methods (e.g., moral judgements, moral choices, AV purchasing behaviours and AV usage behaviours). Importantly, all experiments reported in this thesis involve presenting participants with variations of the AV dilemma, adapted and further developed from Bonnefon et al. (2016), from which participants indicate their moral preferences towards utilitarian and non-utilitarian AVs. Moreover, each of the 9 Experiments in this thesis consisted of an independent sample of participants (none of the participants took part in more than one of the 9 Experiments).

Chapter 2 of this thesis includes Experiment 1, which investigates the influence of variations of contextual accessibility (Kusev et al., 2016) in AV crash scenarios on participants moral judgements and purchasing behaviour (willingness to buy) in AV crash scenarios. Moreover, type of involvement was also manipulated in this experiment (participants could imagine themselves as a character in the scenario or a stranger). Accordingly, I explore the influence of full and partial contextual accessibility and type of involvement (stranger and participant involvement) on participants utilitarian judgements of moral appropriateness and purchasing behaviour.

Chapter 3 comprises of Experiments 2, 3 and 4, which for the first time investigate the influence of full and partial PT accessibility on moral judgements and purchasing behaviours. In these experiments, I accordingly present participants with either full or partial (as in Bonnefon et al., 2016) PT accessibility in order to establish whether such variations in PT accessibility has an influence on participants judgements of moral appropriateness, purchasing behaviour and usage behaviour for each AV. In these experiments, I explore different types of involvement in the scenario (e.g., stranger involvement, participant involvement and participant and family member involvement), and particularly how these types of involvement

can influence moral judgements, purchasing behaviours and usage behaviours for each AV. Different types of judgements tasks are used across the three experiments, which require participants to make judgements about utilitarian and non-utilitarian AVs. These judgement tasks include judgements of moral appropriateness, purchasing behaviours (purchasing value and willingness to buy, and usage behaviour (willingness to ride). Finally, I also investigate whether participants judgements of moral appropriateness inform their purchasing and usage behaviours under different levels of PT accessibility (full and partial).

Since all dependent measures in the three experiments of Chapter 3 involve judgement tasks, in Chapter 4 I explore the influence of full and partial PT accessibility on people's moral choices. I accordingly explore the association between people's moral judgements and their moral choices (Experiment 5) and further investigate whether people's moral choices can inform their moral judgements (Experiment 6). These experiments utilise and further develop the experimental methods and procedures employed in Lichtenstein and Slovic's (1971) preference reversal and Brehm's (1956) free-choice paradigms.

In the final experimental chapter (Chapter 5), three experiments investigate how PT accessibility and type of involvement interacts with different types of psychological processing (immediate, conscious and unconscious). These experiments were based on and follow the methodology employed by unconscious thought theorists (Dijksterhuis & Nordgren, 2006). I accordingly challenge the claim that processing moral dilemmas unconsciously leads to utilitarian moral behaviour (Ham & van den Bos, 2010).

In the final chapter of this thesis, I discuss how my research contributes to moral philosophy (such as Bentham's utilitarianism), informs moral decision-making theories (moral dilemmas and perspective-taking) and more broadly the field of psychology of judgement and decision-making (e.g., normative and descriptive decision-making theories; cognitive accessibility theories). Moreover, I also highlighted the practical opportunities that may arise

as a result of my dissertation project. Finally, I offer some limitations of the experimental methods employed in this thesis and make proposals for follow-up studies that address these limitations.

CHAPTER 2

Contextual Accessibility: Moral Judgements and Purchasing Behaviour

2.1 Overview of Chapter 2

Chapter 2 comprises of a single experiment (Experiment 1) which investigates the influence of contextual accessibility on moral judgements and purchasing behaviour. The purpose of this experiment was to employ and apply Kusev et al.'s (2016) contextual accessibility method (successfully tested and applied to abstract moral tasks) to practical and realistic AV crash scenarios. In particular, I investigated whether presenting moral scenarios with full and partial contextual accessibility influences participants' judgements of moral appropriateness when considering AV dilemmas. Moreover, in Experiment 1, I explored for the first time the influence of full and partial contextual accessibility on moral purchasing behaviour (willingness to buy utilitarian AVs). As the realistic nature of the AV crash scenario offers an opportunity for personal involvement (e.g., participants could be also passengers in AVs), I also explored the influence of personal involvement (in the AV crash scenarios) on participants' judgements of moral appropriateness and purchasing behaviour. The results revealed that providing scenarios with full contextual accessibility enhanced participants' utilitarian judgements of moral appropriateness. However, contextual accessibility (full and partial) had no effect on participants' purchasing behaviour (willingness to buy utilitarian AVs). Furthermore, the results demonstrated that respondents were significantly more willing to buy a utilitarian AV when the AV crash scenario involved a stranger as opposed to the participant themselves. Moreover, this was only the case with moral purchasing behaviour, which was also not informed by contextual accessibility.

These results revealed two important findings: (i) contextual accessibility cannot account for the complexity of moral purchasing behaviour and thus is not a general theory of moral decision-making and (ii) whilst the purchasing behaviour is not informed by contextual accessibility, other factors (e.g., type of involvement) influence and determine utilitarian behaviour. These novel findings call for further research to investigate what constitutes full

accessibility. Accordingly, in this thesis I propose, develop and apply the full PT accessibility theory and method in moral decision-making tasks.

2.2 Experiment 1: The Influence of Contextual Accessibility on Moral Judgements and Purchasing Behaviour

2.2.1 Introduction

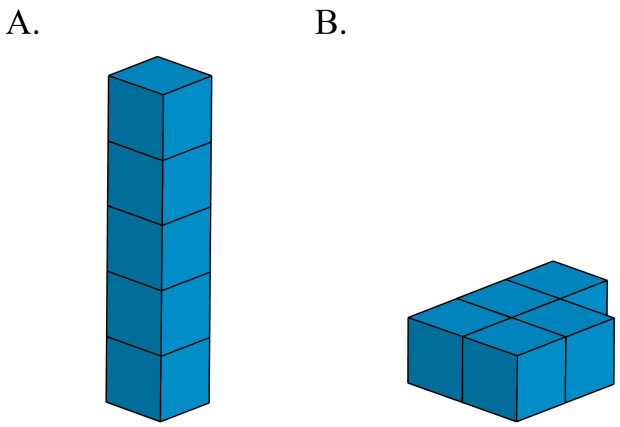
Accessibility has been studied in the context of memory and described as "the ease (or effort) with which particular mental contents come to mind" (Kahneman, 2003, p. 699; see also Kusev et al., 2009; Tulving & Pearlstone, 1966; Tversky & Kahneman 1973). Accessibility is often paired in contrast to *availability,* which refers to whether particular mental contents exist in the mind altogether (Higgins, 1996). Whilst it is impossible to access mental contents that are not available in the mind, it is possible for mental contents that exist in the mind to be inaccessible (Feigenbaum, 1961; Higgins, 1996). This was explored empirically by Tulving and Pearlstone (1966) who first demonstrated the distinction between accessibility and availability. In their experiment, participants were required to memorise a list of words which were associated with corresponding word categories. The participants were then asked to complete two recall tasks. In the first recall task, participants had to recall as many words as they could from memory alone (free recall). In a second recall task, participants were required to recall the words again, however this time they were also presented with the word categories (cued recall). Tulving and Pearlstone (1966) found that participants performed significantly better (recalled more words) in the cued recall task compared to the free recall task. These findings led the authors to infer that whilst information was available in the minds of the participants during the free recall task, not all of it was readily accessible until the subsequent cued recall task. Accordingly, the participants required category cues in order to access the available information. Therefore, Tulving and Pearlstone's (1966) demonstrated empirically

for the first time that information can be available in the mind but it is not always readily accessible; highlighting the importance of accessibility in human behaviour and performance.

Whilst Tulving and Pearlstone's (1966) notion of accessibility refers to accessing internal contents, other scholars such as Kahneman (2003) have noted that elements of the environment can also be accessible or inaccessible. For instance, consider the configuration of the blocks in Figure 3A and Figure 3B. According to Kahneman (2003), we can make assessments relating to these blocks but the ease in which we make these assessments in highly dependent on the accessibility of the information provided. For example, visualising Figure 3A as a tower (or obtaining an impression of the height that 5 blocks can achieve) is easier than it is for Figure 3B. In both Figures, the information is available (they both contain the 5 blocks needed to build a tower), however, Figure 3A provides a fully accessible view of a tower because it is already constructed, whereas Figure 3B must be effortfully imagined by the perceiver. Likewise, forming an impression of the total area that the blocks can cover on a surface is more accessible in Figure 3B than it is in Figure 3A, yet in both figures, this information is available. Accordingly, Kahneman's (2003) example highlights that much like memory instances, when external information is available it may not always be fully accessible.

Figure 3

The Selective Accessibility of Natural Assessments



Note. Adapted from "A perspective on judgement and choice: Mapping bounded rationality" by D. Kahneman, 2003, *American Psychologist, 58*(9), p.700.

Similar to the example presented in Figure 3, written information (e.g., hypothetical scenarios) can also contain *partial accessibility*, where some features are readily accessible whilst others are not. As explored in Chapter 1, Kusev et al. (2016) argued that moral dilemmas commonly employed in moral decision-making research only present partial accessibility to moral scenarios. For example, in the trolley dilemma (e.g., Greene et al., 2001) the action and consequence of hitting the switch are accessible (a detailed description of the consequence is provided). However, the action and consequence of refraining from hitting the switch can only be inferred through reasoning and mental simulation. Moreover, the moral questions related to the dilemma were also partially accessible; they did not explicitly state all decision consequences of hitting the switch or refraining from hitting the switch. Therefore, whilst contextual information is available in the scenario, not all of the information is readily accessible to the decision-maker. Kusev et al. (2016) have therefore argued that making one action and its consequence accessible and another action and its consequence inaccessible in a moral decision-making scenario, may bias respondents' moral choices. Kusev and colleagues accordingly introduced a new variation of the trolley and footbridge dilemmas, where all contextual information is made fully accessible both in the scenario and in the moral questions. The authors demonstrated that when participants were granted full accessibility to dilemma information (full contextual accessibility), the difference in utilitarian choice between personal and impersonal dilemmas was eliminated (a difference that has been described and empirically demonstrated in many studies; e.g., Foot, 1967; Greene et al., 2001, 2009; Kahane, 2013; Nakamura, 2013; Thomson, 1985; Waldmann & Dieterich, 2007). Moreover, there was also an overall increase in utilitarian choices in response to all variations of moral dilemmas that were presented with full contextual accessibility. These novel experimental findings make an important contribution to the moral decision-making literature. In particular, Kusev et al.'s

(2016) findings demonstrate how accessibility to information impacts processing of the information and subsequent moral decisions. Therefore, presenting accessible information to participants means that they equally process all vital aspects of the dilemma, leading to unbiased moral judgements.

Based on Kusev et al.'s (2016) findings, it is plausible that when full contextual accessibility is further applied in an AV dilemma context, there should be a significant increase in utilitarian judgements of moral appropriateness and purchasing behaviours (willingness to buy utilitarian AVs). In the context of AV dilemmas, this means that people will judge utilitarian AVs as relatively more morally appropriate than passenger-protective AVs. Moreover, participants might be more willing to buy utilitarian AVs than passenger-protective models. If this is the case, then Experiment 1 will provide additional evidence for the importance of constructing moral dilemmas with full contextual accessibility in order to elicit unbiased moral preferences. Accordingly, for the first time, Experiment 1 explores the influence of full and partial contextual accessibility on moral purchasing behaviour (willingness to buy utilitarian AVs).

Another important issue addressed in Experiment 1 is the participants involvement in AV crash scenarios. Specifically, the AV dilemma (originally constructed by Bonnefon et al., 2016) presents a different behavioural task to that of traditional trolley and footbridge dilemmas (see Nyholm & Smids, 2016 for additional examples not covered in this thesis). For instance, in trolley-style dilemmas, the participant is often described as an active bystander; someone who can manipulate the situation but will not be affected by the decision outcome. In contrast, the AV dilemma presents a situation where participants are not only the decision-makers but are also directly affected by the outcome of their own decisions (e.g., as passengers in AVs). This makes the moral problem more complex since people are tasked with choosing between their own life vs the lives of the greater number. It is therefore also plausible that type

of involvement (stranger or participant in the AV crash scenarios) influences behaviour in response to AV crash scenarios when/if full contextual accessibility is not informing the decisions. Accordingly, Experiment 1 will explore this possibility and more generally whether type of involvement (stranger or participant in the AV crash scenarios) influence participants' judgements of moral appropriateness and purchasing behaviour.

2.2.2 Method

2.2.2.1 Participants

189 participants were recruited via an online survey panel (PureProfile) and consisted of 103 females and 86 males. The mean age of the participants was 52 (*SD* = 14.79). Prior to data collection, ethical approval was obtained from the Business School Research Ethics Committee (BSREC; The University of Huddersfield). Moreover, all participants were treated in accordance with the British Psychological Society's (BPS) ethical standards.

For statistical testing, a significance level of .05 was set. An effect size was not assumed, but a retrospective power analysis was important to determine whether the sample size would allow the detection of a large effect size (*f* = .40 by convention; Cohen, 1988) of the independent-measures effects of type of accessibility and type of involvement and their interaction. The experiment ran for 14 days to ensure data collection from a (i) sufficiently large sample and (ii) a large effect size, will achieve a statistical power of at least .95. According to the retrospective power analysis, the sample size (*N* = 189) produced a power of .98 which was sufficient to achieve the target.

2.2.2.2 Experimental Design

A 2x2 independent measures design was employed where type of contextual accessibility (full and partial) and type of involvement (participant and stranger) were the independent variables. Accordingly, type of contextual accessibility referred to whether

participants received an AV dilemma that contained a full descriptive account of the scenario and question or a partial descriptive account of the scenario and question (partial contextual accessibility).

The first dependent variable was judgements of moral appropriateness, where participants were required to judge on a 10-point rating scale (0-9) how morally appropriate they believed each AV was (higher numerical ratings denoting a more approving judgement of moral appropriateness). The second dependent variable, willingness to buy, was also measured on a 10-point scale (higher numerical ratings indicating a greater willingness to buy the particular AV).

It is important to note that Bonnefon and colleague's (2016) method of placing swerve and stay AVs on a single unipolar scale did not reveal whether each respondent is overall utilitarian in judging both swerve and stay judgements. Moreover, they were also unable to establish the magnitude of utilitarian/non-utilitarian judgements. In order to avoid this pitfall, separate moral appropriateness scales for swerve and stay judgements were employed. Therefore, the judgements of moral appropriateness were computed as a utilitarian weight – the difference between participants' judgements of moral appropriateness for utilitarian swerve and non-utilitarian stay AVs. Specifically, the judgements for non-utilitarian stay AVs were subtracted from judgements of utilitarian swerve AVs in order to generate a utilitarian weight. Therefore, positive and high difference (utilitarian weight) indicated utilitarian judgements. Accordingly, the same logic was applied to the willingness to buy scales; participants indicated their preferences on two willingness to buy scales (one scale for swerve AVs and one scale for stay AVs).

2.2.2.3 Materials and Procedure

Participants took part in an online computer-based experiment where they were randomly allocated to one of 4 experimental conditions based on the 2 independent variables, type of contextual accessibility (full and partial) and type of involvement (participant and stranger). For example, participants allocated to the condition: partial contextual accessibility with participant involvement were presented with the following scenario and questions:

> *You are the sole passenger in an autonomous self-driving vehicle travelling at the speed limit down a main road. Suddenly, 10 pedestrians appear ahead, in the direct path of the car. The car could be programmed to either **SWERVE** off to the side of the road, where it will impact a barrier, killing you or **STAY** on its current path where it will kill the 10 pedestrians (see the picture illustrating this scenario).*
>
> *Judge the moral appropriateness of programming autonomous self-driving cars to: **SWERVE***
>
> *Judge the moral appropriateness of programming autonomous self-driving cars to: **STAY***
>
> *How would you rate your willingness to BUY an autonomous self-driving car programmed to: **SWERVE?***
>
> *How would you rate your willingness to BUY an autonomous self-driving car programmed to: **STAY?***

Alternatively, participants who were presented with full contextual accessibility and participant involvement were presented with the following scenario and questions:

> *You are the sole passenger in an autonomous self-driving vehicle travelling at the speed limit down a main road. Suddenly, 10 pedestrians appear ahead, in the direct path of the car. The car could be programmed to either **SWERVE** off to the side of the road, where it will impact a barrier, killing you but leaving the 10 pedestrians unharmed or **STAY** on its current path where it will kill the 10 pedestrians, but you will be unharmed (see the picture illustrating this scenario).*
>
> *Judge the moral appropriateness of programming autonomous self-driving cars to: **SWERVE**, killing you but leaving the 10 pedestrians unharmed. If the car does not swerve it will kill the 10 pedestrians but leave you unharmed.*

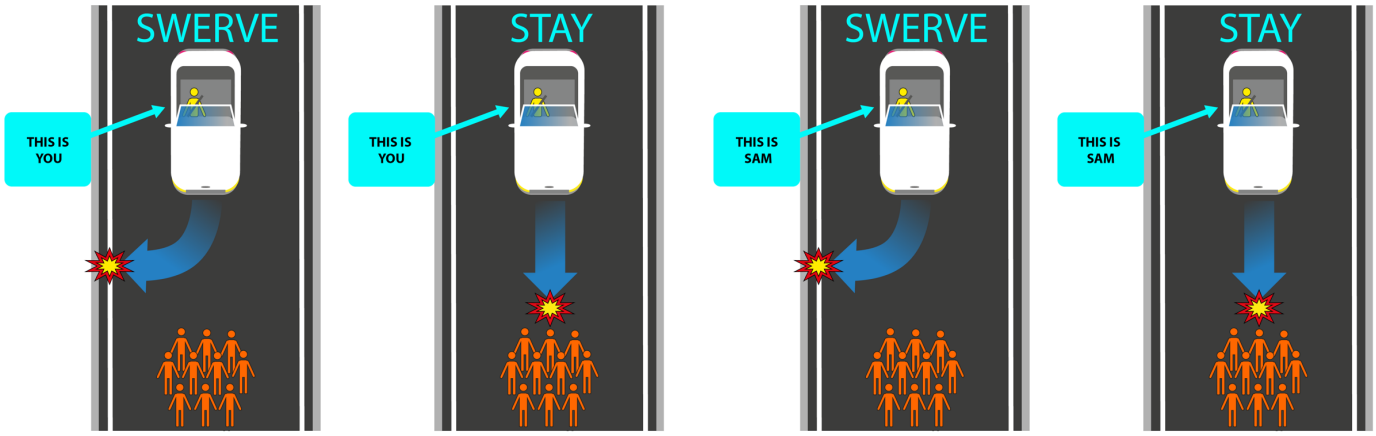
> *Judge the moral appropriateness of programming autonomous self-driving cars to:* ***STAY,*** *killing the 10 pedestrians but leaving you unharmed. If the car does not stay, it will kill you but leave the 10 pedestrians unharmed.*
>
> *How would you rate your willingness to BUY an autonomous self-driving car programmed to:* ***SWERVE?*** *killing you but leaving the 10 pedestrians unharmed. If the car does not swerve it will kill the 10 pedestrians but leave you unharmed.*
>
> *How would you rate your willingness to BUY an autonomous self-driving car programmed to:* ***STAY?*** *killing the 10 pedestrians but leaving you unharmed. If the car does not stay, it will kill you but leave the 10 pedestrians unharmed.*

Accordingly, participants who were in the *stranger involvement* condition received an identical AV dilemma about a stranger called Sam (with either full or partial contextual accessibility; see Appendix A). All participants also received visual stimuli that accompanied the scenario. Visual stimuli only differed between type of involvement conditions, where the arrow indicating the scenario agent is labelled *This is you* (see Figure 4A) in the participant involvement condition, and *This is Sam* (see Figure 4B) in the stranger involvement condition. The experiment was over once participants had independently indicated their judgements of moral appropriateness and willingness to buy each AV model (a utilitarian swerve AV, and a non-utilitarian stay AV).

Figure 4

The Visual Stimuli Presented to Participants in Experiment 1

A. B.



Note. Panel A. Visual stimulus used in the participant involvement condition. Panel B. Visual stimulus used in the stranger involvement condition (see Appendix B for all visual stimuli).

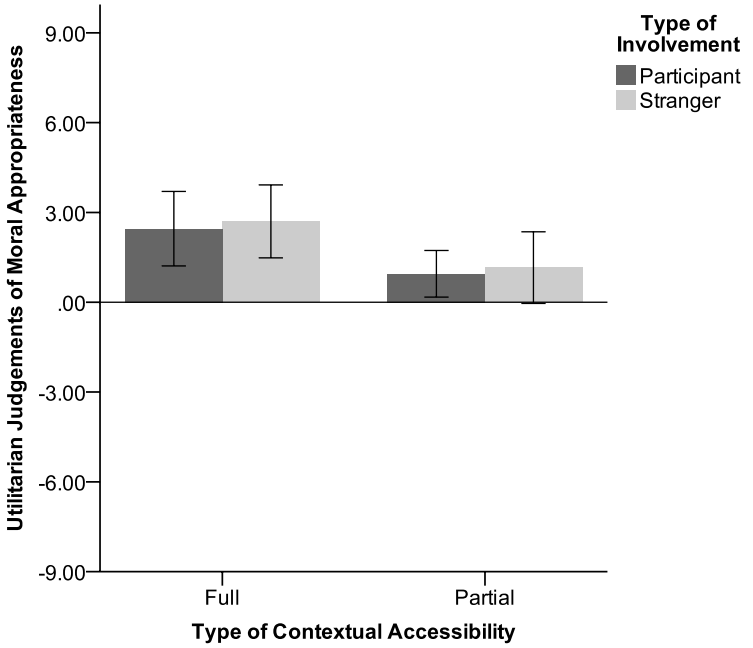
2.3.3 Results

2.3.3.1 Judgements of Moral Appropriateness

A 2x2 independent measures analysis of variance was conducted to explore the influence of the independent variables (type of contextual accessibility and type of involvement) on judgements of moral appropriateness. The results revealed a significant main effect of type of contextual accessibility, $F(1, 185) = 7.41$, $p = .007$, $\eta_p^2 = .039$ on judgements of moral appropriateness. However, the main effect of type of involvement on judgements of moral appropriateness and the interaction effect of type of involvement by type of contextual accessibility were not significant ($F < 1$); see Figure 5. Specifically, with full contextual accessibility, respondents' judgements of moral appropriateness were significantly more utilitarian ($M = 2.58$; $SD = 4.22$) than the judgements of moral appropriateness with partial contextual accessibility ($M = 1.05$; $SD = 3.39$); see Figure 5.

Figure 5

Participants' Utilitarian Judgements of Moral Appropriateness in Experiment 1



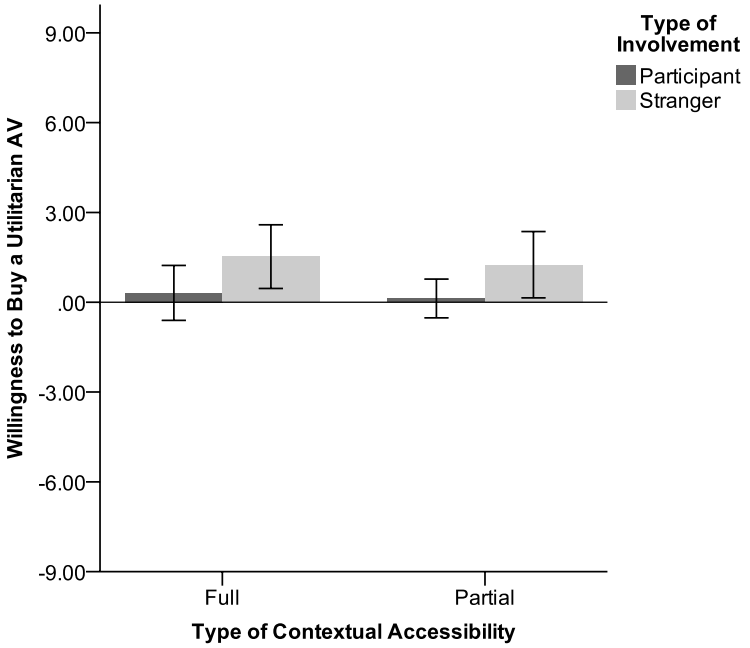
Note. Positive mean values indicate participants' preference for utilitarian swerve AVs. Negative mean values indicate participants' preference for non-utilitarian stay AVs. Error bars represent 95% confidence intervals of the mean.

2.3.3.2 Willingness to Buy

A 2x2 independent measures analysis of variance was conducted to explore the influence of the independent variables (type of contextual accessibility and type of involvement) on willingness to buy an AV. The results revealed a significant main effect of type of involvement $F(1, 185) = 6.16$, $p = .014$, $\eta_p^2 = .032$. Specifically, respondents were significantly more willing to buy a utilitarian swerve AV when they read a dilemma involving a stranger ($M = 1.39$; $SD = 3.66$) than when they read a dilemma involving a participant ($M = .22$; $SD = 2.74$); see Figure 6. Moreover, the results also revealed that the main effect type of contextual accessibility, as well as the two-way interaction effect of type of involvement by type of contextual accessibility ($F < 1$), were not statistically significant (see Figure 6).

Figure 6

Participants' Willingness to Buy a Utilitarian AV in Experiment 1



Note. Positive mean values indicate participants' preference for utilitarian swerve AVs. Negative mean values indicate participants' preference for non-utilitarian stay AVs. Error bars represent 95% confidence intervals of the mean.

2.3.4 Discussion

As predicted, presenting full contextual accessibility induced utilitarian judgements of moral appropriateness. In particular, participants were more likely to judge utilitarian swerve AVs as the most morally appropriate vehicle when they received full contextual accessibility compared to when they received partial contextual accessibility. These findings support Kusev et al.'s (2016) article, since the present experiment demonstrates that full contextual accessibility induces utilitarian behaviour outside of the trolley dilemma paradigm. However, in contrast, full contextual accessibility had no effect on willingness to buy judgements and this was the case for both types of involvement. Hence, whilst full contextual accessibility induced utilitarian judgements of moral appropriateness, it did not affect the participants' willingness to buy utilitarian swerve AVs. This could be explained by the qualitative

distinctions between judgements of moral appropriateness and judgements related to moral action (e.g., choosing to buy a utilitarian or non-utilitarian AV; e.g., Garrigan, et al., 2016; Tassy et al., 2013). For instance, when offering a judgement of moral appropriateness, participants are merely required to evaluate the morality of a particular behaviour (e.g., from a philosophical standpoint). However, when offering a moral action-related judgement (e.g., purchasing an AV), participants are involving themselves in the judgement outcomes which may prompt them to behave more egoistically. Moreover, it is plausible that this egoistic tendency may be accentuated when scenarios are described with partial PT accessibility, since partial PT accessibility presented in Bonnefon et al. (2016) biases participants towards only one perspective of the situation.

Interestingly, participants were more utilitarian in their willingness to buy judgements when they read scenarios involving a stranger compared to when they read scenarios involving themselves. Importantly that was only the case with moral purchasing behaviour, which was also not informed by contextual accessibility. This could be accounted for by the distinguishing features of imagining how another person feels in their situation (e.g., the stranger) versus imagining how *you* would feel in their situation (e.g., the participant). In particular, the former has been associated with feelings of empathy, inducing moral behaviour (e.g., fairness towards others) whilst the latter elicits feelings of personal distress, inducing immoral behaviour (unfairness towards others; Batson et al., 1997b, 2003; Stotland, 1969). Consequently, this effect of type of involvement will be explored further in Chapters 3-5.

Experiment 1 has therefore demonstrated that offering full contextual accessibility to AV crash scenarios influences participants' moral judgements but not their action-related judgements (e.g., willingness to buy AVs). Specifically, the results from Experiment 1 revealed that contextual accessibility is not a general account for moral decision-making (unable to determine both moral judgements and purchasing behaviours). Moreover, when the purchasing

behaviour is not informed by contextual accessibility, other factors (e.g., type of involvement) influence and determine participants' utilitarian behaviour. Consequently, I propose, develop and apply the full PT accessibility theory and method for moral decision-making tasks. Accordingly, decision-makers that receive full PT accessibility to moral tasks have access to not only all explicit details of decision scenarios (full contextual accessibility) but to all situational perspectives (or full PT accessibility). Since Experiment 1 has only employed moral scenarios with partial PT accessibility, it is worthwhile pursuing experiments that offer full PT accessibility. It is plausible that partial PT offered in moral AV dilemmas biases participants' responses towards one situational perspective. These findings highlight the need for an experimental exploration into the enhancement of PT accessibility, particularly in AV dilemmas where PT is a major feature but is currently limited to a single perspective.

CHAPTER 3

Perspective-Taking Accessibility: Moral Judgements and Purchasing Behaviours

3.1 Overview of Chapter 3

Chapter 3 comprises of 3 Experiments (Experiments 2-4) which aim to test the effect of full and partial PT accessibility on participants responses to various moral judgement tasks whilst keeping full contextual accessibility constant. In Experiment 2, I explore the influence of *type of PT accessibility* (whether the participant has partial or full access to perspective-taking in the scenario) and *type of involvement* (whether the participant takes the perspective of themselves or a stranger in the scenario) on participants' judgements of moral appropriateness for utilitarian swerve and non-utilitarian stay AVs. Accordingly, Experiment 2 revealed that participants' judgements of moral appropriateness were influenced by type of PT accessibility but not by type of involvement. In particular, full PT accessibility enhanced participants' utilitarian judgements of moral appropriateness.

In Experiment 3, I explored the influence of type of accessibility and type of involvement on participants' judgements of moral appropriateness and purchasing behaviour (purchasing values; how much money participants spent on utilitarian swerve and non-utilitarian stay AVs). Accordingly, the results revealed that full PT accessibility enhanced utilitarian judgement of moral appropriateness and purchasing values. Interestingly, with full PT accessibility, participants' purchasing values were informed by their judgements of moral appropriateness. Moreover, in contrast to the results with partial PT accessibility, presenting the participants with full PT accessibility eliminated the difference in purchasing values between participant involvement and stranger involvement conditions.

In the final experiment of this chapter (Experiment 4), I explored how PT accessibility and type of involvement influences participants' judgements of moral appropriateness, purchasing behaviour (willingness to buy an AV) and usage behaviour (willingness to ride an AV). Moreover, in the independent variable type of involvement, the level *stranger involvement* was replaced with *participant and family member involvement* (where participants

had to take the perspective of themselves and a family member in the scenario). The results of Experiment 4 revealed that full PT accessibility enhanced participants' judgements of moral appropriateness and induced participants' willingness to buy and ride utilitarian AVs. Moreover, with full PT accessibility, participants' judgements of moral appropriateness informed their willingness to buy and ride AVs. As with Experiment 3, and in contrast to conditions with partial PT accessibility, full PT accessibility eliminated the difference in judgements of moral appropriateness between participant involvement and participant and family member involvement conditions.

3.2 Experiment 2: The Influence of Perspective-Taking Accessibility on Judgements of Moral Appropriateness

3.2.1 Introduction

Experiment 1 demonstrated that providing full accessibility to dilemma actions and consequences (full contextual accessibility; Kusev et al., 2016), enhanced participants' utilitarian judgements of moral appropriateness. In particular, participants who received full contextual accessibility were more likely to morally approve of utilitarian swerve AVs and less likely to morally approve of non-utilitarian stay AVs. In Experiment 2 as well as all subsequent experiments in this thesis, full contextual accessibility will be kept constant. This is because the findings from Kusev et al. (2016) and Experiment 1 confirm that full contextual accessibility is necessary to provide unbiased contextual information in moral dilemmas. However, full PT accessibility will be developed and manipulated in this experiment; therefore, whilst all participants will receive full contextual accessibility, only half will receive full PT accessibility and the other half will receive partial PT accessibility. Accordingly participants in the full PT accessibility condition will be presented with scenarios that offer the perspective of both the AV passenger and the pedestrians, whereas participants in the partial PT accessibility condition will be presented with scenarios that offer the perspective of the AV

passenger only (see Experiment 2 method section). Therefore, the purpose of Experiment 2 is to test (for the first time) the influence of variations in PT accessibility in AV dilemmas on participants' judgements of moral appropriateness for AVs. Although participants' judgements of moral appropriateness are generally utilitarian (particularly when contextual accessibility is made full) it is anticipated that providing full PT accessibility will further enhance this utilitarian moral pattern of results.

3.2.2 Method

3.2.2.1 Participants

Participants ($N = 320$) were recruited via PureProfile's online survey panel. The sample consisted of 168 females and 152 males and the mean age of the participants was 45 ($SD = 15.35$). Prior to data collection, ethical approval was obtained from the BSREC and participants were treated in accordance with BPS ethical standards. A significance level of .05 was set for statistical testing. Moreover, a retrospective power analysis was conducted on the independent-measures effects of *type of PT accessibility* and *type of involvement* and their interaction. The experiment was live for 14 days to ensure data collection from both a sufficiently large sample, and a large effect size, will achieve a statistical power of at least .95. According to the retrospective power analysis, the sample size ($N = 320$) produced a power of 1.00 which was sufficient to achieve the target.

3.2.2.2 Experimental Design

A 2 (type of PT accessibility) x 2 (type of involvement) independent measures design was employed. The first independent variable, type of PT accessibility, had two levels (full and partial) where full PT accessibility refers to a scenario and question that allows the decision-maker to take the perspective of both the AV passenger and one of the pedestrians in the road, whereas partial PT accessibility allows only the perspective of the AV passenger. As in

Experiment 1, the second independent variable, type of involvement, had two levels (participant and stranger involvement). There was one dependent measure in this experiment: judgements of moral appropriateness. Accordingly, the dependent variable was measured using the same method as in Experiment 1 (separate judgements for utilitarian swerve and non-utilitarian stay AV models, each rated on a 10-point scale [0-9], from which the utilitarian weight was later calculated).

3.2.2.3 Materials and Procedure

Each participant took part in one of the four experimental conditions (based on all combinations of the 2 independent variables) where they were presented with an AV dilemma. For example, participants who received partial PT accessibility with participant involvement received the following dilemma and questions that provides access to one perspective of the scenario (the car passenger):

> *You are the sole passenger in an autonomous self-driving vehicle travelling at the speed limit down a main road. Suddenly, 10 pedestrians appear ahead, in the direct path of the car. The car could be programmed to either **SWERVE** off to the side of the road, where it will impact a barrier, killing you but leaving the 10 pedestrians unharmed or **STAY** on its current path where it will kill the 10 pedestrians, but you will be unharmed (see the picture illustrating this scenario).*
>
> *Judge the moral appropriateness of programming autonomous self-driving cars to: **SWERVE**, killing you but leaving the 10 pedestrians unharmed. If the car does not swerve it will kill the 10 pedestrians but leave you unharmed.*
>
> *Judge the moral appropriateness of programming autonomous self-driving cars to: **STAY,** killing the 10 pedestrians but leaving you unharmed. If the car does not stay, it will kill you but leave the 10 pedestrians unharmed.*

Alternatively, participants who received full PT accessibility with participant involvement received the following scenario and questions that provides access to multiple perspectives of the situation (the car passenger and one of the pedestrians):

*You could be the sole passenger in an autonomous self-driving vehicle travelling at the speed limit down a main road. Or you could be one of the 10 pedestrians that have appeared ahead, in the direct path of the car. The car could be programmed to either **SWERVE** off to the side of the road, where it will impact a barrier, killing the passenger (that could be you) but leaving the 10 pedestrians unharmed or **STAY** on its current path where it will kill the 10 pedestrians (that could include you), but the passenger will be unharmed (see the picture illustrating this scenario).*

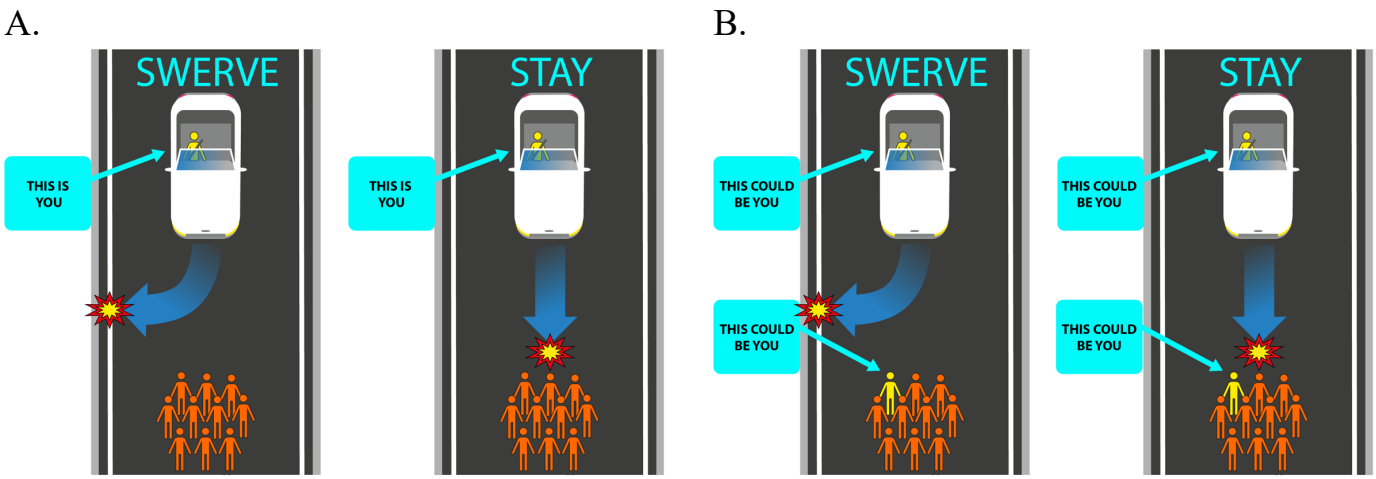
*Judge the moral appropriateness of programming autonomous self-driving cars to: **SWERVE**, killing the sole passenger (this could be you) but leaving the 10 pedestrians unharmed (this could include you). If the car does not swerve it will kill the 10 pedestrians (this could include you) but leave the sole passenger unharmed (this could be you).*

*Judge the moral appropriateness of programming autonomous self-driving cars to: **STAY**, killing the 10 pedestrians (this could be you) but leaving the sole passenger unharmed (this could include you). If the car does not stay, it will kill the sole passenger (this could be you) but leave the 10 pedestrians unharmed (this could include you).*

Note that full contextual accessibility employed in Experiment 1 was kept as a constant throughout all experimental conditions in Experiment 2. Moreover, as in Experiment 1, participants in the stranger involvement condition received a scenario, question and visual stimuli about a stranger called ‘Sam’ (see Appendix A and B for all materials used in the experiment). Importantly, in Experiment 2, the visual stimuli also differed across the two PT accessibility conditions in order to accommodate multiple perspectives in visual format (see Figure 7A and 7B for a comparison).

Figure 7

The Visual Stimuli Presented to Participants in the Participant Involvement Condition of Experiment 2



Note. Panel A. A visual representation of partial PT accessibility. Panel B. A visual representation of full PT accessibility.

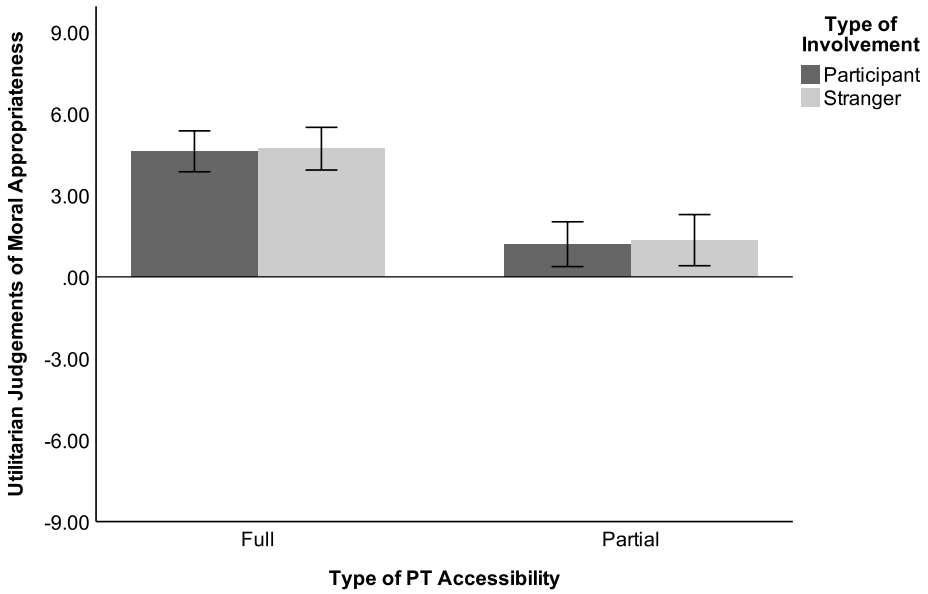
3.2.3 Results

3.2.3.1 Judgements of Moral Appropriateness

A 2x2 independent measures analysis of variance was conducted to explore the influence of the independent variables (type of PT accessibility and type of involvement) on judgements of moral appropriateness. The results revealed that type of PT accessibility significantly influenced respondents judgements of moral appropriateness $F(1, 316) = 66.28$, $p < .001$, $\eta_p^2 = .17$. Specifically, participants were more utilitarian in their moral judgements with full PT accessibility ($M = 4.66$; $SD = 3.44$) than with partial PT accessibility ($M = 1.27$; $SD = 3.96$). However, the results revealed that main effect of type of involvement ($F < 1$), as well as the two-way interaction of type of involvement by type of PT accessibility ($F < 1$) on judgements of moral appropriateness were not statistically significant (see Figure 8).

Figure 8

Participants' Utilitarian Judgements of Moral Appropriateness in Experiment 2



Note. Positive mean values indicate participants' preference for utilitarian swerve AVs. Negative mean values indicate participants' preference for non-utilitarian stay AVs. Error bars represent 95% confidence intervals of the mean.

3.2.4 Discussion

Kusev et al. (2016) demonstrated the importance of debiasing traditional moral dilemmas by introducing full accessibility to all decision actions and consequences. Accordingly, future research that utilises hypothetical scenarios for the purpose of behavioural elicitation should apply full accessibility in order to acquire unbiased judgements. However, some hypothetical scenarios involve PT where the participant must take the perspective of a character in the scenario who will be affected by decision outcomes (such as the AV dilemma; Bonnefon et al., 2016). Therefore, whilst contextual accessibility provides clear information relating to the actions and consequences of each choice option, it does not additionally provide full access to PT. Experiment 2 has demonstrated that presenting AV crash scenarios with full PT accessibility further enhances participants' utilitarian judgements of moral appropriateness

for AVs. The findings from Experiment 2 therefore demonstrates the requirement of PT accessibility in behavioural tasks that involve PT. On the other hand, type of involvement (stranger and participant) had no effect on participants' judgements of moral appropriateness.

Whilst these novel findings demonstrate that unbiased presentation of PT information and contextual information leads to an increase in utilitarian moral judgement it is also important to investigate further how PT accessibility impacts on the moral hypocrisy between participants' judgements of moral appropriateness and purchasing behaviours (Bonnefon et al., 2016; Martin et al., 2017).

3.3 Experiment 3: The Influence of Perspective-Taking Accessibility on Moral Purchasing Value

3.3.1 Introduction

Utilitarian AVs that aim to prioritise the lives of the greatest number of people are considered by potential consumers to be the most morally appropriate vehicle when compared to passenger-protective models (Bonnefon et al., 2016). Moreover, as explored in Experiments 1 and 2, this moral preference for utilitarian vehicles is enhanced when participants have access to all contextual information and situational perspectives of the moral crash scenarios. However, according to Bonnefon et al.'s (2016) empirical findings, people do not want to purchase utilitarian AVs despite judging them to be the most morally appropriate vehicle. As outlined in Chapter 1 and argued in Martin et al. (2017), the AV dilemmas presented to participants in Bonnefon et al.'s (2016) research were biased in PT, emphasising the small and negligible risk of being a utilitarian AV owner. For instance, in all formulations of Bonnefon et al.'s (2016) AV dilemmas, participants were required to take the perspective of the AV passenger (by imagining themselves and/or someone else in the AV). However, the authors did not offer the alternative perspective (that of the pedestrian). Therefore, by using a PT task,

Bonnefon et al.'s (2016) AV dilemma emphasises the 'danger' of being a passenger inside a utilitarian AV, but does not correspondingly emphasise the danger of being a pedestrian crossing the road in front of a non-utilitarian stay AV. Accordingly, Experiment 3 serves as an exploration into the influence of offering PT accessibility (full and partial) in AV crash scenarios on participants' judgements of moral appropriateness as well as participants' perceived purchasing value of utilitarian swerve and non-utilitarian stay AVs. It is accordingly anticipated that offering moral scenarios with full PT accessibility will induce utilitarian judgements of moral appropriateness (as with Experiment 2) and utilitarian purchasing values.

In addition to exploring whether PT accessibility independently influences participants' judgements of moral appropriateness and purchasing values, Experiment 3 will also explore the relationship between these 3 variables. In particular, a mediation analysis will be run to explore the relationship between the predictor (PT accessibility) and the outcome variable (purchasing behaviour), when the mediator (judgements of moral appropriateness) is included in the model. The purpose of this mediation analysis is to establish whether the moral hypocrisy explored in Bonnefon et al. (2016) remains under conditions of full PT accessibility. If the moral hypocrisy remains, then the mediation analysis should reveal that participants' judgements of moral appropriateness do not mediate the relationship between full PT accessibility and participants' purchasing values. However, if the moral hypocrisy is eliminated by the presentation of full PT accessibility then judgements of moral appropriateness should mediate the relationship between PT accessibility and purchasing values. In other words, when full PT accessibility is presented to participants, then those who judge utilitarian AVs as morally appropriate should also want to spend more money on utilitarian swerve AVs (compared to non-utilitarian stay AVs).

Similar to Experiments 1 and 2, in Experiment 3, participants will receive one of two types of involvement (participant or stranger involvement). As with the previous experiments,

participant involvement refers to conditions where the participant imagines themselves in the AV dilemma scenario. Alternatively, stranger involvement refers to conditions where the participants imagines a stranger in the AV dilemma scenario. These types of involvement were manipulated in Bonnefon et al. (2016) but have also been manipulated in other PT research (e.g., Batson, 1997b, 2003; Stotland, 1969). For example, Stotland (1969) distinguished between 'imagine-self' PT (where participants imagine how they would feel in a situation) and imagine-other PT (where participants imagine how someone else is feeling in their situation). Therefore, imagine-self PT is equivalent to the 'participant involvement' condition and imagine-other PT is equivalent with the 'stranger involvement' condition. Interestingly, in Experiment 1 of this thesis, there was a significant difference in participants' utilitarian purchasing behaviour between the two type of involvement conditions. In particular, participants in the stranger involvement (imagine-other) condition were more utilitarian in their purchasing behaviour (willingness to buy an AV) than participants in the participant involvement condition. However, this was under conditions of partial PT accessibility. Therefore, Experiment 3 will also investigate the interaction between type of PT accessibility (full and partial) and type of involvement (stranger and participant).

3.3.2 Method

3.3.2.1 Participants

In Experiment 3, Participants (N = 300; 165 females and 135 males) were recruited via PureProfile's survey panel. The mean age of the participants was 52 (SD = 14.93). All participants were treated in accordance with BPS ethical standards. A significance level of .05 was set for statistical testing. Moreover, a retrospective power analysis was conducted on the independent-measures effects of *type of PT accessibility* and *type of involvement* and their interaction. The experiment was live for 14 days to ensure data collection will achieve a

statistical power of at least .95. According to the retrospective power analysis, the sample size (N = 300) produced a power of 1.00 which was sufficient to achieve the target.

3.3.2.2 Experimental Design

A 2x2 independent measures design was employed. The first independent variable, *type of involvement*, had two levels including: (i) participant involvement (where the participant is described as an agent in the scenario) and (ii) stranger involvement (where a stranger is described as an agent in the scenario). The second independent variable, *type of PT accessibility*, also had two levels including: (i) partial PT accessibility (where the agent is described as being inside an AV in the scenario) and (ii) full PT accessibility (where it is made clear that the agent could potentially be inside the AV or part a group of pedestrians in the scenario). The first dependent variable was *judgements of moral appropriateness* where participants were required to judge on a 10-point rating scale (0-9) how morally appropriate they believed each AV (utilitarian swerve and non-utilitarian stay) model was (higher numerical ratings denoting a higher judgement of moral appropriateness). As in Experiments 1 and 2, a utilitarian weight was calculated by subtracting the judgements for non-utilitarian stay AV models from the judgements for utilitarian swerve AV models.

The second dependent variable was *purchasing value,* where participants were given a budget of £50,000 to distribute between the two AV models. Accordingly, purchasing value was calculated as the difference between the amount of money participants are willing to spend on utilitarian swerve and non-utilitarian stay AV models (where value for non-utilitarian stay AVs was subtracted from the value for utilitarian swerve AVs).

3.3.2.3 Materials and Procedure

Experiment 3 followed the same procedure and used the same materials as Experiment 1 (see Appendix A and B), where participants received a variation of the AV dilemma

(containing a scenario and visual stimuli), dependent on the condition in which they were allocated to and had to provide their judgements of moral appropriateness for each AV described. However, in Experiment 3, a second question was introduced (purchasing value), which required participants to indicate how much money they would spend on each AV from a £50,000 budget. The question was formulated as follows (see Appendix A for the exact presentation of this question in accordance with each PT and type of involvement condition):

> *Using a budget of £50,000, please indicate how much you would pay for the following autonomous self-driving cars (the entire £50,000 budget must be spent):a car that is programmed to swerve and a car that is programmed to stay.*

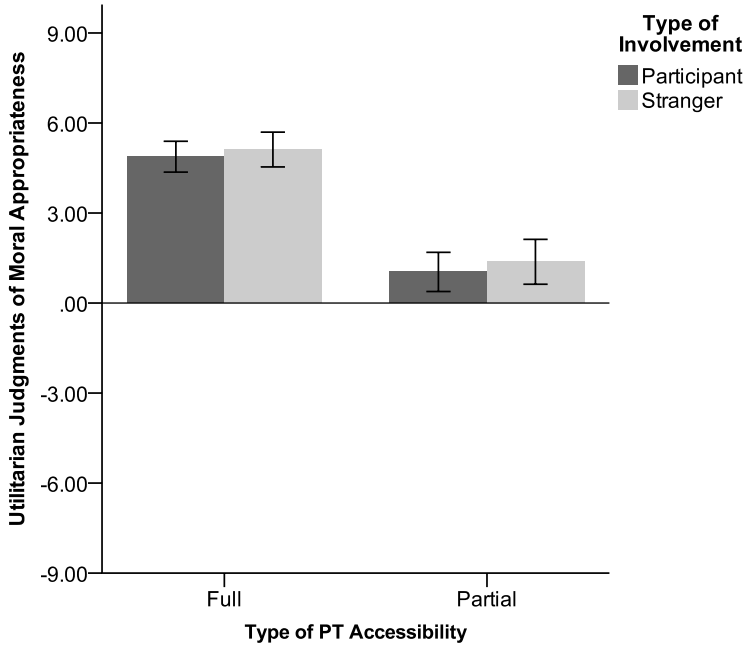
3.3.3 Results

3.3.3.1 Judgements of Moral Appropriateness

A 2x2 independent measures analysis of variance was conducted to explore the influence of the independent variables (type of PT accessibility and type of involvement) on judgements of moral appropriateness. The results revealed that type of PT accessibility significantly influenced respondents judgements of moral appropriateness $F(1, 296) = 143.90$, $p < .001$, $\eta_p^2 = .33$. Specifically, participants were more utilitarian in their moral judgements with full PT accessibility ($M = 5.00$; $SD = 2.37$) than with partial PT accessibility ($M = 1.21$; $SD = 3.05$). However, the results revealed that main effect of type of involvement ($F < 1$), as well as the two-way interaction of type of involvement by type of PT accessibility ($F < 1$) on judgements of moral appropriateness were not statistically significant (see Figure 9).

Figure 9

Participants' Utilitarian Judgements of Moral Appropriateness in Experiment 3



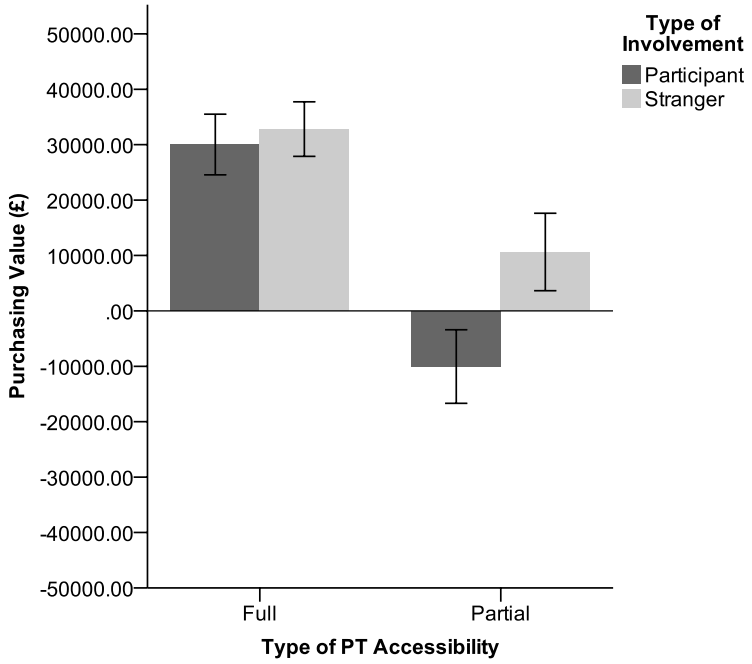
Note. Positive mean values indicate participants' preference for utilitarian swerve AVs. Negative mean values indicate participants' preference for non-utilitarian stay AVs. Error bars represent 95% confidence intervals of the mean.

3.3.3.2 Purchasing Value

A 2x2 independent measures analysis of variance was conducted to explore the influence of type of PT accessibility and type of involvement on purchasing value. The results revealed a significant main effect of both type of PT accessibility, $F(1, 296) = 104.52$, $p < .001$, $\eta_p^2 = .26$, and type of involvement, $F(1, 296) = 14.86$, $p < .001$, $\eta_p^2 = .05$, on purchasing value. Moreover, there was also a significant two-way interaction effect of type of PT accessibility by type of involvement $F(1, 296) = 8.64$, $p = .004$, $\eta_p^2 = .03$ (see Figure 10). Due to the significant 2-way interaction, follow-up simple-effect tests were conducted by type of PT accessibility.

Figure 10

Participants' Reported Purchasing Values for Utilitarian AVs in Experiment 3



Note. Positive purchasing values indicate utilitarian behaviour (more money spent on swerve AVs than stay AVs from the budget of £50,000) and negative purchasing values indicate non-utilitarian behaviour (more money spend on stay AVs than swerve AVs from the budget of £50,000). Error bars represent 95% Confidence Intervals of the mean.

Partial PT Accessibility. A follow-up simple-effect test revealed that with partial PT accessibility, the main effect of type of involvement significantly influenced respondents purchasing value $F(1, 148) = 18.27$, $p < .001$, $\eta_p^2 = .11$ (see Figure 10). Specifically, participants indicated that they would pay £10,640.08 (M = £10,640; SD = £30,374.08) more for utilitarian swerve AV than non-utilitarian stay AVs (when the scenario they read involved a stranger) and £10,040 (M = -£10,040; SD = £28,865.22) less for a utilitarian swerve AV than non-utilitarian stay AVs (when the scenario they read involved themselves). Accordingly, with partial PT accessibility, participants were relatively more utilitarian when a stranger was the agent in the scenario compared to when participant was the agent in the scenario.

Full PT accessibility. A second follow-up simple-effect test revealed that with full PT accessibility, type of involvement did not significantly influence purchasing value ($F < 1$). Accordingly, with full PT accessibility, participants indicated that they would pay £30,026.67 (M = £30,026.67; SD = £23,804.37) more for a utilitarian swerve AV than a non-utilitarian stay AV (when the scenario they read involved themselves) and £32,813.33 (M = £32,813.33; SD = £21,384.32) more for a utilitarian swerve AV than a non-utilitarian stay AV (when the scenario they read involved a stranger; see Figure 10). Therefore, with full PT accessibility (i) respondents were utilitarian in their purchasing behaviour for both types of involvement (when the agent described in the scenario is a stranger or the participant in the study) and (ii) the difference in purchasing value between type of involvement was eliminated.

3.3.3.3 Predicting Purchasing Value

Two mediation analyses (by type of involvement: stranger and participant) were conducted with macro PROCESS (Hayes, 2017) to test whether the respondents' judgements of moral appropriateness mediates the relationship between type of PT accessibility and reported purchasing values. The predictor variable was PT accessibility, the mediator was participants' judgements of moral appropriateness and the outcome variable was respondents reported purchasing values. The indirect effect of PT accessibility through the mediator judgements of moral appropriateness was tested by bootstrapping with N = 5000. The results established that decision-makers' judgements of moral appropriateness is a mediator of the relationship between type of PT accessibility (full and partial) and their reported purchasing values. Hence, respondents' purchasing behaviour was informed by their moral judgements (the utilitarian weight of moral appropriateness).

Stranger Involvement. The mediation model was significant, $F(2, 147) = 60.49$, $p < .001$; the model explained 45% of the variance in purchasing values ($R^2 = .45$). In addition, the standardized total effect of PT accessibility on purchasing value was also significant (β = -

.39, t = -5.17, p < .001). The results also revealed that with stranger involvement, the standardized indirect effect of PT accessibility through the mediator judgements of moral appropriateness was significant and negative, β = -.35, BCa CI(.95) = [-.479; -.248], indicating that judgements of moral appropriateness is a mediator of the relationship between PT accessibility and purchasing values. Moreover, the results revealed that judgements of moral appropriateness fully mediated the relationship between type of PT accessibility and purchasing value as the standardized direct effect of PT accessibility on purchasing value was not significant in the mediation model (β = -.04, t = -.51, p = .608). Specifically, respondents reported purchasing values for swerve AVs were fully mediated by the judgements of moral appropriateness and were higher in the full PT accessibility condition than in the partial PT accessibility condition.

Participant Involvement. The mediation model was significant, F(2, 147) = 86.23, p <.001; the model explained 54% of the variance in purchasing values (R^2 = .54). In addition, the standardized total effect of PT accessibility on purchasing value was also significant (β = -.60, t = -9.27, p < .001). The results also revealed that with participant involvement, the standardized indirect effect of PT accessibility through the mediator judgements of moral appropriateness was significant and negative, β = -.31, BCa CI(.95) = [-.420; -.220], indicating that judgements of moral appropriateness is a mediator of the relationship between PT accessibility and purchasing value. We found that judgements of moral appropriateness partially mediated the relationship between PT accessibility and purchasing value as the standardized direct effect of PT accessibility on purchasing value was significant in the mediation model; however, this effect was weakened from (standardized total effect β = -.60, t = -9.27, p < .001) to (standardized direct effect β = -.29, t = -4.16, p < .001) when the mediator was included as a predictor. Specifically, with participant involvement, participants' reported purchasing values for utilitarian swerve AVs were partially mediated by the judgements of

moral appropriateness and were higher in the full PT accessibility condition than in the partial PT accessibility condition. As predicted, participants' utilitarian purchasing values were influenced by PT accessibility and informed by the moral judgements.

3.3.4 Discussion

In contrast to Bonnefon et al. (2016), who demonstrated a moral hypocrisy between participants' judgements of moral appropriateness and purchasing behaviour, Experiment 3 revealed that with full PT accessibility, participants demonstrate a preference for utilitarian AVs across judgements tasks and type of involvement conditions. For example, Experiment 3 demonstrated that offering full PT accessibility to participants enhanced their utilitarian judgements of moral appropriateness. This was also true for participants' purchasing values, where offering full PT accessibility to participants resulted in them wanting to spend more money on utilitarian-swerve AVs than non-utilitarian-stay AVs.

Perhaps even more interestingly, when participants received full PT accessibility to AV crash scenarios, participants' purchasing values were informed by their judgements of moral appropriateness. For example, participants who received full PT accessibility judged utilitarian-swerve AVs as more moral than non-utilitarian-stay AVs, and subsequently spend more money on utilitarian-swerve AVs than non-utilitarian-stay AVs. Therefore, offering participants full PT accessibility to crash scenarios eliminated the moral hypocrisy between participants' moral judgements and purchasing behaviours observed in Bonnefon et al. (2016).

In addition to eliminating moral hypocritical behaviour, presenting crash scenarios with full PT accessibility also eliminated previously observed differences between the type of involvement conditions (e.g., Batson, 1997b; Bonnefon et al., 2016). For example, when participants received partial PT accessibility there was a significant difference in purchasing values between participants that read scenarios about themselves and participants that read

scenarios about a stranger. However, in the full PT accessibility condition, this difference no longer existed; participant demonstrated preferences for utilitarian AVs across judgement tasks and regardless of the type of involvement. Accordingly, these findings not only inform Bonnefon et al. (2016) but also previous research related to moral hypocrisy and PT tasks (Batson et al., 1997a, 1997b, 1999, 2003; Batson & Thomson, 2001; Lönnqvist et al., 2014).

Experiment 3 therefore demonstrates that presenting an unbiased representation of AV crash scenarios results in consistent utilitarian responses, indicating normative utilitarian behaviour (von Neuman & Morgenstern, 1944). Accordingly, these findings further highlight the importance of presenting unbiased information to participants when attempting to elicit behavioural responses (e.g., moral judgements regarding AVs crashes). Previous research has demonstrated that the behavioural elicitation methods can determine participants choices (e.g., Kusev et al., 2020). However, for the first-time, variations in accessibility to PT has been experimentally manipulated and demonstrated the undeniable affect that biased information has on people's moral judgements.

3.4 Experiment 4: The Influence of Perspective-Taking Accessibility on Participants' Willingness to Buy and Ride AVs

3.4.1 Introduction

One of the findings from Experiment 3 revealed that presenting participants with full PT accessibility to AV crash scenarios eliminated behavioural differences between the participant involvement and stranger involvement conditions. This finding demonstrates that differences between type of involvement conditions often exhibited in PT studies (e.g., Batson et al., 2003; Bonnefon et al., 2016) may be the result of the experimenters offering only partial PT accessibility to participants. However, in another variation of Bonnefon et al.'s (2016) study, participants were required to imagine themselves and a family member inside an AV. It

was in this variation that participants demonstrated most strongly their aversion to purchasing utilitarian-swerve AVs; they demonstrated a willingness to purchase non-utilitarian stay AVs (see Bonnefon et al., 2016; study 3). This effect of family member involvement on non-utilitarian preferences has also been demonstrated in moral decision-making research involving the trolley paradigm (O'Neill & Petrinovich, 1998; Tassy et al., 2013). In particular, Tassy et al. (2013) found that when considering moral moral dilemmas, people are more protective of close family members over distant family members and strangers even when this conflicts with a utilitarian choice (Tassy et al., 2013). Therefore, it is important to test whether offering scenarios will full PT accessibility will result in utilitarian preferences for AVs even when participants must imagine themselves alongside a family member in the scenario. Accordingly, in Experiment 4, stranger involvement is replaced with participant and family member involvement. It is anticipated that presenting AV crash scenario's with partial PT accessibility will induce utilitarian preferences in participants who took the perspective of themselves and a family member in the AV.

In addition to the new independent variable, two new dependent variables are introduced in this experiment in order to overcome a potential shortcoming of the dependent variable *purchasing value* in Experiment 3. For instance, whilst Experiment 3 measured participants' purchasing behaviour in terms of how much they would spend on a particular AV (purchasing value), it did not capture whether they want to purchase an AV in the first place. Therefore, Experiment 3 did not measure participants' willingness to buy utilitarian swerve and non-utilitarian stay AVs. Accordingly, Experiment 4 measured participants' willingness to buy each AV model. Moreover, in addition to eliciting participants' willingness to buy judgements, Experiment 4 also investigates usage behaviour, or participants' willingness to ride AVs. The logic behind measuring participants' usage behaviour is to remove a potential confounding variable with willingness to buy judgements. It is plausible that whilst some

people do not want to buy a utilitarian AV that could potentially sacrifice them in a crash scenario because it does not make intuitive sense to pay for a negative outcome. Therefore, introducing willingness to ride as an additional variable may capture more clearly people's utilitarian intentions regarding the appropriate programming of AVs.

The purpose of Experiment 4 is therefore to establish the effect of type of PT accessibility (full and partial) and type of involvement (participants or participant and family member) on participants' judgements of moral appropriateness, willingness to buy and willingness to ride each utilitarian-swerve and non-utilitarian-stay AV model.

3.4.2 Method

3.4.2.1 Participants

For Experiment 4, participants (N = 300) were recruited through PureProfile's survey panel. The sample consisted of 160 females and 140 males and the mean age was 51 (SD = 14.35). Ethical approval was granted by the BSREC prior to data collection and all participants were treated in accordance with BPS ethical standards. According to power analysis, the statistical power of 2x2 ANOVA with 300 participants was identical to that presented for Experiment 3.

3.4.2.2 Experimental Design

A 2x2 independent measures design was employed to measure the influence of type of involvement (participant or participant with family member) and PT accessibility (full or partial) on *judgements of moral appropriateness, willingness to buy* and *willingness to ride* for each AV model. All dependent variables were measured on a 10-point rating scale where 0 indicated the lowest moral appropriateness/lowest willingness. Judgements of moral appropriateness were computed as utilitarian weight – the difference between participants' judgements of moral appropriateness for utilitarian swerve and non-utilitarian stay AVs.

Similarly, the outcome variables willingness to buy and willingness ride were computed as utilitarian weights too. Specifically, the judgements (willingness to buy and ride) for non-utilitarian stay AVs were subtracted from judgements (willingness to buy and ride) of utilitarian swerve AVs in order to generate a utilitarian weight. Therefore, positive and high difference (utilitarian weight) indicated utilitarian judgements. The first independent variable, type of PT accessibility was identical to that of Experiments 2 and 3. The second independent variable, type of involvement, was similar to the first independent variable in the previous 3 experiments, except stranger involvement was replaced with participant and family involvement (where both the participant and their family member are described as being agents in the scenario).

3.4.2.3 Materials and Procedure

Participants were presented with a scenario (and visual stimuli) depending on the condition they were allocated to and provided their judgements of moral appropriateness (see Appendix A and B for all experimental materials). In scenarios that involved a participant and family member, the number of pedestrians were doubled from 10 to 20. We did this, following the same logic as used in (Bonnefon et al., 2016) by presenting 10 pedestrians for every 1 passenger (see also Figure 11). In addition to providing judgements of moral appropriateness, participants also answered the following questions (see Appendix A for the exact presentation of these questions in accordance with each PT and type of involvement condition):

How would you rate your willingness to BUY an autonomous self-driving car programmed to swerve?;

How would you rate your willingness to BUY an autonomous self-driving car programmed to stay?;

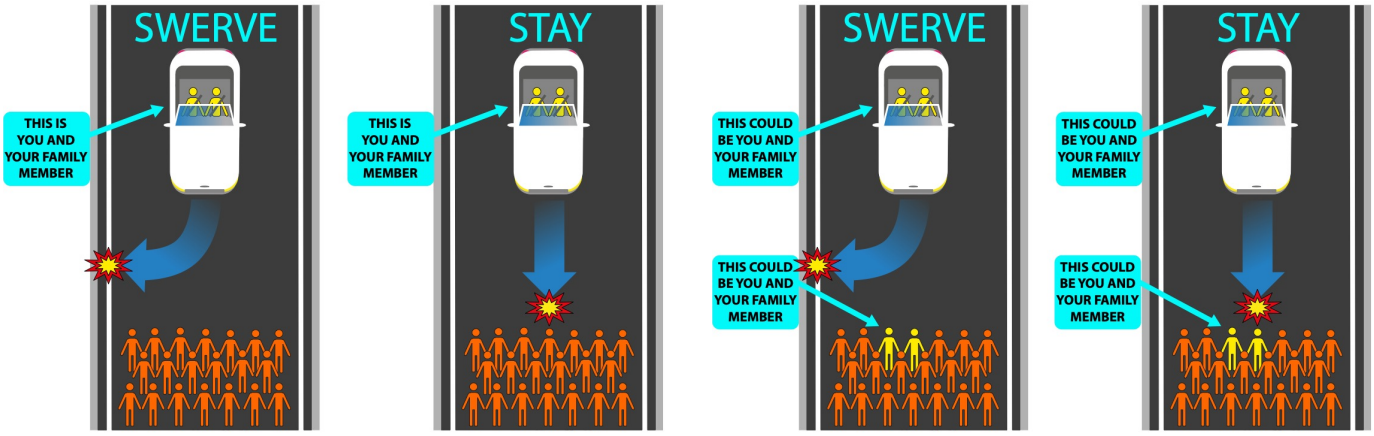
How would you rate your willingness to RIDE inside an autonomous self-driving car programmed to swerve?;

How would you rate your willingness to RIDE inside an autonomous self-driving car programmed to stay?

Figure 11

The Visual Stimuli Presented to Participants in Experiment 4

A. B.



Note. Example of visual stimuli presented to participants who received participant and family involvement scenarios. Panel A. Visual stimulus presented to participants in the Partial PT accessibility condition. Panel B. Visual stimulus presented to participants in the Full PT accessibility condition.

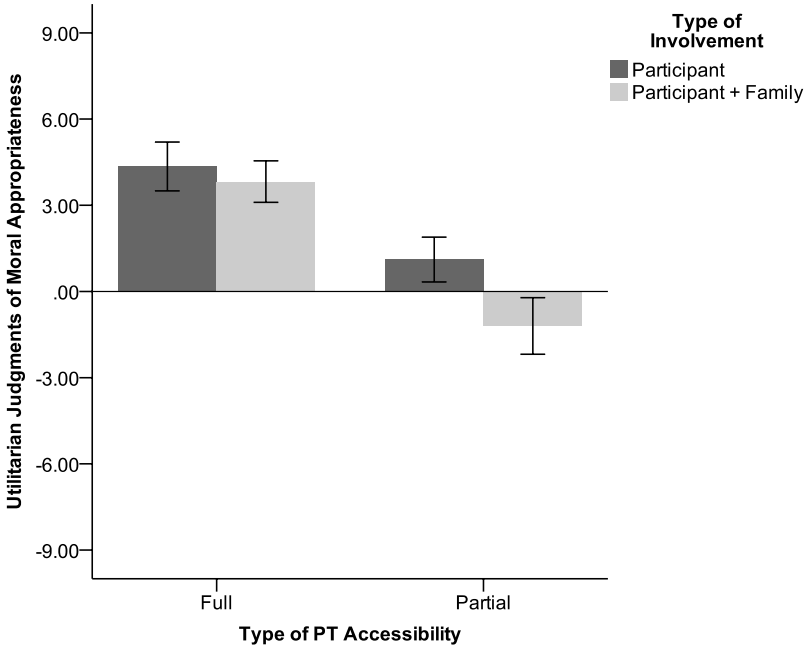
3.4.3 Results

3.4.3.1 Judgements of Moral Appropriateness

A 2x2 independent measures analysis of variance was conducted to explore the influence of the independent variables (type of PT accessibility and type of involvement) on judgements of moral appropriateness. The results revealed a significant main effect of type of PT accessibility, $F(1, 296) = 96.16$, $p < .001$, $\eta_p^2 = .25$, and type of involvement $F(1, 296) = 11.35$, $p = .001$, $\eta_p^2 = .04$ on judgements of moral appropriateness. Moreover, the results also revealed a significant two-way interaction effect of type of involvement by type of PT accessibility on judgements of moral appropriateness $F(1, 296) = 4.48$, $p = .035$, $\eta_p^2 = .02$ (see Figure 12). Accordingly, due to the significant two-way interaction, two follow-up analyses of variance were conducted by type of PT accessibility (partial and full PT accessibility).

Figure 12

Participants' Utilitarian Judgements of Moral Appropriateness in Experiment 4



Note. Positive mean values indicate participants' preference for utilitarian swerve AVs. Negative mean values indicate participants' preference for non-utilitarian stay AVs. Error bars represent 95% confidence intervals of the mean.

Partial PT Accessibility. A follow-up simple-effect test revealed that with partial PT accessibility, the main effect of type of involvement on judgements of moral appropriateness was significant $F(1, 148) = 13.45$, $p < .001$, $\eta_p^2 = .08$. Specifically, with participant involvement, respondents' judgements of moral appropriateness were utilitarian ($M = 1.11$ $SD = 3.39$) and significantly different than the non-utilitarian judgements of moral appropriateness with participant-and-family involvement ($M = -1.20$; $SD = 4.28$); see Figure 12.

Full PT Accessibility. In contrast to the pattern of judged moral appropriateness with partial PT accessibility, with full PT accessibility the main effect of type of involvement on judgements of moral appropriateness was not statistically significant ($F < 1$); see Figure 12. Accordingly, participants' judgements of moral appropriateness with participant-and-family involvement ($M = 3.83$ $SD = 3.13$) as well as participants' judgements of moral appropriateness

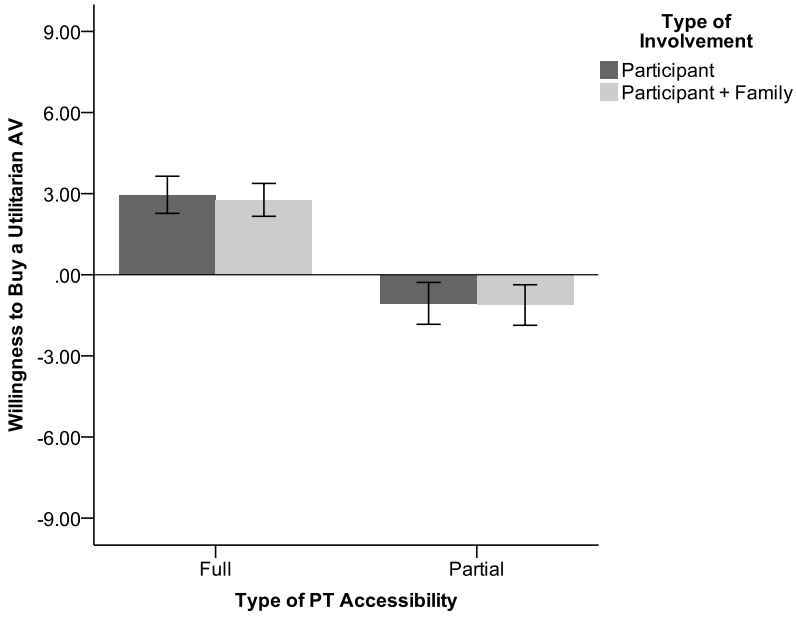
with participant involvement (M = 4.35 SD = 3.70) were both utilitarian and not statistically different.

3.4.3.2 Willingness to Buy

A 2x2 independent measures analysis of variance was conducted to explore the influence of the independent variables (type of PT accessibility and type of involvement) on willingness to buy each AV model. The results revealed a significant main effect of type of PT accessibility $F(1, 296) = 123.87$, $p < .001$, $\eta_p^2 = .30$. Specifically, with full PT accessibility, respondents' willingness to buy an AV was utilitarian (M = 2.86; SD = 2.81) and significantly different from the non-utilitarian willingness to buy an AV with partial PT accessibility (M = -1.09; SD = 3.30); see Figure 13. Moreover, the results also revealed that the main effect type of involvement ($F < 1$), as well as the two-way interaction effect of type of involvement by type of PT accessibility ($F < 1$), were not statistically significant (see Figure 13).

Figure 13

Participants' Willingness to Buy a Utilitarian AV in Experiment 4



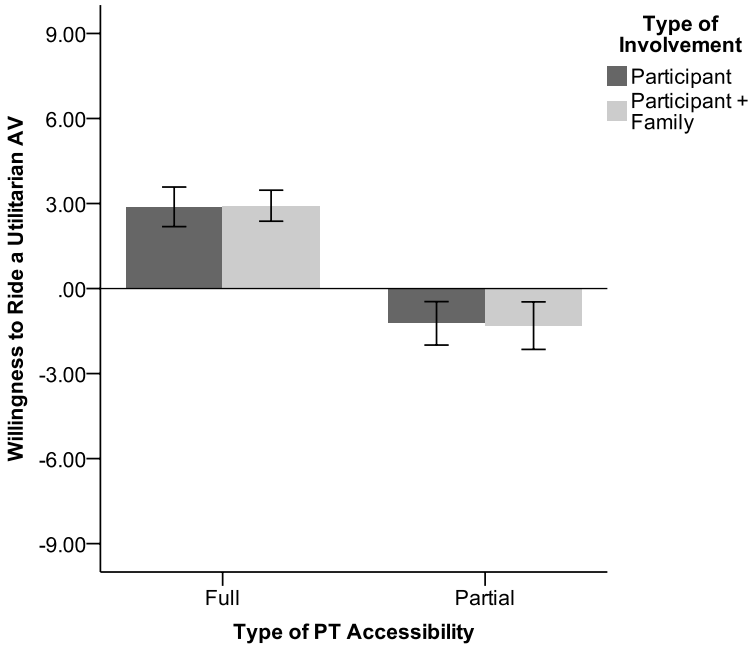
Note. Positive mean values indicate participants' preference for utilitarian swerve AVs. Negative mean values indicate participants' preference for non-utilitarian stay AVs. Error bars represent 95% confidence intervals of the mean.

3.4.3.3 Willingness to Ride

A 2x2 independent measures analysis of variance was conducted to explore the influence of type of PT accessibility (full and partial) and type of involvement (participant and participant and family) on willingness to ride each AV model. The results revealed that type of PT accessibility significantly influenced respondents willingness to ride an AV $F(1, 296) = 132.80, p < .001, \eta_p^2 = .31$ (see Figure 14). Accordingly, with full PT accessibility, respondents' willingness to ride an AV were utilitarian (M = 2.90; SD = 2.72) and significantly different from their non-utilitarian willingness to ride an AV with partial PT accessibility (M = -1.27; SD = 3.48); see Figure 14. Furthermore, the main effect of type of involvement ($F < 1$), as well as the two-way interaction effect of type of involvement by type of PT accessibility ($F < 1$) were statistically not significant (see Figure 14).

Figure 14

Participants' Willingness to Ride a Utilitarian AV in Experiment 4



Note. Positive mean values indicate participants' preference for utilitarian swerve AVs. Negative mean values indicate participants' preference for non-utilitarian stay AVs. Error bars represent 95% confidence intervals of the mean.

3.4.3.4 Predicting Willingness to Buy and Ride

Four mediation analyses (by type of involvement and type of willingness to buy and ride) were conducted with macro PROCESS (Hayes, 2017) to test whether the participants' judgements of moral appropriateness mediates the relationship between type of PT accessibility (full and partial) and reported willingness to buy and willingness to ride each model of AV (utilitarian swerve and non-utilitarian stay). The predictor variable was type of PT accessibility, the mediator was participants' judgements of moral appropriateness and the outcome variable were respondents' reported willingness to buy and ride. The indirect effect of PT accessibility through the mediator judgements of moral appropriateness was tested by bootstrapping with N = 5000. We found that decision-makers' judgements of moral appropriateness is a mediator of the relationship between type of PT accessibility (full and partial) and judgements for willingness to buy and willingness to ride. Moreover, we found that respondents' willingness to buy and ride judgements were informed by their moral judgements of appropriateness.

Willingness to Buy: Participant Involvement. The mediation model was significant, $F(2, 147) = 54.44$, $p < .001$. Moreover, the model explained 43% of the variance in willingness to buy ($R^2 = .43$). In addition, the standardized total effect of PT accessibility on willingness to buy was also significant ($\beta = -.54$, $t = -7.74$, $p < .001$). The results also revealed that with participant involvement, the standardized indirect effect of type of PT accessibility through the mediator judgements of moral appropriateness was significant and negative, $\beta = -.17$, BCa $CI(.95) = [-.250; -.107]$, indicating that judgements of moral appropriateness is a mediator of the relationship between type of PT accessibility and willingness to buy. Moreover, judgements

of moral appropriateness partially mediated the relationship between PT accessibility and willingness to buy as the standardized direct effect of PT accessibility on willingness to buy was significant in the mediation model; however, this effect was weakened from (standardized total effect β = -.54, t = -7.74, p < .001) to (standardized direct effect β = -.37, t = -5.32, p < .001) when the mediator was included as a predictor. Specifically, with participant involvement, respondents' reported willingness to buy swerve AVs were partially mediated by the judgements of moral appropriateness and were higher in the full PT accessibility condition than in the partial PT accessibility condition.

Willingness to Buy: Participant and Family Involvement. The second mediation model was also significant, $F(2, 147) = 51.40$, p <.001. Moreover, the model explained 41% of the variance in willingness to buy (R^2 = .41). In addition, the standardized total effect of PT accessibility on willingness to buy was also significant (β = -.55, t = -8.02, p < .001). The results also revealed that with participant and family member involvement, the standardized indirect effect of type of PT accessibility through the mediator judgements of moral appropriateness was significant and negative, β = -.22, BCa CI(.95) = [-.336; -.135], indicating that judgements of moral appropriateness is a mediator of the relationship between type of PT accessibility and willingness to buy. Furthermore, judgements of moral appropriateness partially mediated the relationship between PT accessibility and willingness to buy as the standardized direct effect of PT accessibility on willingness to buy was significant in the mediation model; however, this effect was weakened from (standardized total effect β = -.55, t = -8.02, p < .001) to (standardized direct effect β = -.33, t = -4.30, p < .001) when the mediator was included as a predictor. Specifically, with participant-and-family involvement, respondents' reported willingness to buy swerve AVs were partially mediated by the judgements of moral appropriateness and were higher in the full PT accessibility condition than in the partial PT accessibility condition.

Willingness to Ride: Participant Involvement. The third mediation model was also found to be significant, $F(2, 147) = 59.94$, $p < .001$; the model explained 45% of the variance in willingness to ride ($R^2 = .45$). In addition, the standardized total effect of PT accessibility on willingness to ride was also significant ($\beta = -.55$, $t = -7.91$, $p < .001$). The results also revealed that with participant involvement, the standardized indirect effect of type of PT accessibility through the mediator judgements of moral appropriateness was significant and negative, $\beta = -.18$, BCa $CI(.95) = [-.254; -.119]$, indicating that judgements of moral appropriateness is a mediator of the relationship between type of PT accessibility and willingness to ride. Moreover, judgements of moral appropriateness partially mediated the relationship between PT accessibility and willingness to ride as the standardized direct effect of PT accessibility on willingness to ride was significant in the mediation model; however, this effect was weakened from (standardized total effect $\beta = -.55$, $t = -7.91$, $p < .001$) to (standardized direct effect $\beta = -.37$, $t = -5.42$, $p < .001$) when the mediator was included as a predictor. Specifically, with participant involvement, respondents' reported willingness to ride swerve AVs were partially mediated by the judgements of moral appropriateness and were higher in the full PT accessibility condition than in the partial PT accessibility condition.

Willingness to Ride: Participant and Family Involvement. Finally, the fourth mediation model was significant, $F(2, 147) = 59.33$, $p < .001$. Moreover, the model explained 45% of the variance in willingness to buy ($R^2 = .45$). In addition, the standardized total effect of PT accessibility on willingness to ride was also significant ($\beta = -.57$, $t = -8.40$, $p < .001$). The results also revealed that with participant and family member involvement, the standardized indirect effect of type of PT accessibility through the mediator judgements of moral appropriateness was significant and negative, $\beta = -.24$, BCa $CI(.95) = [-.351; -.147]$, indicating that judgements of moral appropriateness is a mediator of the relationship between type of PT accessibility and willingness to ride. Furthermore, judgements of moral

appropriateness partially mediated the relationship between PT accessibility and willingness to ride as the standardized direct effect of PT accessibility on willingness to ride was significant in the mediation model; however, this effect was weakened from (β = -.57, t = -8.40, p < .001) to (β = -.33, t = -4.47, p < .001) when the mediator was included as a predictor. Specifically, with participant-and-family involvement, respondents' reported willingness to ride swerve AVs were partially mediated by the judgements of moral appropriateness and were higher in the full PT accessibility condition than in the partial PT accessibility condition.

3.4.4 Discussion

The results of Experiment 4 demonstrate that providing participants with full PT accessibility to moral crash scenarios results in them making consistent utilitarian judgements across all judgement tasks and all types of involvement. Hence, with unbiased information, people judge utilitarian-swerve AVs as more moral than non-utilitarian stay AVs and are subsequently more willing to buy and ride utilitarian-swerve AVs than non-utilitarian stay AVs. Conversely, when participants are presented with partial PT accessibility, they are overall less utilitarian in response to all judgement tasks and also demonstrate a moral hypocrisy; they judge utilitarian-swerve AVs as the most moral vehicle yet are more willing to buy and ride non-utilitarian stay (passenger-protective) AVs. Moreover, the results from the mediation analyses also confirmed that when participants received full PT accessibility, their level of willingness to buy and ride AVs was informed by their judgements of moral appropriateness. For example, when participants received crash scenarios containing full PT accessibility, they were more likely to judge utilitarian-swerve AVs as morally appropriate compared to non-utilitarian stay AVs and in turn were more willing to buy and ride utilitarian AVs.

Taken together, these findings indicate that the moral hypocrisy exemplified in Bonnefon et al. (2016) is likely the result of presenting limited PT accessibility (biased information) to participants. Therefore, any method that employs partial PT accessibility does

not reveal participants' (or consumers') true preferences for the ethical programming of AVs. Instead, presenting crash scenario from one narrow perspective causes a framing effect, exaggerating the danger of utilitarian AVs and inducing non-utilitarian preferences.

Experiment 4 also further illustrates the influence of variations in PT accessibility on type of involvement conditions. In particular, participants in the participant only condition approved of utilitarian-swerve AVs over non-utilitarian stay AVs, whereas participants in the participant and family member condition demonstrated the reverse; they judged non-utilitarian stay AVs as more morally appropriate than utilitarian-swerve AVs. However, under conditions of full PT accessibility this difference was eliminated, all participants regardless of the type of involvement employed, judged utilitarian AVs as the most morally appropriate vehicle. Therefore, offering full PT accessibility eliminates inconsistencies between type of involvement conditions, resulting in consistent utilitarian behaviour.

3.5 General Discussion

For decades, psychologists have implemented hypothetical scenarios as a safe and ethical way to measure human moral preferences (e.g., Bonnefon et al., 2016; Greene et al., 2001; Tversky & Kahneman, 1981). However, the way in which hypothetical scenarios are constructed have an (often unintended) influence on such preferences (Kusev at al., 2016; Martin et al., 2017; Tversky & Kahneman, 1981). For example, providing participants with only partial contextual accessibility to dilemma actions and consequences biases them toward non-utilitarian behaviour as well as utilitarian behavioural inconsistencies between dilemma types (Kusev et al., 2016). Accordingly, as a result of contextual accessibility not being accounted for in moral decision-making research (such as in Greene et al., 2001), researchers are grounding behavioural interpretations on preferences that have been biased by the decision-making task itself. Therefore, in the three Experiments of Chapter 3 I have developed PT

accessibility and empirically investigated the influence of PT accessibility on participants' moral judgements and purchasing behaviour.

Across 3 Experiments in Chapter 3, I introduced and experimentally manipulated PT accessibility in AV crash scenarios. Accordingly, the three experiments established that restricting PT accessibility to its partial form (as employed in Bonnefon et al., 2016) resulted in an overall low moral approval of utilitarian AVs (Experiments 2, 3 and 4) as well as inconsistencies in utilitarian preferences between behavioural tasks and types of involvement (Experiments 3 and 4). Importantly, these findings also replicated Bonnefon et al.'s (2016) study, who also found moral hypocrisies between participants' judgements of moral appropriateness and purchasing behaviours and moral inconsistencies between type of involvement conditions. However, in Experiments 2-3 of this thesis, PT accessibility was experimentally manipulated, so that some participants could experience full PT accessibility instead of the standard partial PT accessibility to dilemma information. Importantly, participants who received full PT accessibility were overall relatively utilitarian in the judgements of moral appropriateness. Moreover, these utilitarian judgements of moral appropriateness matched and informed their AV purchasing and usage behaviours. These findings indicate that moral hypocrisies demonstrated in Bonnefon et al.'s (2016) studies are a direct result of the method employed (partial PT accessibility). Therefore, Bonnefon and colleagues (2016) proposal does not account for or reveal consumers true moral preferences.

It is important to note that whilst partial PT accessibility makes utilitarian behaviour inaccessible to participants, full PT accessibility does not make non-utilitarian behaviour inaccessible. If one wanted to bias participants to making utilitarian judgements (nudging), this could be achieved by offering only utilitarian options (the perspective of the *pedestrian*) in AV crash scenarios. Full PT accessibility on the other hand, offers both the perspective of the passenger and the pedestrians, effectively balancing the available information, debiasing

respondents' interpretation of the scenario and revealing their true preferences which happens to be prosocial and utilitarian. Therefore, full PT accessibility is not a tool to make people behave a certain way, but rather a method of reducing bias in participants moral judgements and purchasing behaviour when partial PT information presented.

In light of these findings, claims that utilitarian AVs will be difficult to market to consumers and should therefore be scrapped altogether (e.g., Bonnefon et al., 2016; Greene, 2016, Sharif et al., 2017) sound inappropriate and far-fetched. This is because research that indicates a public rejection of utilitarian ethical algorithms (Bonnefon et al., 2016) is based on experiments that display limited PT accessibility to participants. Therefore, according to the findings from Experiments 2-4, it is plausible that marketing utilitarian AVs will not be a particularly challenging feat, since as informed by their judgements of moral appropriateness, people are willing to buy and ride utilitarian AVs (Experiment 4). Therefore, the Experiments in Chapter 3 reveal the importance of making ethically sensitive information fully accessible in order to prevent unnecessary biases in moral judgements and purchasing behaviour.

CHAPTER 4

Perspective-Taking Accessibility: Moral Judgements and Moral Choices

4.1 Overview of Chapter 4

A plethora of research offered evidence that human judgements and choices are psychologically dissociated (independent) and do not share similar psychological processes or properties (e.g., Hsee et al., 1996, 1999; Hsee & Zhang, 2010; Lichtenstein & Slovic, 1971; Slovic & Lichtenstein, 1968). Experiments 2, 3 and 4 in Chapter 3 provide evidence for the influence of full PT accessibility on participants' moral behaviour with judgement tasks. Accordingly, in Chapter 4 (Experiment 5) I investigate whether the full PT accessibility is a general psychological phenomenon that informs participants behaviour across different behavioural elicitation methods. The results in Experiment 5 confirmed that PT accessibility is indeed a fundamental psychological phenomenon. When participants were provided with full PT accessibility to both choice and judgement tasks the likelihood of moral utilitarian behaviour increased substantially. Importantly, there was a strong positive association between moral utilitarian judgements and moral utilitarian choices when the PT accessibility was full.

In Experiment 6 I further explore the relationship between moral choices and judgements of moral appropriateness using modified methodological approach developed and employed in the free-choice paradigm research (e.g., Brehm, 1956; Sharot et al, 2012). Researchers using the free-choice method predict that once people commit to a difficult decision (choose an option from a set of two which are similarly judged/evaluated) they tend to value and appreciate this option more than before; in other words, evidence for a choice-induced change in preferences (e.g., Brehm, 1956; Festinger, 1964; Izuma et al., 2010; Sharot et al., 2012). In Chapter 3 (Experiments 2, 3 and 4) I have established that when full PT accessibility is available, participants' judgements of moral appropriateness informed their moral purchasing behaviour. In Experiment 5 I also found that full PT accessibility influences both participants' choices and judgements, and that full PT accessibility made the association between choice and judgement strong. Accordingly, in Experiment 6, I explore whether full a

PT accessibility informed decision induces change in preferences for judgements of moral appropriateness with partial PT accessibility. The results established that when respondents made decisions informed by full PT accessibility, their judgements of moral appropriateness (in AV crash scenarios with partial PT accessibility) changed in the direction of the choice they made. This result provides evidence for a full PT accessibility transfer, from choice (when participants received AV crash scenarios with full PT accessibility) to judgements (when participants received AV crash scenarios with partial PT accessibility). Importantly, and in contrast to predictions from free-choice researchers, this full PT accessibility transfer takes place irrespective of whether decisions were difficult or easy. Moreover, this effect was large and superior to the effect of choice-induced change in preferences (small effect size) and the nonsignificant effect of dilemma difficulty (easy or difficult decisions).

Taken together the results in Chapter 3 and Chapter 4, for the first time, established that with full PT accessibility participants' behaviour was informed by their moral judgements. Moreover, full PT accessibility informed behaviour, changed participants judgements of moral appropriateness even when they judged these AV crash scenarios with partial PT accessibility (full PT accessibility transfer).

4.2 Experiment 5: The Influence of Perspective-Taking Accessibility on Moral Judgements and Moral Choices

4.2.1 Introduction

So far, all psychological measures across all Experiments in this thesis have required participants to make judgements. Appropriately, judgements provide an indication of an individuals' preference towards a target (e.g., an object, experience or outcome state) and as described by Lichtenstein and Slovic (1971), a variety of methods can be employed in order to obtain judgements. For instance, a rating scale can be employed in order to measure an individual's judgement of personal valence towards a target (e.g., the targets attractiveness).

Accordingly, this method follows the same logic as the judgements of moral appropriateness and willingness to buy/ride scales as developed in this thesis. Moreover, other types of judgements require an individual to indicate how much money they will spend on the target (or in Lichtenstein and Slovic's [1971] example, a gamble), as implemented as *purchasing value* in Experiment 3 of this thesis. Whilst these tasks measure different aspects of behaviour (e.g., moral appropriateness and purchasing value), they both elicit respondents' judgements, and are therefore both judgement tasks. However, judgement tasks are not the only behavioural elicitation method that can be employed in behavioural research. One can also simply ask people to make a choice between two decision options (binary choice). In the current context, this could mean deciding, in absolute terms, whether utilitarian swerve or non-utilitarian stay AVs are the most morally appropriate vehicle.

Rather surprisingly and according to several lines of research, people's choices do not necessarily match their judgements (Lichtenstein & Slovic, 1971, 1973; Slovic & Lichtenstein, 1968). As described in Chapter 1 of this thesis, Lichtenstein and Slovic (1971) found that when offering participants to both choose and evaluate hypothetical bets, participants often judged the bet that they did not choose as the most valuable. This psychological phenomenon – known as a preference reversal - demonstrates a dissociation between the processing of choice, and the processing of judgements in decision-makers' cognitive system (Slovic & Lichtenstein, 1968). However, other authors have argued that the preference reversal phenomenon may be the result of the difference in presentations formats for judgement tasks and choice tasks (Hsee, 1996). For instance, in Lichtenstein and Slovic's (1971) experiments, choices were presented side-by-side, prompting a joint evaluation of choice options, whereas judgements were presented sequentially, prompting a separate evaluation of each option. Previous experimental findings have demonstrated that manipulating whether participants make joint or separate

evaluations of particular targets, can result in a reversal in preferences between tasks (see also Bazerman et al., 1999; Hsee et al., 1999).

According to Hsee et al.'s (1996, 1999; Hsee & Zhang, 2010) Evaluability Theory the preference reversal between separate and joint evaluations is caused by the *evaluability* of the attributes of the decision options. For example, joint evaluations are relatively easy as decision-makers are able to compare directly the values in each attribute and make a trade-off between the attributes, selecting the option superior on most attributes. In contrast, separate evaluations are difficult to make as the attribute value comparison between options are not contextually available (each separate option has a single value per attribute). Accordingly, despite a possible joint-separate evaluations difference, it is anticipated that providing full PT accessibility to choice and judgement tasks will enhance participants' utilitarian moral behaviour in both. Moreover, it is also anticipated that full PT accessibly influences the strength of the association between participants' judgements and choices; it is plausible that with full PT accessibility the association between respondents' judgements and choices is strong.

Whilst it is not the aim of this Chapter to determine the cause of preference reversals between tasks, it is important to establish whether there is a difference between moral choices and moral judgements. Therefore, Experiment 5 will measure each participants' moral judgements (separate-evaluation; as employed in all previous experiments) and moral choices (joint-evaluation). Moreover, having established that full PT accessibility results in consistent behaviour across judgement tasks (judgements of moral appropriateness, purchasing value, willingness to buy and ride AVs) it is also important to explore the influence that full PT accessibility has across different behavioural elicitation methods (moral judgements and moral choices). It is accordingly plausible that full PT accessibility will enhance utilitarian preferences (and hence behavioural consistency) across behavioural elicitation methods.

4.2.2 Method

4.2.2.1 Participants

Participants (N = 307) were recruited via PureProfile online survey panels. The sample consisted of 184 females and 123 males and the mean age was 54 (SD = 13.98). Importantly, prior to data collection, ethical approval was granted by the BSREC, and BPS ethical guidelines were followed in the treatment of all participants.

A significance level of .05 was set for statistical testing. Moreover, a retrospective power analysis was conducted on the independent-measures effect of *type of PT accessibility*. The experiment was live for 14 days to ensure that data collection from a sufficiently large sample (and a large effect size) will achieve a statistical power of at least .95. According to the retrospective power analysis, the sample size (N = 307) produced a power of 1.00 which was sufficient to achieve the target.

4.2.2.2 Experimental Design

A one-factor independent measures design was employed to measure the effect of type of PT accessibility on moral judgements and moral choices. Accordingly, moral judgements (judgements of moral appropriateness) were measured in the same way as previous experiments, where participants made separate judgements (separate-evaluations) for utilitarian swerve and non-utilitarian stay AV models on 10-point rating scales. As with the previous experiments, the utilitarian weight for moral judgements was computed by subtracting the ratings for non-utilitarian stay AV models from utilitarian swerve AV models. In contrast, moral choices, unlike moral judgements were presented in joint-evaluation format (Hsee, 1996), enabling participants to make a binary choice between a utilitarian swerve and non-utilitarian stay AV model, indicating which model is the most morally appropriate. Moreover,

to allow for choice-judgement pattern comparisons, the utilitarian weight for moral judgements was coded as 0 (non-utilitarian) and 1 (utilitarian).

4.2.2.3 Materials and Procedure

Participants were presented with a moral AV scenario and visual stimuli that consisted of either full or partial PT accessibility. Unlike previous experiments, type of involvement was kept constant and followed the logic of 'participant involvement'. Therefore, all participants received scenarios that involved themselves as the agent in the scenario. Participants were then asked to separately judge the moral appropriateness of programming AVs to swerve and stay in situations like the one described in the scenario. Following the judgement task, participants were asked to make a choice between a utilitarian-swerve AV model and a non-utilitarian stay AV model on the basis of which model was the most morally appropriate (see Appendix A and B for the exact presentation of these questions in accordance with each PT accessibility condition).

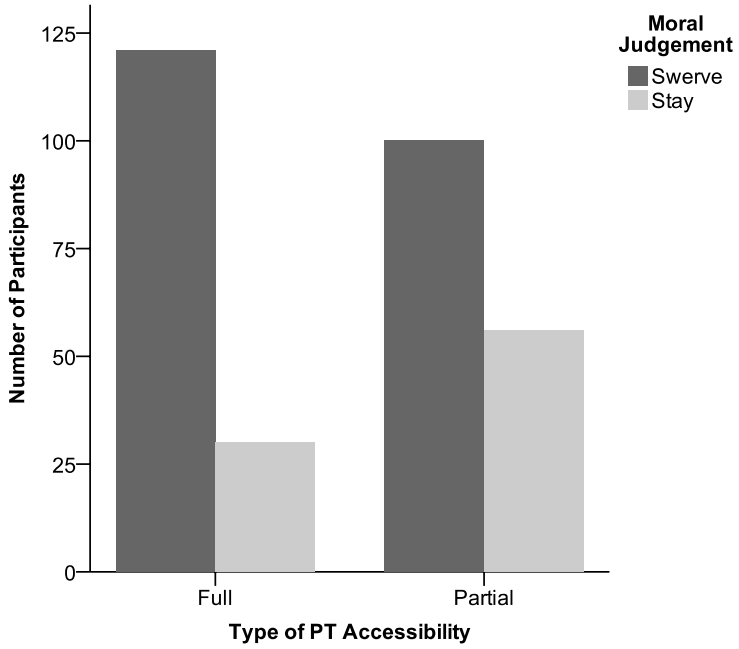
4.2.3 Results

4.2.3.1 Predicting Moral Judgements and Moral Choices

A test of the full model against a constant only model was statistically significant, indicating that the predictor reliably distinguished between utilitarian and non-utilitarian judgements $\chi^2(1) = 9.90, p = .002$. The results revealed that PT accessibility made a significant contribution to the model Wald $z = 9.56$, $p = .002$. Specifically, PT accessibility was a significant predictor, positively associated with respondents' moral judgements, odds ratio (*OR EXP[B]*) = 2.26, *CI*(.95) = (1.347; 3.786). Accordingly, these results revealed that the odds of a utilitarian judgement were 2.26 times larger when the PT accessibility was full than when the PT accessibility was partial (see Figure 15).

Figure 15

The Number of Participants Indicating their Preference for Utilitarian Swerve and Non-Utilitarian Stay AV Models as a Function of Type of PT Accessibility

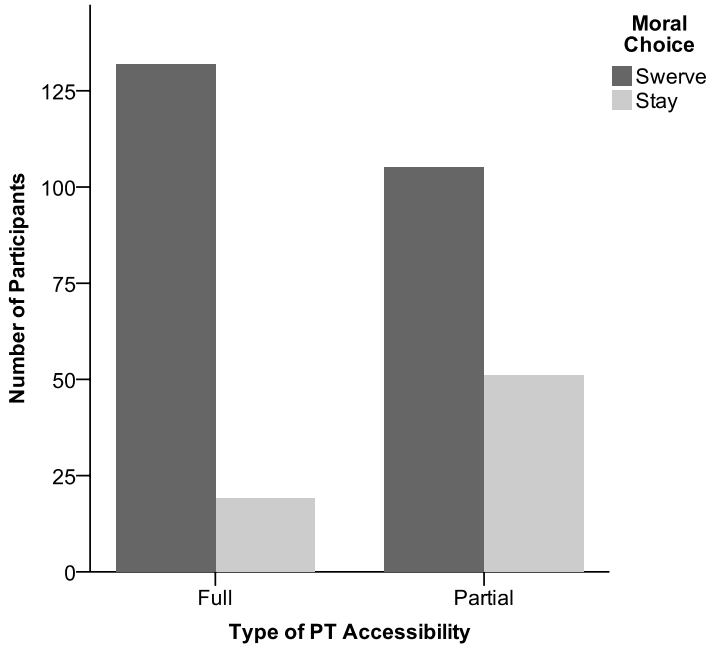


Note. Participants' preferences were calculated as a utilitarian weight of moral judgement.

Similarly, with moral choice, the test of the full logistic model against a constant only model was statistically significant $\chi^2(1) = 18.19$, $p < .001$. Accordingly, the results revealed that PT accessibility made a significant contribution to the model Wald $z = 16.56$, $p < .001$. Specifically, PT accessibility was a significant predictor, positively associated with respondents' moral choices, odds ratio (*OR EXP[B]*) = 3.37, *CI*(.95) = (1.878; 6.062). Accordingly, these results revealed that the odds of a utilitarian choice were 3.37 times larger when the PT accessibility was full (full PT accessibility) than when the PT accessibility was partial (partial PT accessibility; see Figure 16).

Figure 16

The Number of Participants Indicating their Preference (Binary Choice) for Utilitarian Swerve and Non-Utilitarian Stay AV Models as a Function of Type of PT Accessibility



4.2.3.2 Analysis of Associations: Moral Judgement and Moral Choice by PT Accessibility

Importantly, the results also show evidence for an association between the moral choices and moral judgements when the PT accessibility was full $\chi^2(1) = 56.52$, $p < .001$, and when the PT accessibility was partial $\chi^2(1) = 7.49$, $p = .008$. Crucially, the strength of the association between moral choice and judgement was strong with full PT accessibility ($\Phi = .612$, $p < .001$) and weak with partial PT accessibility ($\Phi = .219$, $p = .008$). In other words, utilitarian judgements were positively associated with utilitarian choices even more so when PT accessibility was full, indicating that PT accessibility is a fundamental decision-making phenomenon which informs respondents' preferences across different behavioural elicitation methods and tasks. The positive association between participant's utilitarian judgements and utilitarian choices was stronger when PT accessibility was full, indicating that PT accessibility

is an underlying decision-making strategy that respondents employ across behavioural elicitation methods.

4.2.4 Discussion

According to many lines of decision-making research, people's preferences are highly dependent on the behavioural elicitation method employed (Kusev et al., 2020; Pedroni et al., 2017). One example of this has been demonstrated by the preference reversal phenomenon; where people's judgements do not reflect the choices they have made (Lichtenstein & Slovic, 1971, 1973). Moreover, theorists have argued that in comparison to separate evaluations, joint evaluations are easier to make (Hsee et al., 1996, 1999; Hsee & Zhang, 2010). This *evaluability* effect invites opportunities for preference reversals as the attributes are treated differently when their values are available for direct comparisons (e.g., in choice task) and when they are not (e.g., in judgement task). Experiment 5, explored participants' utilitarian moral behaviour; in particular how full PT accessibility impacts respondents' judgements and choices (despite a possible joint-separate evaluations difference). Accordingly, the goal of this study was to explore (i) whether full PT accessibility influence both respondents' judgements and choice, (ii) whether full PT accessibility determines the strength of the association between participants' judgements and choice.

The preference reversal pattern of behaviour was not exhibited in any of the experimental conditions in this experiment. This could be explained by differences in the nature of moral decision-making tasks compared to hypothetical risky gambles (using probability and money). For instance, it is plausible that decision-makers treat human life utility very differently from monetary utilities (Gold et al., 2013). Importantly, the results revealed that the likelihood of moral utilitarian behaviour increased substantially for both choice and judgement tasks when participants were given tasks with full PT accessibility.

Furthermore, when applied to a moral context, as in Experiment 5, judgements and choices related to the moral appropriateness of utilitarian and non-utilitarian models AVs did correlate but only very weakly. However, this was the case with partial PT accessibility (where participants received crash scenarios that involved taking the perspective of the AV passenger). In contrast, when participants received crash scenarios with full PT accessibility (where they were offered the perspective of the AV passenger and the perspective of the pedestrians) the association between moral judgements and moral choices was moderate to strong. Specifically, there was a strong positive association between moral utilitarian judgements and moral utilitarian choices when PT accessibility was full. This result is important as it provides evidence that full PT accessibility is a fundamental decision-making phenomenon which informs respondents' preferences across different behavioural elicitation methods and tasks.

These findings together indicate that offering full PT accessibility does not only improve prosocial utilitarian behavioural consistencies across judgement tasks (as in Experiments 2-4), but also across behavioural elicitation methods. Moreover, when considered separately, respondents' utilitarian moral choices and moral judgements increased in full PT accessibility conditions when compared to partial PT accessibility conditions. That is, people who received moral crash scenario's containing full PT accessibility were more likely to morally approve of utilitarian AVs over non-utilitarian AVs, as highlighted by their moral judgements and moral choices. These findings therefore support the notion that full PT accessibility leads to consistent prosocial utilitarian preferences, where people morally approve of utilitarian AVs regardless of the preference elicitation method.

4.3 Experiment 6: How Perspective-Taking Behaviour Informs Moral Judgements

4.3.1 Introduction

Researchers using methodological variations of the free-choice paradigm commonly explore the influence that a difficult choice has on participants subsequent judgements (Brehm,

1956). Unlike Lichtenstein and Slovic's (1971) preference reversal phenomenon, which reveals a psychological dissociation between choice and judgements, Brehm's (1956) free-choice paradigm demonstrates that people's binary choices can sometimes inform their subsequent judgements. Specifically, once people commit to a difficult decision (choose an option from a set of two which are similarly judged/evaluated) they tend to value and appreciate this option more than before; in other words, evidence for choice-induced change in preferences (e.g., Brehm, 1956; Festinger, 1964; Izuma et al., 2010; Sharot et al., 2012). In this experimental paradigm, first, participants are required to rate several items (pre-choice judgement task). Participants are then asked to choose between 2 of the items they rated similarly in the judgement task (in many studies they are led to believe they will keep their chosen item). Participants are then required to repeat the first judgement task (post-choice judgement task). Brehm (1956) found that chosen items are given higher ratings in the post-choice judgement task than the pre-choice judgement task and accordingly, rejected items are given poorer ratings in the post-choice judgement task than the pre-choice judgement task. Accordingly, theorists have argued and debated over the years as to why this effect occurs. One explanation comes from the cognitive dissonance theory (e.g., Festinger, 1957); making difficult decisions induces cognitive dissonance, which is supressed/reduced by the participants with re-evaluation of their judgements after the decision is made. Interestingly, this proposal is suggesting that there are not any stable and available preferences that guide human judgements (e.g., Brehm, 1956; Festinger, 1957; Gerard & White, 1983; Sharot et al., 2012).

However, other researchers have argued that this choice-induced preference change is independent from cognitive dissonance and happens even when the experimental method does not permit for cognitive dissonance to appear; for example, when participants make two judgements and then a choice (Chen & Risen, 2010; Izuma et al., 2010; Izuma & Murayama, 2013; Sharot et al., 2012). The authors have argued that the post-choice changes in judgements

were simply evidence for pre-existing (prior to the choice) *true* preferences rather than choice induced change in judgement (Chen & Risen, 2010; Izuma et al., 2010; Izuma & Murayama, 2013). They also argued that true preferences are most likely to be captured by the choice that people make (decision measures) and that participants' pre-choice judgements are noisy and their second (or post-choice) judgements are less noisy (and more likely to represent true preferences), which according to the authors is the actual reason for the change in judgements. In contrast to the cognitive dissonance explanation, this proposal suggests that there are stable and available preferences that guide human judgements.

These two opposing views regarding the relationship between human choices and judgements are interesting and important, but they are not the only plausible accounts. As it is evident, in the results of this dissertation project, there is also a full PT accessibility explanation. True preferences are not always psychologically accessible or may not even exist (e.g., Kusev et al., 2020); human preferences accordingly require full PT accessibility. For example, the experimental results in this thesis prove that when participants read AV crash scenarios with partial PT accessibility, their judgements and decisions are dissociated and do not represent their true/actual moral preferences. In contrast, full PT accessibility eliminates behavioural inconsistency; making moral judgements with full PT accessibility informs participants' moral behaviour (see Experiments 2, 3, 4 and 5). Moreover, with partial PT accessibility the association between moral judgements and moral behaviour is weak or non-existent, and strong when participants are offered full PT accessibility. In Experiments 2, 3, 4, I found evidence for full PT accessibility judgement-induced behavioural change (participants judgements of moral appropriateness informed their purchasing behaviour). Accordingly, in Experiment 6, I will explore the possibility for full PT accessibility choice-induced judgement change. It is plausible that with full PT accessibility, the expected choice-induced judgement change will be stronger than the choice-induced judgement change itself, and irrespective of

whether the decisions are difficult or easy. In other words, I argue for full PT accessibility transfer rather than simply a choice-induced judgement change (or judgement-choice induced change) or a change motivated by judgement noise, as predicted by previous accounts (Chen & Risen, 2010; Izuma et al., 2010; Izuma & Murayama, 2013; Sharot et al, 2012).

4.3.2 Method

4.3.2.1 Participants

The participants for Experiment 6 consisted of 340 (194 females and 146 males) users of PureProfile's survey panel. The mean age of the participants was 48 (*SD* = 13.99). Prior to data collection, ethical approval was granted by the BSREC, and accordingly, all participants were treated in accordance with BPS ethical guidelines. A retrospective power analysis was conducted on the independent-measures effects of *type of PT accessibility* and *dilemma difficulty*. The experiment was live for 14 days to ensure that data collection from a sufficiently large sample (and a large effect size) will achieve a statistical power of at least .95. According to the retrospective power analysis, the sample size (*N* = 340) produced a power of 1.00 which was sufficient to achieve the target.

4.3.2.2 Experimental Design

A 2x2 independent measures design was employed. The first independent variable, type of PT accessibility had two levels (full and partial) and the second independent variable, dilemma difficulty also had two levels (easy or difficult). It is important to note that PT accessibility was only manipulated in stage 2 of the experiment (see materials and methods for more details). The dependent measure of this experiment was judgements of moral appropriateness which was measured for each AV model (utilitarian swerve and non-utilitarian stay) on separate 10-point rating scales. Accordingly, as with all of the previous experiments in this thesis, these separate ratings were combined (by subtracting judgements for non-

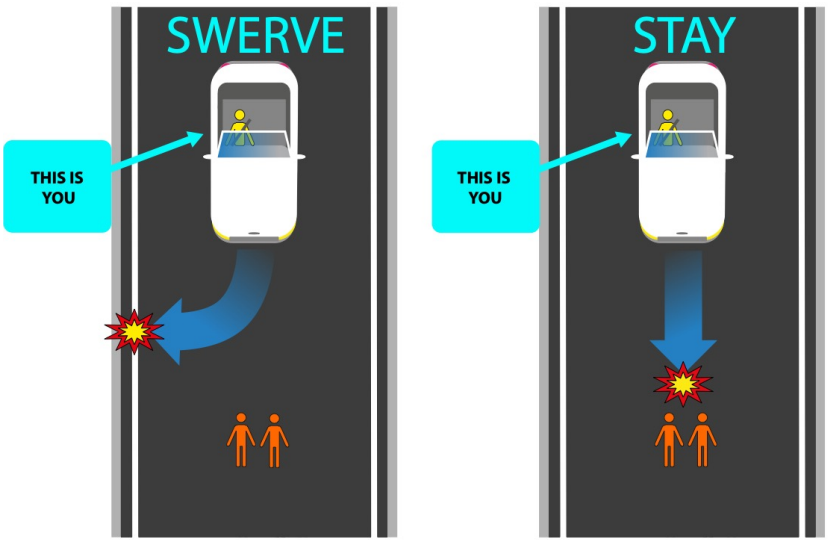
utilitarian stay AVs from utilitarian swerve AVs) in order to achieve a utilitarian weight. Importantly, type of accessibility was only manipulated during the choice stage of the experiment (see materials and procedure for more details).

4.3.2.3 Materials and Procedure

Participants were first allocated to either an easy or difficult dilemma condition. The difficulty of a dilemma was characterised by the utility ratios between the passenger and pedestrian. For example, an easy decision would be a standard 1 passenger vs. 10 pedestrians' trade-off (as employed across all of the experiments in this thesis), whereas a difficult decision involved a trade-off between 1 passenger and 2 pedestrians (see Figure 17). Previous research has demonstrated that smaller utility ratios induce non-utilitarian choices, suggesting that smaller utility ratios render the dilemma more difficult to solve (Faulhaber et al., 2019; Martin & Kusev, 2016; Nakamura, 2012). Importantly the dilemma difficulty (easy or difficult) in which the participant was assigned to remained the same throughout all stages of the experiment.

Figure 17

Visual Stimuli that Depicts a Difficult Dilemma

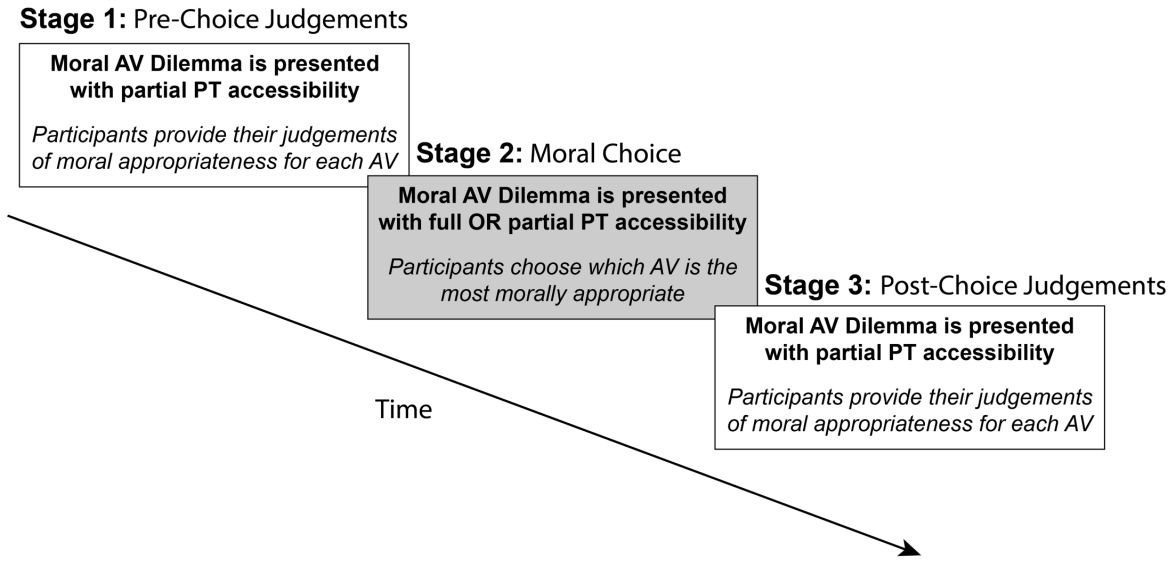


Note. This example also contains partial PT accessibility.

The experiment involved 3 stages as depicted in Figure 18. In the first stage (pre-choice judgements) participants received an AV scenario and were required to make separate judgements of moral appropriateness for each AV model on two 10-point scales. In the second stage (moral choice) participants were then given the same scenario but this time with either full or partial PT accessibility. Participants then made a binary choice between each AV in the in order to indicate the 'most' morally appropriate AV model. In Stage 3, participants repeated the same task as in stage 1, once again receiving an AV scenario that contained partial PT accessibility.

Figure 18

The 3-Stage Experimental Procedure for Experiment 6



Note. See Appendix A and B for full experimental materials.

4.3.3 Results

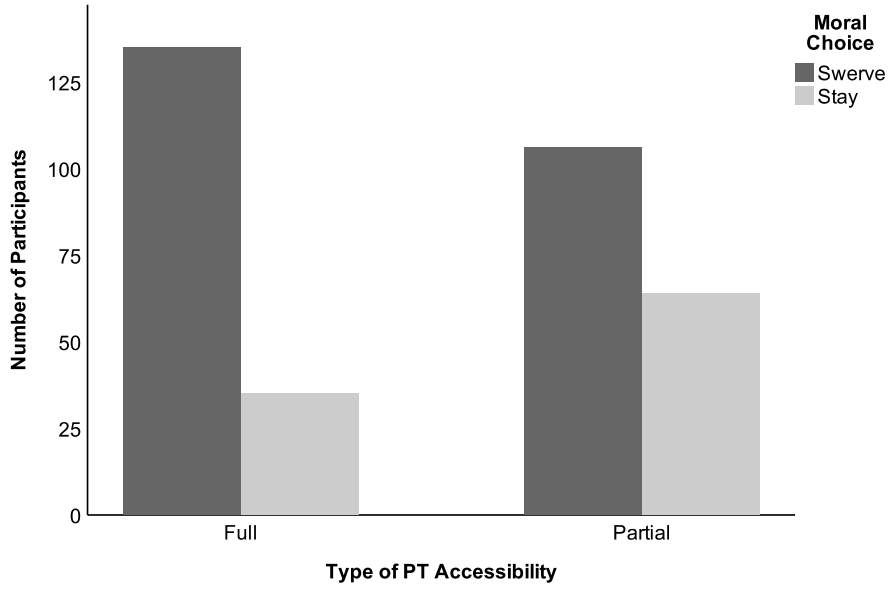
4.3.3.1 Predicting Moral Choices.

As in Experiment 5 utilitarian moral choices (rational utilitarian behaviour to swerve and save the majority) were more commonly made when the PT accessibility was full (see Figure 19) than when the PT accessibility was partial. A binary logistic regression analysis was

conducted to predict moral utilitarian behaviour (choice) using type of PT accessibility (full or partial) and dilemma difficulty (easy or difficult) as predictors. A test of the full model against a constant only model was statistically significant, revealing that the predictors as a set reliably distinguished between utilitarian and non-utilitarian choices, $\chi^2(3) = 13.121$, $p = .004$.

Figure 19

The Number of Participants Indicating their Preference (binary choice) for Utilitarian Swerve and Non-Utilitarian Stay AV Models as a Function of Type of PT Accessibility



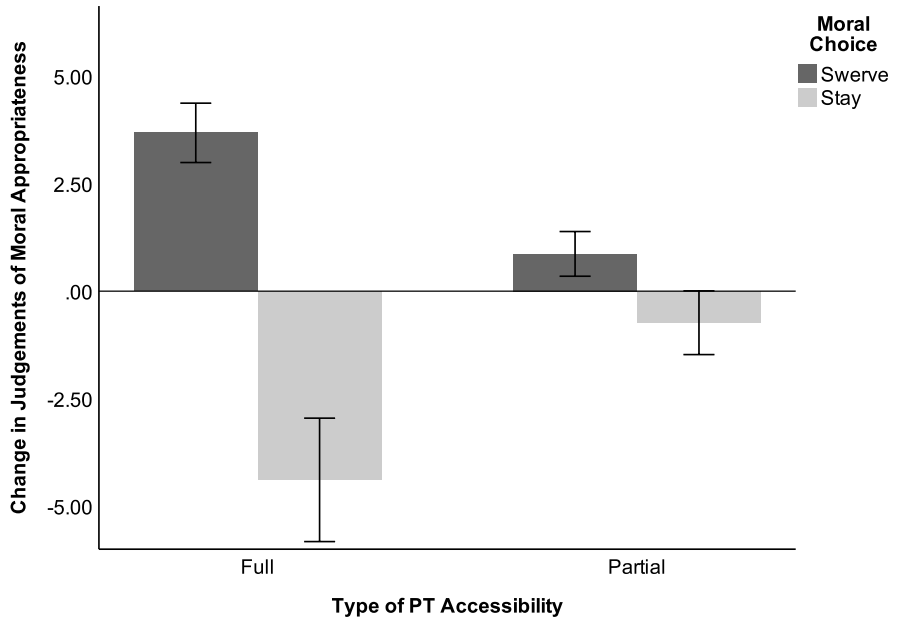
The results revealed that PT accessibility made a significant contribution to the model Wald $z = 7.37$, $p = .007$. Specifically, PT accessibility was a significant predictor, positively associated with respondents' moral choices, odds ratio (*OR EXP[B]* = 2.68, *CI*(.95) = (1.315; 5.456). Accordingly, these results revealed that the odds of a utilitarian choice were 2.68 times larger when the PT accessibility was full (full PT accessibility) than when the PT accessibility was partial (partial PT accessibility). However, neither the predictor dilemma difficulty, odds ratio (*OR EXP[B]*) = .91, *CI*(.95) = (.486; 1.683), *p*= .752, nor the interaction PT accessibility by dilemma difficulty odds ratio (*OR EXP[B]*) = .77, *CI*(.95) = (.291; 2.038), *p*= .598, made a significant contribution to the regression model.

4.3.3.2 Utilitarian Change of Judgements of Moral Appropriateness.

A 2x2x2 independent measures analysis of variance was conducted to explore the influence of the independent variables (type of PT accessibility, dilemma difficulty and choice made – utilitarian or non-utilitarian) on respondents' change in judgements of moral appropriateness. The results revealed that respondents' choices (utilitarian or non-utilitarian), $F(1, 332) = 127.56$, $p < .001$, $\eta_p^2 = .28$, as well as the interaction choice by type of PT accessibility, $F(1, 332) = 57.89$, $p < .001$, $\eta_p^2 = .15$, significantly influenced respondents' change in judgements of moral appropriateness (see Figure 20). However, the results revealed that main effects of type of PT accessibility ($F[1, 332] = 1.29$, $p = .258$), dilemma difficulty ($F[1, 332] = 2.09$, $p = .149$), and the two-way interactions of type of PT accessibility by dilemma difficulty ($F[1, 332] = 1.55$, $p = .214$), dilemma difficulty by choice ($F[1, 332] = 1.34$, $p = .247$), as well as the three-way interaction type of PT accessibility by dilemma difficulty by choice ($F[1, 332] = 1.25$, $p = .265$) on respondents' change in judgements of moral appropriateness were not statistically significant.

Figure 20

Change in Participants Judgements of Moral Appropriateness



Note. Positive values indicate utilitarian change; negative values indicate non-utilitarian change. Error bars represent 95% confidence intervals of the mean.

Due to the significant two-way interaction, follow-up simple-effect tests were conducted by type of PT accessibility.

Partial PT Accessibility. A follow-up simple-effect test revealed that with partial PT accessibility, the main effect of choice $F(1, 166) = 12.93$, $p < .001$, $\eta_p^2 = .07$ significantly influenced respondents' change in judgements of moral appropriateness. Specifically, the results revealed that the effect of choice was significant, with non-utilitarian choices leading to a small non-utilitarian change in judgements of moral appropriateness ($M = -.74$; $SD = 2.96$) and utilitarian choices leading to a small utilitarian change in judgements of moral appropriateness ($M = .87$; $SD = 2.70$). However, the main effect of dilemma difficulty ($F<1$), as well as the interaction dilemma difficulty by choice ($F<1$) on respondents' change in judgements of moral appropriateness were not statistically significant.

Full PT Accessibility. A follow-up simple-effect test revealed that with full PT accessibility, the main effect of choice $F(1, 166) = 111.19$, $p < .001$, $\eta_p^2 = .40$, significantly influenced respondents' change in judgements of moral appropriateness. Specifically, the results revealed that the effect of choice was significant, with non-utilitarian choices leading to a substantial non-utilitarian change in judgements of moral appropriateness ($M = -4.39$; $SD = 4.18$) and utilitarian choices leading to a substantial utilitarian change in judgements of moral appropriateness ($M = 3.68$; $SD = 4.06$). However, the main effect of dilemma difficulty ($F[1, 166] = 2.25$, $p = .135$), as well as the interaction dilemma difficulty by choice ($F[1, 166] = 1.61$, $p = .206$) on respondents' change in judgements of moral appropriateness were not statistically significant.

4.3.4 Discussion

Over the years, researchers using Brehm's (1956) free-choice paradigm revealed evidence that participants' choices inform their subsequent judgements (see for review Izuma et al., 2013). Specifically, once people make a difficult choice they tend to value and appreciate their chosen option more than they used to. This effect provides evidence for choice-induced change in preferences (e.g., Brehm, 1956; Festinger, 1964; Izuma et al., 2010; Sharot et al., 2012). There are two dominant theoretical accounts that provide an explanation for this interesting phenomenon. According to the *dissonance* explanation (e.g., Festinger, 1957), participants making difficult choices experience cognitive dissonance, which is reduced by a shift in post-choice preferences (informed by the preference expressed in the difficult choice). In contrast, Chen and Risen (2010) and Izuma et al. (2010) argued that this choice-induced preference is instead an artefact of the method used in the free-choice paradigm. They provided experimental evidence for choice-induced preference change even when the cognitive dissonance explanation is methodologically impossible. They explained the original free-choice results by suggesting that participants' pre-choice judgements are noisy and their second (or post-choice) judgements are less noisy, and more likely to represent true preferences. This proposal implies that, in contrast to the cognitive dissonance explanation, there are true, stable and available preferences that guide human judgements and choices.

In Chapter 3 (Experiments 2, 3, 4), I found evidence for full PT accessibility judgement-induced behavioural change. In Experiment 6, I have further explored the possibility for full PT accessibility choice-induced judgement change. In contrast to previous free-choice results and arguments, I proposed that (i) with full PT accessibility, the expected choice-induced judgement change will be stronger than the choice-induced judgement change itself, (ii) full PT accessibility transfer will take place irrespective of whether the choices are difficult or easy. Moreover, I have argued for full PT accessibility transfer rather than simply a choice-induced

judgement change (or judgement-choice induced change) or a change motivated by judgement noise (Chen & Risen, 2010; Izuma et al., 2010; Izuma & Murayama, 2013; Sharot et al, 2012).

Similar to the results in Experiment 5, the results in Experiment 6 confirmed that the likelihood of making utilitarian moral choice was higher when PT accessibility was full. However, the dilemma difficulty did not predict participants utilitarian choice. The main results established that when respondents made choices informed by full PT accessibility, their judgements of moral appropriateness (in AV crash scenarios with partial PT accessibility) changed in the direction of the choices they made (utilitarian or non-utilitarian). It is important to note that with full PT accessibility, participants making non-utilitarian decisions were the minority (as confirmed in Experiments, 2, 3, 4 and 5). This result provides evidence for full PT accessibility transfer from choice (when participants received crash scenarios with full PT accessibility) to judgements (when participants received crash scenarios with partial PT accessibility). Importantly, and in contrast to predictions from free-choice researchers, this full PT accessibility transfer takes place irrespective of whether choices were difficult or easy. Moreover, this effect was large and superior to the effect of choice-induced change in preferences (which had a small effect size) and the nonsignificant effect of dilemma difficulty (easy or difficult decisions).

4.4 General Discussion

People's choices are dissociated from their judgements. Despite the numerous explanations for this dissociation, many researchers agree with this statement and have established this phenomenon in risky and non-risky decision-making tasks (see Hsee et al., 1996, 1999; Hsee & Zhang, 2010; Lichtenstein & Slovic, 1971; Slovic & Lichtenstein, 1968). However, in moral decision-making tasks people's judgements are associated (positively so) albeit very weakly. Experiment 5 exemplified this association, utilising the AV dilemma as a hypothetical moral scenario - based on which, participants made their moral judgements and

moral choices. In particular, when participants received AV crash scenarios with partial PT accessibility, their choices and judgements were indeed very weakly associated. However, when participants received AV crash scenarios with full PT accessibility, both their moral judgements and moral choices were moderate-strongly associated (and utilitarian). In other words, with full PT accessibility, the majority of participants morally approved of the utilitarian AV over non-utilitarian AV and this was observed in both their judgements and choices. This pattern of behaviour importantly builds on a similar finding from Chapter 3 (Experiments 3 and 4) which demonstrates that presenting participants with crash scenarios containing full PT accessibility results in consistent utilitarian behaviour across different types of judgements tasks (e.g., judgements of moral appropriateness, purchasing value, willingness to buy and willingness to ride). Moreover, despite predictions that judgements and choices are psychologically dissociated (Lichtenstein & Slovic, 1971), Experiment 5 demonstrates for the first time that offering AV crash scenarios with full PT accessibility to participants also results in consistent moral preferences across preference elicitation methods. Therefore, people require full accessibility to PT information in order to be both consistent across their moral judgements and moral choices. The novel findings from Experiment 5 demonstrate that PT accessibility is general psychological phenomenon that can inform preferences across different judgements tasks and behavioural elicitation methods (e.g., both judgements and choices).

Having established an association between people's moral judgements and choices under conditions of full PT accessibility, I explored further whether people's moral choices could go as far as to inform their moral judgements. This was taken from Brehm's (1956) free-choice paradigm, which in contrast to preference reversals (Lichtenstein & Slovic, 1971, 1973; Slovic & Lichtenstein, 1968) demonstrates that people's choices can sometimes inform their subsequent judgements. Utilising a modified version of the free-choice paradigm, under conditions of partial PT accessibility, I replicated Brehm's (1956) effect; people's post-choice

moral judgements were informed by the choices they made and were slightly different from their original pre-choice judgements (small effect size). However, in conditions of full PT accessibility (where participants made their pre- and post-choice judgements with partial PT accessibility but their choices with full PT accessibility) this effect of choice-induced judgement change was large. In particular, the results established that when respondents made choices informed by full PT accessibility, their judgements of moral appropriateness (in AV crash scenarios with partial PT accessibility) changed in the direction of the choice they made. In other words, participants transferred their experience with a full PT to a task without full PT accessibility and used their previous experience with full PT to inform their moral judgements. This result therefore provides evidence for a full PT accessibility transfer, from the choice task to the judgement task. Importantly, and in contrast to predictions from free-choice researchers (Brehm, 1956; Sharot et al., 2012), this full PT accessibility transfer takes place irrespective of whether decisions are difficult or easy. Moreover, this effect was large and superior compared to the effect of choice-induced change in preferences (which had a small effect size) and the nonsignificant effect of dilemma difficulty (easy or difficult decisions). Therefore, Experiment 6 also builds on Chapter 3's Experiments 3-4 where one of the main findings demonstrated that with full PT accessibility, participants' judgements of moral appropriateness informed their purchasing and usage behaviours (full PT accessibility judgement-induced behavioural change). However, in Experiment 6, I have found that with full PT accessibility, participants' utilitarian moral choices inform their utilitarian moral judgements judgement-induced behavioural change (full PT accessibility choice-induced judgement change).

Taken together these findings demonstrate that full PT accessibility is a fundamental psychological phenomenon that results in consistent utilitarian preferences across behavioural elicitation methods. With full PT accessibility, people are not only utilitarian in their judgements and choices, but these judgements and choices and strongly associated. Moreover,

full PT strategies can be transferred from choice tasks (where full PT accessibility is available) to subsequent judgements tasks (where only partial PT accessibility is available). Therefore, once exposed to full PT accessibility, participants can employ this strategy in subsequent tasks.

CHAPTER 5

Perspective-Taking Accessibility: Conscious and Unconscious Thinking

5.1 Overview of Chapter 5

In Chapter 5, I introduce and explore Unconscious Thought Theory (Dijsterhuis & Nordgren, 2006) where experimental evidence suggests that people make better decisions after they have processed complex decision-making information unconsciously as opposed to consciously. Consequently, this Chapter comprises of three Experiments that test the influence of three types of processing (immediate, conscious and unconscious) on judgements of moral appropriateness. Moreover, the three experiments in this chapter also test the influence of variations in PT accessibility on judgements of moral appropriateness as well as the interaction between PT accessibility by type of psychological processing and hence their combined influence on people's moral judgements.

Experiment 7 serves as an extended replication of Experiment 2, where judgements of moral appropriateness is the single dependent variable. Moreover, Experiment 8 is an extended replication of Experiment 3, which measures judgements of moral appropriateness and purchasing value. Finally, Experiment 9 is an extended replication of Experiment 4, which measures judgements of moral appropriateness and willingness to buy AVs.

All experiments revealed that when scenarios were presented with partial PT accessibility, participants were more utilitarian in their moral judgements if they had processed the information unconsciously as opposed to immediately or consciously. However, participants presented with full PT accessibility were more utilitarian in their judgements of moral appropriateness regardless of the type of processing employed. These findings demonstrate that when information is presented in an unbiased and fully accessible way, the way in which information is processed does not influence participants' judgements of moral appropriateness.

5.2 Experiment 7: The Influence of Perspective-Taking Accessibility on Conscious and Unconscious Moral Judgements

5.2.1 Introduction

In real life situations as well as in experimental settings, it is assumed that people consciously deliberate about choice options prior to making final decisions. Deliberating consciously involves thinking about task relevant goals whilst attending to the task itself. It is generally accepted that careful deliberations made during conscious thinking will result in people making good decisions. Whilst this remains true for simple decision problems, according to Unconscious Thought Theory (hereafter, UTT; Dijksterhuis & Nordgren, 2006), complex decision problems should be left to unconscious processing if one wants to make the optimal choice. Unlike conscious processing, unconscious processing involves the processing of information beyond conscious awareness, such as when an individual is distracted with an unrelated task.

In order to exemplify the advantage of unconscious thinking of complex decision-problems, UTT theorists have developed experimental paradigm that allows unconscious thinking to be directly manipulated. In the UTT experimental paradigm (see for example, Dijksterhuis et al., 2006), participants are presented with multi-attribute choice options (for example, cars, each with an array of positive and negative characteristics). Accordingly, a good choice would be a car with many positive attributes and few negative attributes, whereas a bad choice would be the reverse. However, some participants receive simple decision problems (4 attributes per car) whilst others receive complex decisions problems (12 attributes per car). Participants are then separated into one of two type of processing conditions: (i) conscious thought (where participants are instructed to think carefully about the options for 4 minutes) and (ii) unconscious thought (where participants are distracted with an anagram task for 4 minutes). The purpose of the distraction task is to induce unconscious processing, by directing

the participants' attention away from the task itself. After the conscious thought/distraction period, participants are then asked to make their choice. An overwhelming number of studies demonstrate the advantage of unconscious thinking when engaging in complex decisions (see Dijksterhuis & Strick, 2016 for a review). That is, participants are more likely to choose options with a large number of favourable attributes and a low number of unfavourable attributes after unconscious deliberation. Moreover, these findings are consistent also across preference elicitation methods (both decision-making and judgement tasks; Dijksterhuis et al., 2006). Based on numerous empirical findings, UTT theorists have therefore developed 6 principles of UTT which outline the distinctions between conscious and unconscious thinking and processing (see Table 2; see also Dijksterhuis & Nordgren, 2006).

Table 2

The 6 Principles of Unconscious Thought Theory

Principle Title	**Principle Description**
Unconscious-Thought Principle	There are two modes of thought: conscious thought (or conscious processing) and unconscious thought (unconscious processing) each with unique characteristics.
Capacity Principle	Conscious processing is constrained by limits to processing capacity whereas unconscious processing is not.
Bottom-Up-Versus-Top-Down-Principle	Unconscious processing relies on bottom-up processing (slowly integrating information to form an objective summary judgement), whereas conscious processing relies on top-down processing (basing judgements on pre-existing stereotypes and schemas).
Weighting Principle	Unconscious processing often results in better weighting of positive/negative attributes of choice options than conscious processing.
Rule Principle	Conscious processing can follow rules (such as rules required to solve mathematical problems) whilst unconscious processing cannot.
The Convergence-Versus-Divergence Principle	Conscious processing is convergent (linear), whereas unconscious processing is divergent (examines multiple ways to deal with a problem).

It is clear from Table 2 that conscious and unconscious thinking involve distinct strategies for processing decision information, which may account for why these two types of processing result in contrasting decision outcomes. As mentioned in Chapter 1, dual process theorists (e.g., Greene et al., 2001; Stanovich & West) also pertain that decision outcomes are highly dependent on how decision-making information is processed. However, contrary to this viewpoint, Experiments 1-6 have consistently demonstrated that variations in accessibility to dilemma information (e.g., contextual accessibility or PT accessibility) result in variations in decision-making behaviour. For instance, full PT accessibility eliminates many behavioural inconsistencies between judgements tasks, types of involvement and behavioural elicitation methods (Experiments 2-6). Accordingly, it is plausible that when people are provided with fully accessible information (e.g., full PT accessibility), they will make consistent decisions regardless of the type of psychological processing (e.g., conscious or unconscious) undertaken. Therefore, the current experiment aims to investigate the influence of PT accessibility and type of processing on judgements of moral appropriateness towards AVs.

Importantly, UTT has been previously applied in moral decision-making tasks where a 'good' choice is determined according to normative utilitarian standards. For example, Ham and van den Bos (2010) applied UTT to a moral decision-making context, offering the footbridge dilemma as the decision problem. The authors found that after unconscious processing of a complex version of the footbridge dilemma, people were more likely to morally approve of pushing 1 man to his death in order to save 5 (the utilitarian choice). Given these findings, it is expected that when applying the UTT paradigm to AV crash scenarios, unconscious processing will result in a greater approval of utilitarian-swerve AVs over non-utilitarian stay AVs. However, it is anticipated that this will only be the case in partial PT accessibility conditions. Conversely, it is expected that under conditions of full PT accessibility, people will approve of utilitarian swerve AVs over non-utilitarian stay AVs

across all types of processing conditions (conscious processing, unconscious processing and immediate judgement).

5.2.2 Method

5.2.2.1 Participants

Participants (N = 360) were recruited to take part in an online computer-based experiment through PureProfile's online survey panels. The sample consisted of 190 females and 170 males. The mean age of the participants was 47 (SD = 13.59). Prior to data collection, ethical approval was obtained from the BSREC. Moreover, all participants were treated in accordance with BPS ethical guidelines.

For statistical testing, a significance level of .05 was set. Moreover, a retrospective power analysis was conducted on the independent-measures effects of *type of PT accessibility* and *type of psychological processing* and their interaction. The experiment was available online for 14 days to ensure that data collection will achieve a statistical power of at least .95. According to the retrospective power analysis, the sample size (N = 360) produced a power of 1.00 which was sufficient to achieve the target.

5.2.2.2 Experimental Design

A 2 (type of PT accessibility) x 3 (type of psychological processing) independent measures design was employed. The first independent variable, type of PT accessibility, had two levels (full and partial) and was manipulated using the same logic as previous experiments. The second independent variable, type of psychological processing, had 3 levels including immediate judgements (where participants were instructed to make moral judgements immediately after reading the moral AV scenario), conscious processing (where participants were instructed to consciously consider the moral AV scenario for 3 minutes) and unconscious processing (where participants completed a distraction task for 3 minutes in order to induce

unconscious processing of the moral AV scenario). Importantly, type of involvement was kept constant across all conditions and followed the logic of *participants involvement*. Therefore, all participants were required to imagine themselves in the crash scenarios.

The experiment contained one dependent measure (judgements of moral appropriateness) where, identical to the previous experiments in this thesis, participants rated each AV model (utilitarian swerve and non-utilitarian stay) on separate 10-point rating scales, from which the utilitarian weight was calculated.

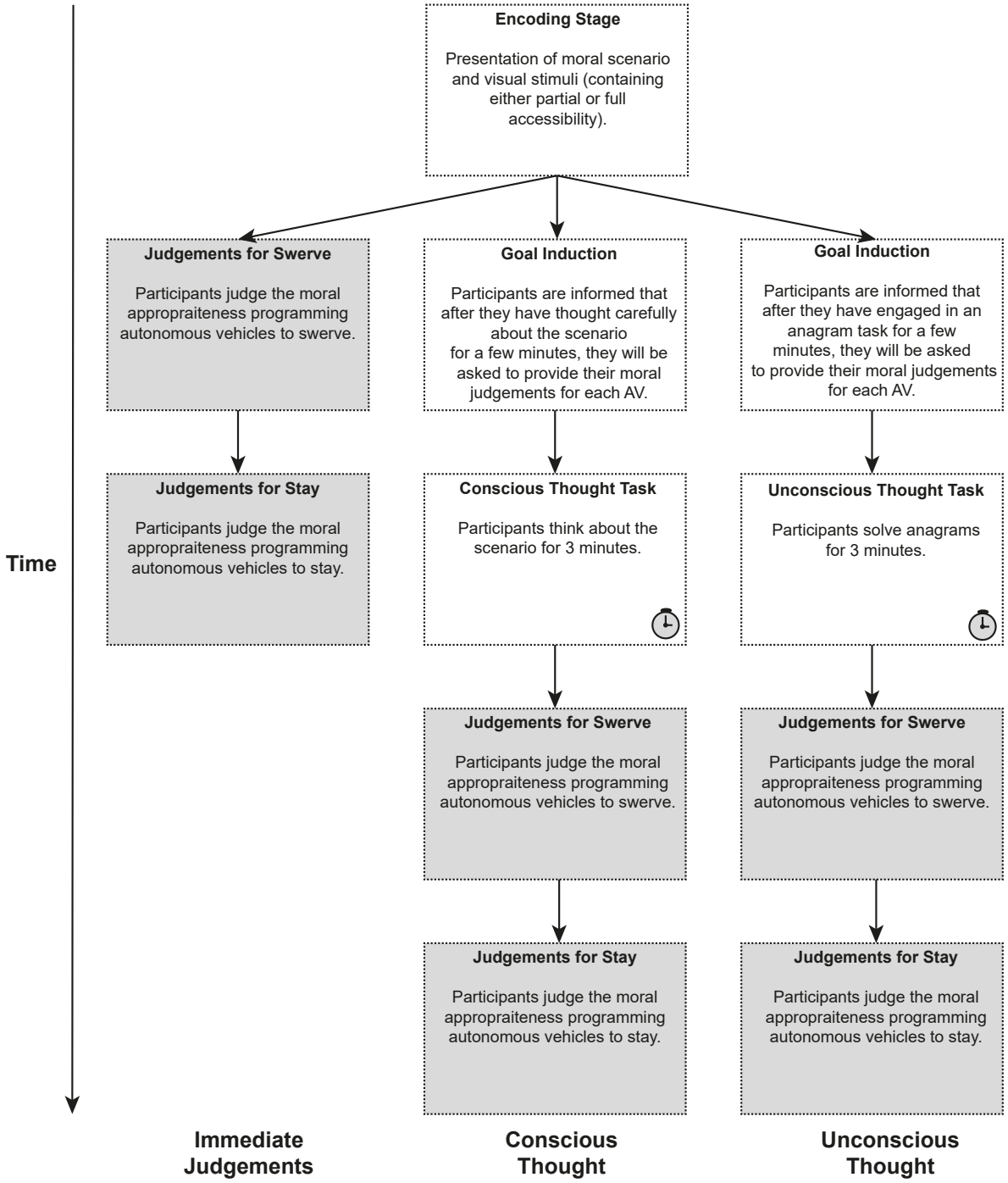
5.2.2.3 Materials and Procedure

All participants were presented with an AV crash scenario and visual stimuli that contained either full or partial PT accessibility (depending on the type of PT accessibility condition they were assigned to). Participants were then assigned to one of 3 types of psychological processing conditions: immediate judgements, conscious processing, and unconscious processing (see Figure 21). In the immediate judgements condition, participants made judgements of moral appropriateness for each AV model immediately after reading the moral AV dilemma. In the conscious processing condition, participants were prompted to think carefully about it for 3 minutes before making their final moral judgements. Specifically, participants were told that they will be asked about their judgements regarding the moral appropriateness of each AV model and should therefore think about this in particular. In the unconscious processing condition, participants were also told that they will need to provide judgements regarding the moral appropriateness of each AV later but will first be required to complete an anagram task. Participants in the unconscious condition were subsequently distracted with an anagram task for 3 minutes to prevent them from thinking consciously about the scenario (see Appendix A for the anagrams employed in this experiment). Previous research in unconscious processing has successfully employed anagram tasks in order to induce unconscious processing in participants (see for example, Dijksterhuis et al., 2006; Strick et al.,

2010, 2011). Accordingly, solving anagram tasks provide enough distraction to prevent participants from consciously processing the main decision task and therefore serves as an ideal distraction task for inducing unconscious processing (Acker, 2008).

Figure 21

The Procedure for all Experimental Conditions in Experiment 7



Note. Adapted from "A case for thinking without consciousness" by A. Dijsterhuis, and M. Strick, 2016, *Perspective on Psychological Science, 11*(1), p. 122. Each box represents one page on the computer screen, therefore participants had to press a button to proceed to the next page in all cases except for the pages that had a time limit (timed pages proceeded to the next page automatically after the timer ran out).

After the 3 minutes surpassed for the participants in the conscious and unconscious condition, they were then automatically redirected to a separate webpage where they could provide their moral judgements. Importantly, in all conditions in Experiment 7, judgements of moral were made in the absence of the moral scenario and visual stimuli. Therefore, participants had to recall scenario information. In order to make sure that participants in all conditions were correctly remembering scenario details, a manipulation check was included at the end of the experiment which asked participants the following (for full experimental materials see Appendix A and B):

In the scenario you read:

1. How many people were **inside** the car?
2. How many people were **outside** of the car?

5.2.3 Results

All 360 participants included in the analysis passed the manipulation check. Therefore, any differences in utilitarian judgements across conditions cannot by accounted for by a decline in memory for utilitarian details.

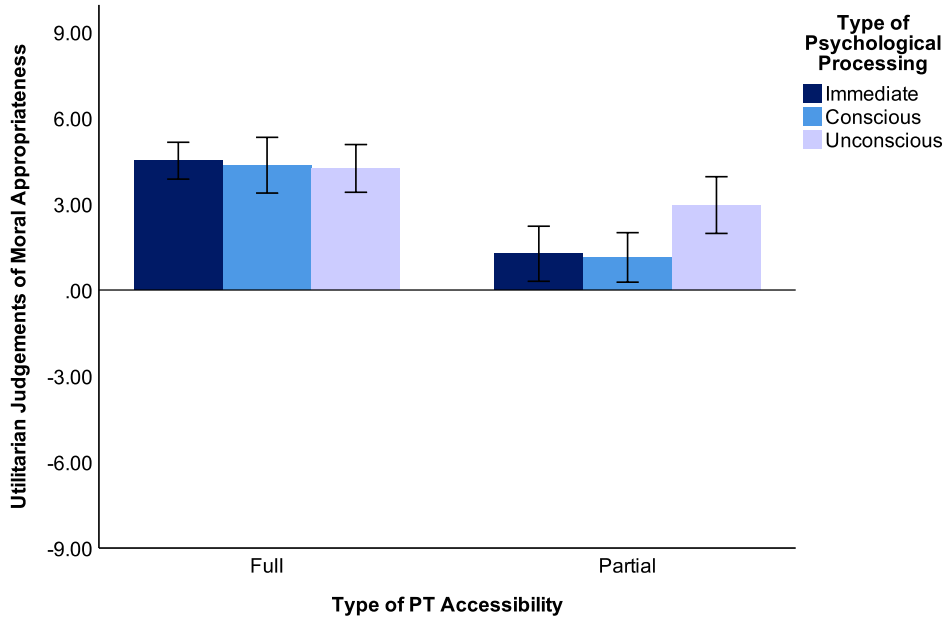
5.2.3.1 Judgements of Moral Appropriateness

A 2x3 independent measures analysis of variance was conducted to explore the influence of the independent variables type of PT accessibility (full or partial) and type of psychological processing (immediate, conscious, or unconscious) on judgements moral appropriateness. The results revealed that type of PT accessibility $F(1, 354) = 51.04$, $p < .001$,

η_p^2 = .13, as well as the two-way interaction type of PT accessibility by type of psychological processing $F(2, 354) = 3.25$, $p = .040$, $\eta_p^2 = .02$ significantly influenced respondents judgements of moral appropriateness. However, the results revealed that main effect of type of psychological processing on judgements of moral appropriateness was not statistically significant $F(2, 354) = 2.16$, $p = .117$, $\eta_p^2 = .01$ (see Figure 22).

Figure 22

Participants' Utilitarian Judgements of Moral Appropriateness in Experiment 7



Note. Positive mean values indicate participants' preference for utilitarian-swerve AVs. Negative mean values indicate participants' preference for non-utilitarian stay AVs. Error bars represent 95% confidence intervals of the mean.

Due to the significant two-way interaction, follow-up simple-effect tests were conducted by type of PT accessibility.

Partial PT Accessibility. A follow-up simple-effect test revealed that with partial PT accessibility, the main effect of type of psychological processing $F(2, 177) = 4.71$, $p = .010$, $\eta_p^2 = .05$ significantly influenced respondents' judgements of moral appropriateness. Specifically, the results revealed that with partial PT accessibility, participants were more

utilitarian in their moral judgements with unconscious psychological processing ($M = 2.97$; $SD = 3.84$) than with conscious psychological processing ($M = 1.14$; $SD = 3.35$), $p = .020$, and immediate psychological processing ($M = 1.27$; $SD = 3.72$), $p = .034$. Moreover, the results revealed no statistically significant difference between conscious and immediate processing ($p>.05$); see Figure 22.

Full PT Accessibility. Importantly and in contrast to the moral judgements with partial PT accessibility, with full PT accessibility respondents' moral appropriateness judgements were not influenced by the type of psychological processing ($F<1$); see Figure 22.

5.2.4 Discussion

According to UTT, complex decisions are better solved during unconscious processing as opposed to conscious processing (Dijksterhuis & Nordgren, 2006). Whilst complexity of the AV crash scenarios was not manipulated in the current experiment, it can be assumed that the scenarios were complex, since under conditions of partial PT accessibility, utilitarian choice was greater in the unconscious processing condition compared with the conscious processing and immediate judgement conditions. Accordingly, when participants were presented with partial PT accessibility to crash scenarios, the results replicated Ham and van den Bos (2010), where unconscious processing led to normative utilitarian judgements of moral appropriateness. These findings also replicate the general behavioural pattern predicted when employing the UTT paradigm (Dijksterhuis & Strick, 2016). However, participants that received full PT accessibility to AV crash scenarios were equally utilitarian in their judgements of moral appropriateness across type of processing conditions. Accordingly, the results revealed that full PT accessibility eliminates the effect of type of psychological processing on judgements of moral appropriateness. Moreover, and in accordance with the experiments reported in this thesis, respondents' judgements of moral appropriateness with full PT

accessibility were overall more utilitarian than the respondents' judgements with partial PT accessibility.

Importantly, one can rule out the possibility that decision problems become simple when full PT accessibility is offered, since under full PT accessibility there was no advantage of conscious processing over unconscious processing. Moreover, full PT accessibility required the participant to read more detailed (complex) information that partial PT accessibility. Thus, whilst full PT accessibility increased the complexity of the decision information, it did not result in unconscious processing producing greater moral approval for utilitarian-swerve AVs than conscious processing, as would be predicted by UTT (Dijsterhuis & Nordgren, 2006).

In conclusion, Experiment 7 has not only replicated the enhancement in utilitarian choice between type of PT accessibility conditions, but also demonstrated that providing unbiased PT information eliminates the differences in utilitarian choice between type of processing conditions. Therefore, Experiment 7 (for the first time) presents evidence that the decision-making information with full PT accessibility has a greater impact on people's moral judgements than the type of thinking (conscious or unconscious) used to process the information.

5.3 Experiment 8: The Influence of Perspective-Taking Accessibility on Conscious and Unconscious Moral Purchasing Values

5.3.1 Introduction

The results from Experiment 7 revealed for the first time that the way in which people process information (immediately, consciously or unconsciously) has no influence on their judgements of moral appropriateness when the PT information is fully accessible. However, another interesting line of enquiry is whether moral judgements (as influenced by PT accessibility and type of psychological processing) inform participants' purchasing values (how much they would spend on utilitarian-swerve and non-utilitarian stay AVs). In Experiment 3 demonstrated

that when participants received full PT accessibility to AV crash scenarios, judgements of moral appropriateness informed their subsequent purchasing values. However, it is not yet clear whether this mediation will remain across the different types of psychological processing. Accordingly, the goal of Experiment 8 is to investigate whether variations in type of processing will influence the mediating effect of judgements of moral appropriateness on the relationship between on PT accessibility and purchasing values.

5.3.2 Method

5.3.2.1 Participants

Three hundred and sixty participants (199 females, 161 males) were recruited to take part in an online experiment through PureProfile's online survey panels. The mean age of the participants was 48 (*SD* = 14.11). Importantly, ethical approval was granted by the BSREC prior to data collection and all participants were treated in accordance with APA ethical guidelines.

A significance level of .05 was set for statistical testing. Moreover, a retrospective power analysis was conducted on the independent-measures effects of *type of PT accessibility* and *type of psychological processing* and their interaction. The experiment was live for 14 days to ensure that data collection with a large sample size will achieve a statistical power of at least .95. According to the retrospective power analysis, the sample size ($N = 360$) produced a power of 1.00 which was sufficient to achieve the target.

5.3.2.2 Experimental Design

Identical to Experiment 7, a 2 (type of PT accessibility) x 3 (type of psychological processing) experimental design was employed. The independent variables were manipulated in the same way as Experiment 7. However, unlike Experiment 7, in Experiment 8 there were 2 dependent variables. The first dependent variable was judgements of moral appropriateness

and was measured for each AV model (utilitarian swerve and non-utilitarian stay) on separate 10-point rating scales (and the utilitarian weight was calculated by subtracting judgements for non-utilitarian stay AVs from utilitarian swerve AVs). The second dependent variable was purchasing value, where participants were required distribute a budget of £50,000 between two AV models (indicating how much they would spend on each model).

5.3.2.3 Materials and Procedure

The experiment followed the same procedure as Experiment 7. Participants received a moral AV scenario and visual stimuli that contained either full or partial PT accessibility and were then randomly allocated to a type of psychological processing condition (immediate, conscious or unconscious). Participants in the immediate condition provided their judgements of moral appropriateness and indicated their purchasing values immediately after reading the AV moral scenario (but on a separate webpage from the scenario). Participants in the conscious condition redirected to a separate webpage where they were encouraged to think carefully about the moral appropriateness of each AV for 3 minutes. Participants in the unconscious condition were distracted with an anagram task (see Appendix A for the anagrams employed in this experiment) for 3 minutes to prevent conscious processing but induce unconscious processing of the moral judgement task. After 3 minutes elapsed in both the conscious and unconscious processing conditions, participants were redirected to another webpage where they could make their judgements of moral appropriateness for each AV model and indicate their purchasing values. After making judgements participants completed a manipulation check to ensure that they could correctly recall the moral AV scenario's utilitarian details.

It is important to note that following the logic of unconscious processing experiments (e.g., Dijksterhuis et al., 2006) participants in the conscious and unconscious condition were informed that they will be making judgements of moral appropriateness before the 3 minutes of conscious or unconscious processing. However, at no point during the experiment were

participants informed that they will be making purchasing value judgements. The purpose of this was to further test though mediation analysis whether conscious/unconscious processing of moral judgements would, in turn, affect purchasing value judgements (see Appendix A and B for full experimental materials).

5.3.3 Results

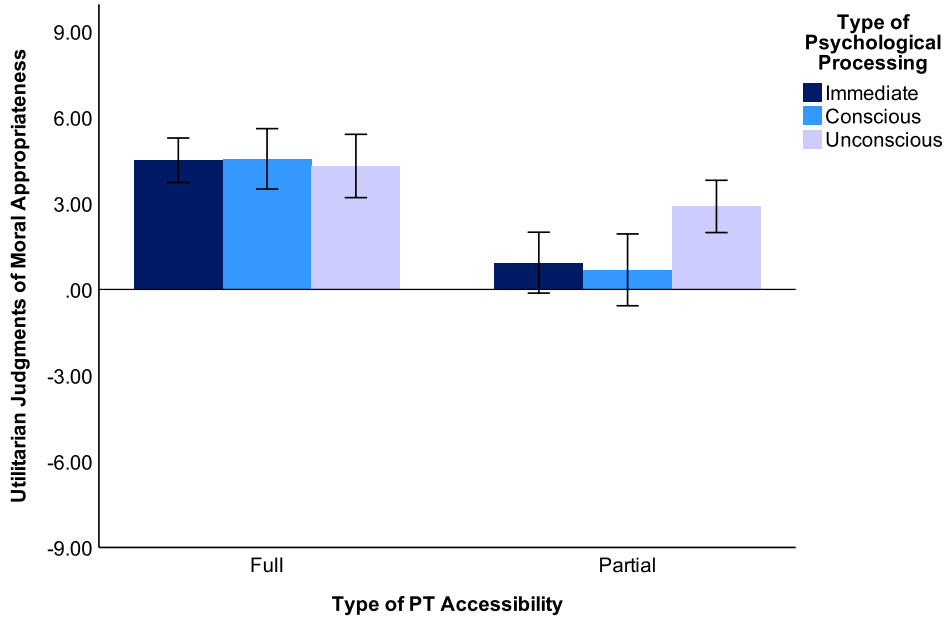
All 360 participants included in the analysis passed the manipulation check. Therefore, any differences in utilitarian judgements across conditions cannot by accounted for by a decline in memory for utilitarian details.

5.3.3.1 Judgements of Moral Appropriateness

A 2x3 independent measures analysis of variance was conducted to explore the influence of the independent variables type of PT accessibility (full or partial) and type of psychological processing (immediate, conscious, or unconscious) on judgements of moral appropriateness. The results revealed that type of PT accessibility $F(1, 354) = 48.51$, $p < .001$, $\eta_p^2 = .12$, as well as the two-way interaction type of PT accessibility by type of psychological processing $F(2, 354) = 3.35$, $p = .036$, $\eta_p^2 = .02$ significantly influenced respondents judgements of moral appropriateness. However, the results revealed that main effect of type of psychological processing on judgements of moral appropriateness was not statistically significant $F(2, 354) = 2.16$, $p = .116$, $\eta_p^2 = .01$ (see Figure 23).

Figure 23

Participants' Utilitarian Judgements of Moral Appropriateness in Experiment 8



Note. Positive mean values indicate participants' preference for utilitarian-swerve AVs. Negative mean values indicate participants' preference for non-utilitarian stay AVs. Error bars represent 95% confidence intervals of the mean.

Due to the significant two-way interaction, follow-up simple-effect tests were conducted by type of PT accessibility.

Partial PT Accessibility. A follow-up simple-effect test revealed that with partial PT accessibility, the main effect of type of psychological processing $F(2, 177) = 4.98$, $p = .008$, $\eta_p^2 = .05$ significantly influenced respondents' judgements of moral appropriateness. Specifically, the results revealed that with partial PT accessibility, participants were more utilitarian in their moral judgements with unconscious psychological processing ($M = 2.89$; $SD = 3.54$) than with conscious psychological processing ($M = 0.68$; $SD = 4.85$), $p = .013$, and immediate psychological processing ($M = 0.94$; $SD = 4.11$), $p = .034$. Moreover, the results revealed no statistically significant difference between conscious and immediate processing ($p>.05$); see Figure 23.

Full PT Accessibility. Importantly and in contrast to the moral judgements with partial PT accessibility, with full PT accessibility respondents' moral appropriateness judgements were not influenced by the type of psychological processing (F<1); see Figure 23.

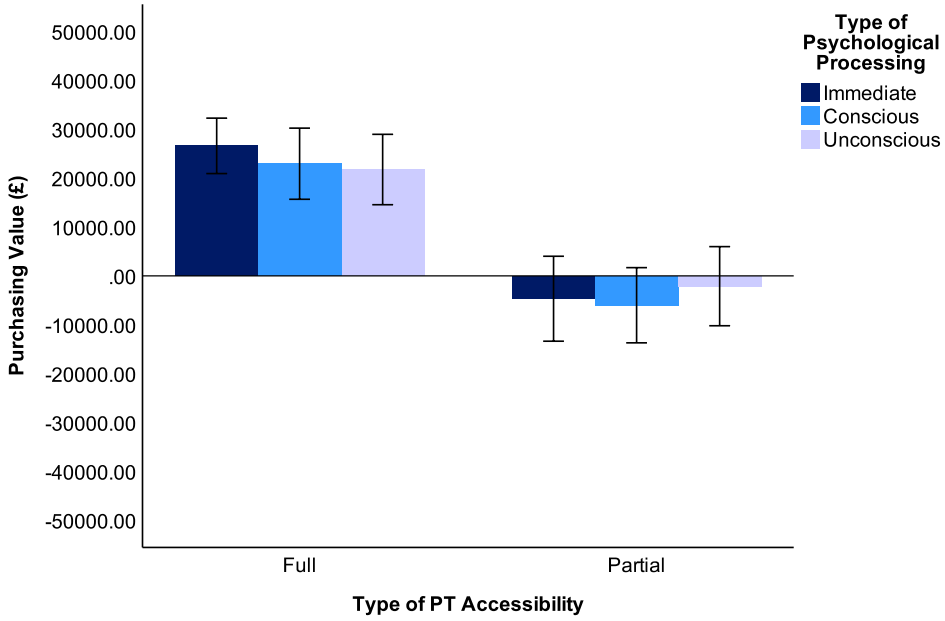
5.3.3.2 Purchasing Value

A 2x2 independent measures analysis of variance was conducted to explore the influence of type of PT accessibility and type of psychological processing (as well as their interaction) on purchasing value. The results revealed a significant main effect of type of PT accessibility $F(1, 354) = 84.00$, $p < .001$, $\eta_p^2 = .19$ on purchasing value. However, the main effect of type of psychological processing (F<1), as well as the two-way interaction type of PT accessibility by type of psychological processing (F<1) were not statistically significant.

Accordingly, participants indicated that they would pay £23,761.11 (M = £23,761.11; SD = £26,104.08) more for a swerve AV than a stay AV (when the PT accessibility was full) and £4,272.22 (M = -£4,272.22; SD = £31,474.89) less for a swerve AV than a stay AV (when the PT accessibility was less; see Figure 24). Therefore, with full PT accessibility (i) respondents were utilitarian in their purchasing behaviour irrespectively of type of psychological processing and (ii) there was no difference in purchasing value between types of psychological processing.

Figure 24

Participants' Reported Purchasing Values for Utilitarian AVs in Experiment 8



Note. Positive purchasing values indicate utilitarian behaviour (more money spent on swerve AVs than stay AVs from the budget of £50,000) and negative purchasing values indicate non-utilitarian behaviour (more money spend on stay AVs than swerve AVs from the budget of £50,000). Error bars represent 95% Confidence Intervals of the mean.

5.3.3.3 Predicting Purchasing Value

Three mediation analyses (by type of psychological processing: immediate, conscious and unconscious) with macro PROCESS were conducted to test whether the respondents' judgements of moral appropriateness mediates the relationship between type of PT accessibility and reported purchasing values. The predictor variable was PT accessibility, the mediator was respondents' judgements of moral appropriateness and the outcome variable was respondents' reported purchasing values. The standardised indirect effect of PT accessibility through the mediator judgements of moral appropriateness was tested by bootstrapping with N = 5000. I found that decision-makers' judgements of moral appropriateness is a mediator of the relationship between type of PT accessibility (full and partial) and reported purchasing

values. Hence, respondents' purchasing behaviour was informed by their moral judgements (the utilitarian weight of moral appropriateness).

Immediate Psychological Processing. The mediation model was significant, $F(2, 117) = 26.71$, $p < .001$; the model explained 31% of the variance in purchasing values ($R^2 = .31$). In addition, the standardized total effect of PT accessibility on purchasing value was also significant ($\beta = -.49$, $t = -6.04$, $p < .001$). The results also revealed that with immediate type of psychological processing, the standardized indirect effect of PT accessibility through the mediator judgements of moral appropriateness was significant and negative, $\beta = -.14$, BCa $CI(.95) = [-.263; -.046]$, indicating that judgements of moral appropriateness is a mediator of the relationship between PT accessibility and purchasing values. The results revealed that judgements of moral appropriateness partially mediated the relationship between PT accessibility and purchasing value as the standardized direct effect of PT accessibility on purchasing value was significant in the mediation model; however, this effect was weakened from (standardized total effect $\beta = -.49$, $t = -6.04$, $p < .001$) to (standardized direct effect $\beta = -.35$, $t = -4.05$, $p < .001$) when the mediator was included as a predictor. Specifically, with immediate type of psychological processing, respondents' reported purchasing values for swerve AVs were partially mediated by the judgements of moral appropriateness and were higher in the full PT accessibility condition than in the partial PT accessibility condition. As predicted, participants' utilitarian purchasing values were influenced by PT accessibility and informed by the moral judgements.

Conscious Psychological Processing. The mediation model was significant, $F(2, 117) = 27.11$, $p < .001$; the model explained 32% of the variance in purchasing values ($R^2 = .32$). In addition, the standardized total effect of PT accessibility on purchasing value was also significant ($\beta = -.45$, $t = -5.47$, $p < .001$). The results also revealed that with conscious type of psychological processing, the standardized indirect effect of PT accessibility through the

mediator judgements of moral appropriateness was significant and negative, β = -.15, BCa *CI*(.95) = [-.269; -.068], indicating that judgements of moral appropriateness is a mediator of the relationship between PT accessibility and purchasing values. The results revealed that judgements of moral appropriateness partially mediated the relationship between PT accessibility and purchasing value as the standardized direct effect of PT accessibility on purchasing value was significant in the mediation model; however, this effect was weakened from (standardized total effect (β = -.45, t = -5.47, p < .001) to (standardized direct effect β = -.30, t = -3.63, p < .001) when the mediator was included as a predictor. Specifically, with conscious type of psychological processing, respondents' reported purchasing values for swerve AVs were partially mediated by the judgements of moral appropriateness and were higher in the full PT accessibility condition than in the partial PT accessibility condition. As I predicted, participants' utilitarian purchasing values were influenced by PT accessibility and informed by the moral judgements.

Unconscious Psychological Processing. The mediation model was significant, F(2, 117) = 20.18, p <.001; the model explained 26% of the variance in purchasing values (R^2 = .26). In addition, the standardized total effect of PT accessibility on purchasing value was also significant (β = -.38, t = -4.41, p < .001). The results also revealed that with unconscious type of psychological processing, the standardized indirect effect of PT accessibility through the mediator judgements of moral appropriateness was significant and negative, β = -.06, BCa *CI*(.95) = [-.153; -.006], indicating that judgements of moral appropriateness is a mediator of the relationship between PT accessibility and purchasing values. The results revealed that judgements of moral appropriateness partially mediated the relationship between PT accessibility and purchasing value as the standardized direct effect of PT accessibility on purchasing value was significant in the mediation model; however, this effect was weakened from (standardized total effect (β = -.38, t = -4.41, p < .001) to (standardized direct effect β =

-.31, t = -3.88, p < .001) when the mediator was included as a predictor. Specifically, with unconscious type of psychological processing, respondents' reported purchasing values for swerve AVs were partially mediated by the judgements of moral appropriateness and were higher in the full PT accessibility condition than in the partial PT accessibility condition. As I predicted, participants' utilitarian purchasing values were influenced by PT accessibility and informed by the moral judgements.

5.3.4 Discussion

Experiment 8 successfully replicated the influence of PT accessibility on participants' judgements of moral appropriateness (a finding consistent across Experiments 2-7). In particular, participants were more likely to morally approve of prosocial utilitarian-swerve AVs when they received AV crash scenarios with full PT accessibility compared to when they received AV crash scenarios with partial PT accessibility. Moreover, as with Experiment 7, psychological processing did not influence participants' judgements of moral appropriateness when PT accessibility was full. However, when PT accessibility was partial, participants who processed the AV dilemma unconsciously were more utilitarian in their responses than participant in the immediate judgements and conscious processing conditions (replicating UTT findings; e.g., Ham & van den Bos, 2010). Accordingly, these findings replicate Experiment 8, demonstrating that the method in which people process information is irrelevant if the information is fully accessible in the first place.

Participants' purchasing values were also influenced by PT accessibility. Specifically, participants indicated that they would spend more money on non-utilitarian stay AVs when they received AV crash scenarios with full PT accessibility and more money on utilitarian-swerve AVs when they received AV crash scenarios with partial PT accessibility. However, the type of processing employed did not influence purchasing values across partial and full PT accessibility conditions. It is plausible that this is the result of not informing participants before

the distraction period that they will be making subsequent purchasing value judgements. Accordingly, participants would have only processed judgements of moral appropriateness, and not the purchasing values during the conscious thought/distraction period.

Finally, participants' judgements of moral appropriateness informed their purchasing values regardless of the type of psychological processing that was employed. In particular, with full PT accessibility, participants were more likely to morally approve of utilitarian-swerve AVs over non-utilitarian stay AVs and in turn indicated that they would spend more money on utilitarian-swerve AVs compared to non-utilitarian stay AVs. Therefore, regardless of type of processing employed, when PT accessibility is full participants are more utilitarian in both their judgements of moral appropriateness and purchasing values.

5.4 Experiment 9: The Influence of Perspective-Taking Accessibility on Conscious and Unconscious Moral Willingness to Buy

5.4.1 Introduction

Numerous experimental studies have demonstrated the advantage of processing complex information unconsciously across a wide variety of behavioural tasks including consumer decision-making (Messner & Wänke, 2011), forecasting accuracy (Dijksterhuis et al., 2009), lie detection accuracy (Reinhard et al., 2013), and moral decision-making (Ham & van den Bos, 2010). Moreover, Experiments 7 and 8 have replicated the unconscious advantage in moral judgements related to the AV dilemma where participants demonstrate a greater moral approval of utilitarian-swerve AVs after unconscious processing of AV crash scenarios. However, the unconscious advantage remained only in conditions of partial PT accessibility, where participants were limited to one situational perspective of the AV crash scenario. Alternatively, when participants had access to all situational perspectives in the AV crash scenario (full PT accessibility), they demonstrated a preference for prosocial utilitarian-swerve AVs regardless of the type of psychological processing employed. Thus, with full accessibility

to unbiased information, people make prosocial utilitarian decisions regardless of the type of psychological processing they have employed.

Another interesting finding from Experiment 8 revealed that when PT accessibility was full, judgements of moral appropriateness informed participants' purchasing values, even though the type of psychological processing of purchasing values was not manipulated. This finding further demonstrates that with full PT accessibility, judgements of moral appropriateness inform participants' purchasing values (as with Experiment 3) regardless of the type of psychological processing employed. However, in order to determine whether this effect remains in all purchasing behavioural tasks, it is important to establish whether judgements of moral appropriateness inform participants' willingness to buy AVs. Accordingly, as a final line of enquiry, Experiment 9 replicates Experiment 8, however, the dependent variable purchasing value is replaced with willingness to buy. Thus, Experiment 9 investigates the influence of PT accessibility, type of psychological processing and their interaction on judgements of moral appropriateness and in turn, participants' willingness to buy utilitarian-swerve and non-utilitarian stay AV models.

5.4.2 Method

5.4.2.1 Participants

Participants (N = 360) were recruited to take part in an online computer-based experiment through PureProfile's online survey panels. The sample consisted of 217 females and 143 males and the mean age of the participants was 50 (SD = 13.53). Prior to data collection, ethical approval was granted by the BSREC. Moreover, all participants were treated in accordance with BPS ethical guidelines.

For the purpose of statistical testing, a significance level of .05 was set. Moreover, a retrospective power analysis was conducted on the independent-measures effects of *type of PT*

accessibility and *type of psychological processing* and their interaction. The experiment was live for 14 days to ensure that data collection with a large sample size will achieve a statistical power of at least .95. According to the retrospective power analysis, the sample size ($N = 360$) produced a power of 1.00 which was sufficient to achieve the target and yielded an identical power to Experiments 7 and 8.

5.4.2.2 Experimental Design

A 2x3 independent measures design was employed to test the effect of the independent variables, type of PT accessibility and type of psychological processing, on participants' judgements of moral appropriateness and willingness to buy each AV (utilitarian swerve and non-utilitarian stay). Judgements of moral appropriateness for each AV model were measured on separate 10-point rating scales (0 indicating 'not at all appropriate', and 9 indicating 'definitely appropriate') and a utilitarian weight was calculated by subtracting judgements for non-utilitarian stay AVs from judgements for utilitarian swerve AVs. Willingness to buy was measured and calculated using the same method; participants rated their willingness to buy each AV (0 indicated that they were not at all willing, and 9 indicated that they were definitely willing) and a utilitarian weight was calculated by subtracting willingness judgements for non-utilitarian stay AVs from willingness judgements for utilitarian swerve AVs.

5.4.2.3 Materials and Procedure

The procedure for Experiment 9 was similar to Experiments 7 and 8 in that participants received a moral AV scenario (and visual stimuli) with either partial PT accessibility or full PT accessibility and were then randomly allocated to one of 3 types of psychological processing conditions (immediate, conscious and unconscious; see Experiments 7 and 8). All participants made judgements of moral appropriateness and indicated their willingness to buy on a separate webpage from the moral AV scenario and visual stimuli.

Following the logic of unconscious processing experiments (e.g., Dijksterhuis et al., 2006) participants in the conscious and unconscious condition were informed that they will be making judgements of moral appropriateness before the 3 minutes of conscious or unconscious processing. However, at no point during the experiment were participants informed that they will be making purchasing willingness to buy judgements. The purpose of this was to further test though mediation analysis whether conscious/unconscious processing of moral judgements would affect willingness to buy judgements. After making judgements, as with Experiments 7 and 8, participants completed a manipulation check to ensure that they could correctly recall the moral AV scenario's utilitarian details (see Appendix A and B for full details of the experimental materials).

5.4.3 Results

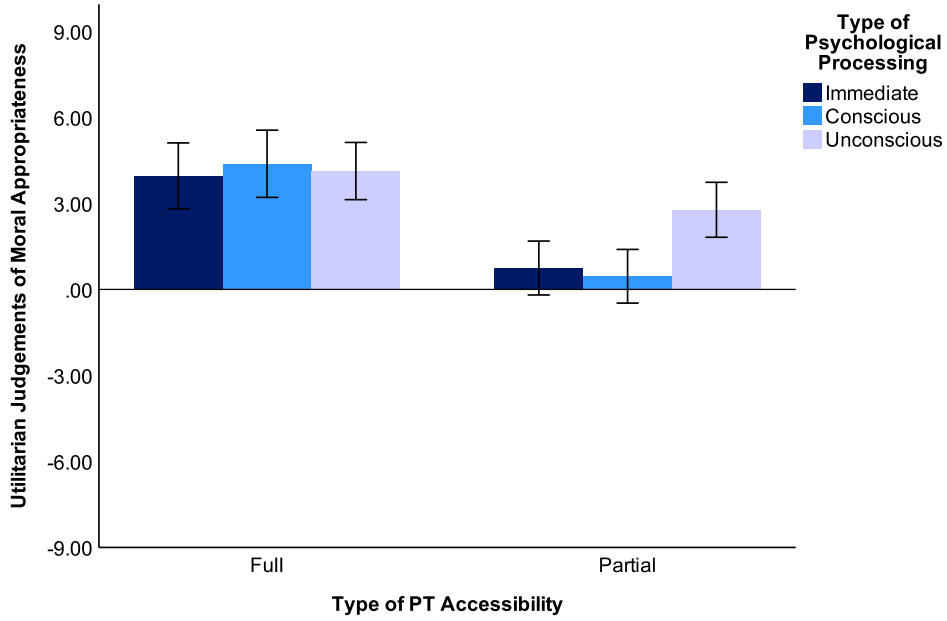
All 360 participants included in the analysis passed the manipulation check. Therefore, any differences in utilitarian judgements across conditions cannot by accounted for by a decline in memory for utilitarian details.

5.4.3.1 Judgements of Moral Appropriateness

A 2x2 independent measures analysis of variance was conducted to explore the influence of the independent variables type of PT accessibility (full or partial) and type of psychological processing (immediate, conscious, or unconscious) on judgements of moral appropriateness. The results revealed that type of PT accessibility $F(1, 354) = 45.30$, $p < .001$, $\eta_p^2 = .11$, as well as the two-way interaction type of PT accessibility by type of psychological processing $F(2, 354) = 3.31$, $p = .037$, $\eta_p^2 = .02$ significantly influenced respondents judgements of moral appropriateness. However, the results revealed that main effect of type of psychological processing on judgements of moral appropriateness was not statistically significant $F(2, 354) = 2.86$, $p = .059$, $\eta_p^2 = .01$ (see Figure 25).

Figure 25

Participants' Utilitarian Judgements of Moral Appropriateness in Experiment 9



Note. Positive mean values indicate participants' preference for utilitarian swerve AVs. Negative mean values indicate participants' preference for non-utilitarian stay AVs. Error bars represent 95% confidence intervals of the mean.

Due to the significant two-way interaction, follow-up simple-effect tests were conducted by type of PT accessibility.

Partial PT Accessibility. A follow-up simple-effect test revealed that with partial PT accessibility, the main effect of type of psychological processing $F(2, 177) = 7.15$, $p = .001$, $\eta_p^2 = .08$ significantly influenced respondents' judgements of moral appropriateness. Specifically, the results revealed that with partial PT accessibility, participants were more utilitarian in their moral judgements with unconscious psychological processing ($M = 2.78$; $SD = 3.71$) than with conscious psychological processing ($M = 0.46$; $SD = 3.63$), $p = .002$, and immediate psychological processing ($M = 0.75$; $SD = 3.64$), $p = .008$. Moreover, the results revealed no statistically significant difference between conscious and immediate processing ($p>.05$); see Figure 25.

Full PT accessibility. Importantly and in contrast to the moral judgements with partial PT accessibility, with full PT accessibility respondents' moral appropriateness judgements were not influenced by the type of psychological processing (F<1); see Figure 25.

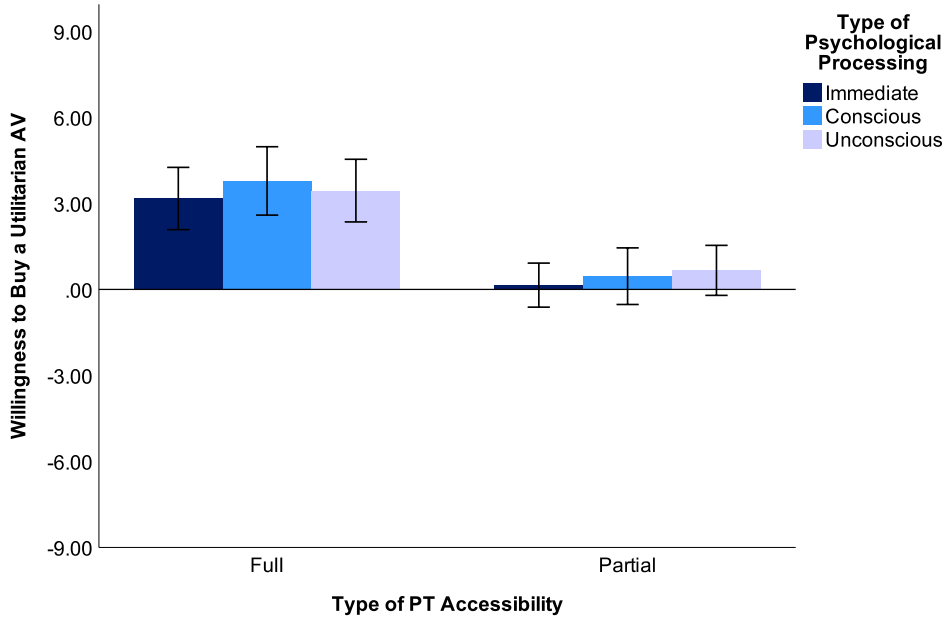
5.4.3.2 Willingness to Buy

A 2x2 independent measures analysis of variance was conducted to explore the influence of type of PT accessibility and type of psychological processing (as well as their interaction) on willingness to buy. The results revealed a significant main effect of type of PT accessibility $F(1, 354) = 54.43$, $p < .001$, $\eta_p^2 = .13$ on willingness to buy.

Importantly, with full PT accessibility, respondents' judgements of willingness to buy an AV were utilitarian ($M = 3.47$; $SD = 4.34$) and significantly different than the non-utilitarian judgements of willingness to buy an AV with partial PT accessibility ($M = 0.43$; $SD = 3.40$); see Figure 26. However, the main effect of type of psychological processing (F<1), as well as the two-way interaction type of PT accessibility by type of psychological processing (F<1) were statistically not significant.

Figure 26

Participants' Willingness to Buy a Utilitarian AV in Experiment 9



Note. Positive mean values indicate participants' preference for utilitarian swerve AVs. Negative mean values indicate participants' preference for non-utilitarian stay AVs. Error bars represent 95% confidence intervals of the mean.

5.4.3.3 Predicting Willingness to Buy.

Three mediation analyses (by type of psychological processing: immediate, conscious and unconscious) with macro PROCESS were conducted to test whether the respondents' judgements of moral appropriateness mediates the relationship between type of PT accessibility (full and partial) and reported willingness to buy each type of AV (utilitarian swerve and non-utilitarian stay). The predictor variable was type of PT accessibility, the mediator was respondents' judgements of moral appropriateness and the outcome variable was respondents' reported willingness to buy. The indirect effect of PT accessibility through the mediator judgements of moral appropriateness was tested by bootstrapping with $N = 5000$. The results revealed that decision-makers' judgements of moral appropriateness is a mediator of the relationship between type of PT accessibility (full and partial) and judgements for

willingness to buy and willingness to ride. Moreover, I found that respondents' willingness to buy judgements were informed by their moral judgements of appropriateness.

Immediate Psychological Processing. The mediation model was significant, $F(2, 117) = 37.23$, $p < .001$; the model explained 39% of the variance in willingness to buy ($R^2 = .39$). In addition, the standardized total effect of PT accessibility on willingness to buy was also significant ($\beta = -.39$, $t = -4.55$, $p < .001$). The results also revealed that with immediate type of psychological processing, the standardized indirect effect of PT accessibility through the mediator judgements of moral appropriateness was significant and negative, $\beta = -.20$, BCa $CI(.95) = [-.307; -.112]$, indicating that judgements of moral appropriateness is a mediator of the relationship between PT accessibility and willingness to buy. The results revealed that judgements of moral appropriateness partially mediated the relationship between PT accessibility and willingness to buy as the standardized direct effect of PT accessibility on willingness to buy was significant in the mediation model; however, this effect was weakened from (standardized total effect $\beta = -.39$, $t = -4.55$, $p < .001$) to (standardized direct effect $\beta = -.19$, $t = -2.45$, $p = .016$) when the mediator was included as a predictor. Specifically, with immediate type of psychological processing, respondents' reported willingness to buy for swerve AVs were partially mediated by the judgements of moral appropriateness and were higher in the full PT accessibility condition than in the partial PT accessibility condition. As I predicted, participants' utilitarian willingness to buy were influenced by PT accessibility and informed by the moral judgements.

Conscious Psychological Processing. The mediation model was significant, $F(2, 117) = 40.98$, $p < .001$; the model explained 41% of the variance in willingness to buy ($R^2 = .41$). In addition, the standardized total effect of PT accessibility on willingness to buy was also significant ($\beta = -.37$, $t = -4.29$, $p < .001$). The results also revealed that with conscious type of psychological processing, the standardized indirect effect of PT accessibility through the

mediator judgements of moral appropriateness was significant and negative, β = -.25, BCa *CI*(.95) = [-.368; -.161], indicating that judgements of moral appropriateness is a mediator of the relationship between PT accessibility and willingness to buy. Moreover, with conscious type of psychological processing the results revealed that judgements of moral appropriateness fully mediated the relationship between PT accessibility and willingness to buy as the standardized direct effect of PT accessibility on willingness to buy was not significant in the mediation model (β = -.11, t = -1.45, p = .149). Specifically, respondents' willingness to buy swerve AVs were fully mediated by the judgements of moral appropriateness and were higher in the full PT accessibility condition than in the partial PT accessibility condition.

Unconscious Psychological Processing. The mediation model was significant, $F(2, 117) = 35.43$, p <.001; the model explained 38% of the variance in willingness to buy (R^2 = .38). In addition, the standardized total effect of PT accessibility on willingness to buy was also significant (β = -.34, t = -3.98, p < .001). The results also revealed that with unconscious type of psychological processing, the standardized indirect effect of PT accessibility through the mediator judgements of moral appropriateness was significant and negative, β = -.10, BCa *CI*(.95) = [-.187; -.004], indicating that judgements of moral appropriateness is a mediator of the relationship between PT accessibility and willingness to buy. The results revealed that judgements of moral appropriateness partially mediated the relationship between PT accessibility and willingness to buy as the standardized direct effect of PT accessibility on willingness to buy was significant in the mediation model; however, this effect was weakened from (β = -.34, t = -3.98, p < .001) to (standardized direct effect β = -.25, t = -3.41, p < .001) when the mediator was included as a predictor. Specifically, with unconscious type of psychological processing, respondents' reported willingness to buy for swerve AVs were partially mediated by the judgements of moral appropriateness and were higher in the full PT accessibility condition than in the partial PT accessibility condition. As predicted, participants'

utilitarian willingness to buy were influenced by PT accessibility and informed by the moral judgements.

5.4.4 Discussion

The results form Experiment 9 demonstrated and confirmed that with full PT accessibility, participants were more consistent and utilitarian in their judgements of moral appropriateness and willingness to buy behaviour than with partial PT accessibility. In line with Experiment 7 and 8, the results also confirmed that full PT accessibility eliminates the effect of type of psychological processing on people's moral judgements. Moreover, under conditions of full PT accessibility, the dependent variable willingness to buy was informed by participants' judgements of moral appropriateness. In other words, when participants received AV crash scenarios containing full PT accessibility, they were more utilitarian in their judgements of moral appropriateness, and in turn were more willing to buy utilitarian-swerve over non-utilitarian stay AVs.

5.5 General Discussion

The main premise of UTT theory is that when judgement or decision-making information is complex, people make better decisions if they process this complex information unconsciously rather than consciously. Decision-making complexity can be manipulated in many ways such increasing the number of choice options (Messner & Wänke, 2011), increasing the number of choice attributes (Dijksterhuis et al., 2006) or increasing the amount of irrelevant information in the description of the task (Ham & van den Bos, 2010). Whilst judgement complexity was not directly manipulated in any of the Experiments in this Chapter, it can be assumed that the information presented in conditions of partial PT accessibility was complex. The rationale for this is that across all Experiments in Chapter 5, participants were more utilitarian (provided normative rational judgements according to EUT; von Neuman & Morgenstern, 1944) in their judgements of moral appropriateness after processing the

information unconsciously compared to consciously or immediately. Accordingly, the weighting principle of UTT theory (see Figure 2) may account for why people are more utilitarian in their judgements of moral appropriateness after processing information unconsciously rather that consciously. For instance, as argued by Kusev et al. (2016), moral scenarios that contain limited accessibility to information are 'cognitively challenging', however, according to the weighting principle of UTT, such cognitively challenging information can be weighted more accurately during unconscious processing as opposed to conscious processing (see Bos et al., 2010). That is, the unconscious mind is more apt at identifying the benefit of saving the greatest number of people than the conscious mind is.

In contrast, to partial PT accessibility, when participants were presented with full PT accessibility, the advantage of unconscious processing disappeared; participants were relatively utilitarian in their moral judgements regardless of the type of psychological processing they had employed. However, this does not mean that the judgement task has become less complex, if this were the case then one would expect people to be more utilitarian after processing judgement information consciously rather than unconsciously (as predicted by UTT; Dijksterhuis & Nordgren, 2006). However, this was not the case, with full PT accessibility people were equally utilitarian in their judgements of moral appropriateness across across types of psychological processing conditions. Therefore, this finding indicates that UTT is not only influenced by the complexity of information (Dijksterhuis et al., 2006) but also by the accessibility to information.

The difference in psychological processing effects between PT accessibility conditions could also be explained by the convergence-divergence principle of UTT. According to the convergence-divergence principle of UTT, conscious processing is convergent (focused on one solution) whereas unconscious processing is divergent (can approach the problem from many angles and find multiple solutions; see also Figure 2). Thus, drawing on Kahneman's (2003,

p.699) definition of accessibility, “the ease (or effort) with which particular mental contents come to mind”, it is possible that scenario’s with partial PT accessibility are difficult to imagine from multiple perspectives (because the information is not accessible), thus with conscious convergent thinking, participants will struggle to establish the prosocial utilitarian strategy. However, when processing information unconsciously, the divergent unconscious mind could potentially take on multiple perspectives, despite this information not being accessible. On the other hand, with full PT accessibility, PT information is fully accessible, inducing divergent thinking in the conscious mind. As a result, both conscious and unconscious processing led to relatively utilitarian judgements of moral appropriateness. This finding has important implications for UTT since it demonstrates that it is the information itself – and not the method in which it is processed – is more important in eliciting ‘good’ (or in this case prosocial, utilitarian) judgements.

Experiments 8-9 demonstrated that when PT accessibility was full, judgements of moral appropriateness informed purchasing behaviour (both purchasing value and willingness to buy judgements). These findings accordingly replicate Experiments 3-4 and additionally demonstrate that this mediating effect remains regardless of how the moral judgements were initially processed. Moreover, as intended by the design of the Experimental method of Experiments 8-9, participants were informed before the conscious thought/distraction period that they will make a moral judgement regarding each AV. This allowed participants to either consciously or unconsciously process the target information during the conscious thought/distraction period. However, purchasing behaviour judgements (purchasing values and willingness to buy judgements) were intentionally not introduced in the same way. The purpose of not introducing them was to prevent any processing of these tasks during the conscious thought/distraction period in order to establish whether they would be informed by participants’

judgements of moral appropriateness. Under full PT accessibility, this was the case; participants' judgements of moral appropriateness did inform their purchasing behaviour.

In light of these findings, two interesting avenues of research could be conducted. The first possibility for future research is related to decision-making complexity. The complexity of information could be directly manipulated in order to establish whether PT accessibility still eliminates the unconscious processing effect when judgement tasks are particularly complex. For example, increasing the number of possible driving trajectories within the AV dilemma may increase the level of complexity of the AV dilemma task (without interfering with the level of contextual or PT accessibility). The second possibility for future research is related to the induction task. In all 3 Experiments in Chapter 5, only the processing judgements of moral appropriateness were manipulated experimentally. It would be interesting to instead directly manipulate purchasing behaviours (purchase intention and willingness to buy judgements) in order to establish whether PT accessibility also eliminates the effect of unconscious processing in tasks related to purchasing behaviour as opposed to just tasks related to moral judgements.

CHAPTER 6

Conclusions and Future Work

6.1 Overview of Chapter 6

In the final Chapter of this thesis, I begin by summarising how Experiment 1 reveals the limits of contextual accessibility in dilemmas that involve PT tasks. I consequently discuss how Experiment 1 demonstrates the need for empirical research to investigate the influence of PT accessibility on moral behaviours. This is followed by a discussion of the main findings of Experiments 2-9, and in particular, how presenting participants with full PT accessibility results in consistent utilitarian behaviours across judgements tasks, behavioural elicitation methods, types of involvement and types of psychological processing. Moreover, I consider the research and practical implications the current thesis has on moral philosophy, moral psychology and the AV industry (AV manufacturers and policymakers). I also make a suggestion that full PT accessibility can support policymakers and car manufacturers' efforts in promoting pro-social life-saving vehicles. The thesis ends with a note on the limitations of the experimental explorations made in the thesis, proposed future research initiatives and final concluding remarks.

6.2 Summary and Discussion of Main Findings

6.2.1 Contextual Accessibility and its Limits in Perspective-Taking Tasks

There is little doubt that way in which decision-making information is presented to humans has an impact on their judgements and choices (Kusev et al., 2016, 2018; Tversky & Kahneman, 1981). When constructing decision-making information such as hypothetical moral dilemmas, authors typically make dilemma information available to participants, from which participants can psychologically process the details and make informed judgements or choices. However, as first demonstrated in research on human memory (e.g., Tulving & Pearlstone, 1966), whilst information may be available, it does not mean that it is readily accessible. Likewise, traditional moral dilemmas based on the trolley paradigm (see Foot, 1967; Thomson, 1985) often lack accessibility to all decision-making actions and consequences (contextual

accessibility; Kusev et al., 2016). Accordingly, Kusev et al.'s (2016) experiment demonstrated that presenting moral dilemmas with partial contextual accessibility induces uncertainty in decision-makers, resulting in inconsistent utilitarian moral preferences across behavioural tasks. However, when clearly presenting the action and consequence of each choice within moral dilemmas and moral questions (full contextual accessibility), this behavioural inconsistency was eliminated (and people were generally more utilitarian in their preferences). Similarly, Experiment 1 of this thesis applied contextual accessibility to a modern moral dilemma: the AV dilemma. However, whilst full contextual accessibility increased participants' utilitarian judgements of moral appropriateness, it did not influence their purchasing behaviour (willingness to buy judgements). Accordingly, I reasoned that this was due to the AV dilemma being characteristically different to traditional moral dilemmas. For instance, unlike traditional moral dilemmas (e.g., the trolley problem) the AV dilemma involves PT, where participants must imagine themselves as character who may be affected by choice outcomes. Moreover, the AV dilemma employed in Experiment 1 was adapted from Bonnefon et al.'s (2016) AV dilemma and offered only partial accessibility to PT. Accordingly, participants were only presented with the perspective of the AV passenger and were offered the corresponding perspective of the pedestrians. Therefore, even when offering full contextual accessibility, PT accessibility was still partial, which may explain why contextual accessibility had no influence on participants' purchase intention. Appropriately, the subsequent experiments (Experiments 2-9) tested the influence of PT accessibility on people's moral judgements related to AV crash scenarios. However, importantly, since full contextual accessibility has been demonstrated to be an important improvement in the presentation of dilemma information (e.g., Kusev et al., 2016 and Experiment 1 of this thesis), it was kept constant in Experiments 2-9.

6.2.2 Perspective-Taking Accessibility and Consistent Utilitarian Moral Preferences

The main finding from Experiments 2-9 revealed that presenting participants with moral dilemmas containing full PT accessibility resulted in participants making consistent utilitarian moral judgements and choices. Accordingly, Experiments 2-9 have demonstrated utilitarian preference consistencies across many domains including judgement tasks, behavioural elicitation methods, types of scenario involvement, and types of psychological processing. Therefore, in this subsection I address how PT accessibility leads to behavioural consistencies in each of these domains.

6.2.2.1 Utilitarian Consistency Across Judgement Tasks

The controversial conclusion of Bonnefon et al.'s (2016) study is that people judge utilitarian AVs as the most morally appropriate vehicle for societal use yet would rather buy non-utilitarian AVs for themselves and their family. As discussed previously this behavioural inconsistency is known in the social psychology literature as a moral hypocrisy – where people want to appear moral whilst avoiding the cost of actually being moral (Batson, 2011). However, in this thesis I have argued and demonstrated that this moral hypocrisy is the result of limited accessibility to PT in AV crash scenarios. Several Experiments (3, 4, 8 and 9) in this thesis have accordingly demonstrated that when PT accessibility is full, participants do not demonstrate moral hypocrisies in their behaviour. Specifically, when participants receive full PT accessibility to AV crash scenarios, they are utilitarian in their judgements of moral appropriateness, purchasing behaviours (purchasing values and willingness to buy judgements), and usage behaviours (willingness to ride judgements). Moreover, with full PT accessibility, participants' utilitarian judgements of moral appropriateness even inform their purchasing and usage behaviours. Thus, presenting full PT accessibility eliminates behavioural inconsistencies between judgements tasks and purchasing behaviours.

6.2.2.2 Utilitarian Consistency Across Behavioural Elicitation Methods

Presenting scenarios with full PT accessibility does not only result in consistent utilitarian behaviour across judgement tasks but also across behavioural elicitation methods. Previous research has demonstrated that the behavioural elicitation method employed can influence their behaviour (e.g., Kusev et al., 2020; Lichtenstein & Slovic, 1971; Pedroni et al., 2017). In particular, in some studies, participants' choices are distinct from their judgements (e.g., what they choose is not necessarily what they judge to be the most valuable; Lichtenstein & Slovic, 1971). However, the results from Experiment 5 revealed a weak positive association between participants' moral judgements and moral choices under conditions of partial PT accessibility; yet, this association became moderate-strong under conditions of full PT accessibility. In other words, participants who received full PT accessibility to crash scenarios were more utilitarian in both their moral choices and moral judgements. Further exploration into the relationship between participants' moral judgements and moral choices (Experiment 6) revealed that with full PT accessibility, participants moral choices inform their moral judgements. Thus, revealing a PT accessibility choice-induced change in moral judgements as opposed to simply a choice-induced change in judgements as predicted by free-choice theorists (e.g., Brehm, 1956; Izuma & Murayama, 2013; Sharot et al., 2012).

6.2.2.3 Utilitarian Consistency Across Types of Involvement

Another interesting example of utilitarian preference consistency was demonstrated between types of involvement (see Experiments 2-4). Type of involvement refers to the type of PT task participants engaged in when reading the scenarios (whether participants imagined themselves or a stranger in the AV crash scenarios). Moreover, stranger involvement was equivalent to imagine-other PT as described by Batson et al. (1997b), whereas participant involvement was equivalent to imagine-self PT. Accordingly, previous research has

established behavioural and affective differences between these types of PT tasks (Stotland, 1969; Batson et al., 1997b; Bonnefon et al., 2016), where imagine-other PT leads to moral behaviour whilst imagine-self PT does not (e.g., Batson et al., 2003). However, these tasks all contained partial PT accessibility (the participant was required to take the perspective of only one person in the scenario or task). Similarly, Experiment 3 of this thesis demonstrated that with partial PT accessibility, participants were more utilitarian in their purchasing values when they had engaged in imagine-other PT (stranger involvement) than when they had engaged in imagine-self PT (participant involvement). However, under conditions of full PT this difference was eliminated; participants were consistently utilitarian in their purchasing values across both types of involvement. Moreover, this effect was also replicated when the type of PT become even more emotionally salient and involved imagining themselves with a family member (Experiment 4). Specifically, under conditions of full PT accessibility, participants were generally utilitarian in their purchasing behaviour and usage behaviour regardless of whether they imagined themselves with or without the presence of a family member in the scenario. Therefore, contrary to previous findings (Batson et al., 1997b; Bonnefon et al., 2016; Stotland, 1969) full PT accessibility eliminates differences between types of involvement, resulting in consistent utilitarian preferences across tasks that present different types of involvement.

6.2.2.4 Utilitarian Consistency Across Types of Psychological Processing

The final example of full PT inducing utilitarian preference consistencies is demonstrated across types of psychological processing. According to unconscious thought theorists, the way in which we psychologically process information determines our judgements and decisions related to that information (Dijksterhuis & Nordgren, 2006). In particular, processing complex information unconsciously leads to better decisions than processing the same information consciously (see Dijksterhuis et al., 2016 for a review). However, the current

thesis supports a different proposal; the information itself (e.g., how accessible the information is) determines our judgements and decisions. Accordingly, Experiments 7-9 demonstrate that will partial PT accessibility, participants' judgements of moral appropriateness are more utilitarian if they processed AV crash scenarios unconsciously, and less utilitarian if they processed AV crash scenarios consciously (replicating UTT predictions). However, when presented with full PT accessibility this utilitarian inconsistency is eliminated and participants are generally utilitarian in their moral judgements regardless of the method they employed to process the information.

6.2.2.5 Conclusion of Findings and an Important Clarification

The findings from Experiments 2-9 accordingly demonstrate that presenting participants with scenarios that contain full PT accessibility results in behavioural consistencies across judgement tasks, behavioural elicitation methods (judgement and choice), types of involvement (participants, stranger and family member) and types of psychological processing (immediate, conscious and unconscious). In other words, regardless of the behaviour being measured, the method used to measure behaviour, the contextual details in the scenario or the way people process information, people approve of utilitarian AVs when full PT accessibility is provided. Thus, PT accessibility is a fundamental psychological phenomenon that results in consistent normative (and prosocial) utilitarian preferences.

In light of these findings, it is important to clarify how PT accessibility influences people's moral behaviours. In this thesis, partial PT accessibility has been found to bias respondents towards non-utilitarian moral preferences. This is because partial PT accessibility requires the participants to take on only the perspective of the AV passenger, which results in participants overestimating the potential dangers of owning utilitarian AVs, and thus favouring passenger-protective (non-utilitarian) vehicles. One might argue then, that full PT accessibility simply nudges people away from making non-utilitarian judgements and towards making

utilitarian judgements. However, this is not the case, since full PT accessibility does not present participants with inaccessible information regarding any situational perspectives in AV crash scenarios. If one wanted to accordingly nudge people into making utilitarian decisions, they could present partial PT accessibility which only allows the respondent to take on the perspective of the pedestrian. This would accordingly emphasise the dangers of passenger-protective vehicles and may lead to utilitarian preferences. However, full PT accessibility offers representations of all situational perspectives in AV crash scenarios, allowing participants can make their judgements and choices in the absence of bias. Therefore, rather than nudging participants into making utilitarian decisions, full PT accessibility removes the bias in the decision-making information (partial PT accessibility) and reveals people's actual preferences, which happen to be utilitarian. These important findings reveal an opportunity for full PT accessibility to be used as an educational tool and decision support system. Future research should accordingly explore these applications.

6.3 Implications of the Current Thesis: Current and Future Directions

6.3.1 Theoretical and Practical Contributions of the Current Thesis

This section addresses how the experimental and theoretical explorations made in this thesis contributes to existing research. I accordingly address how this thesis informs Jeremy Bentham's philosophy of utilitarianism as well as normative and descriptive theories of decision-making psychology. Moreover, I also address the practical implications this thesis has on AV car manufacturers and policymakers in terms of how to market AVs.

6.3.1.1 Jeremy Bentham's Moral Philosophy of Utilitarianism

Jeremy Bentham (1789/1970) theory of utilitarianism is informed by the experimental explorations made in the present thesis. One of the tenets of utilitarianism put forward by Bentham (1789/1970) was that in order to make a utilitarian choice, the decision-maker should not be directly affected by decision outcomes. Bentham reasoned that if decision-makers know

they will be affected by the decisions they make, they may deviate from the utilitarian option and fall prey to their egoistic tendencies. Experimental research in psychology has accordingly provided evidence for Bentham's prediction; people who know they will be affected by decision outcomes tend to make non-utilitarian purchasing judgements (Bonnefon et al., 2016). However, as demonstrated throughout this thesis, when respondents have full access to all situational perspectives in moral scenarios, their judgements reflect that of an objective utilitarian. Hence, the current thesis informs Bentham's utilitarian theory; when people have full access to PT, they can make utilitarian judgements (and choices), even when they will be directly affected by the judgements and choices they make.

6.3.1.2 Normative and Descriptive (Moral) Decision-Making Psychology

As described in Chapter 1, normative decision-making theorists assume that human decision-makers are rational and consistent in their preferences (Sugden, 1991). Whilst these theories do not account for actual human behaviour (Tversky & Kahneman, 1992), there are some circumstances where consistent rational preferences can have positive outcomes. The AV dilemma is a prime example, where making a utilitarian (rational) purchasing judgements reflects prosocial behaviour that is anticipated to lead to a reduction in the number of deaths caused during unavoidable road accidents. The current thesis demonstrates further the lability in human preferences as a function of the contextual construction and accessibility of dilemma information (Tversky & Kahneman, 1981; Kusev et al., 2009, 2016). When presented with partial PT accessibility, people are not only generally less utilitarian in all judgement tasks but also demonstrate inconsistencies between their moral judgements and purchasing behaviours; they do not want to buy the AV that they judge to be the most moral (a moral hypocrisy). However, with full PT accessibility, people are more normative in their behaviour. In particular, people overall approve of harm-minimising utilitarian AVs, and they display consistent moral preferences across judgements tasks, behavioural elicitation methods, types

of involvement, and types of psychological processing. Thus, the current thesis contributes to descriptive theories of human judgement and decision-making with a novel example of how the presentation of information during moral scenarios can influence people's preferences.

This thesis further informs descriptive theories that account for the relationship between choices and judgements (e.g., Bazerman et al., 1999; Brehm, 1956; Hsee et al., 1999, Hsee & Zhang, 2010; Izuma et al., 2010; Lichtenstein & Slovic, 1971, 1973; Sharot et al., 2012; Slovic & Lichtenstein, 1968). For example, the preference reversal phenomenon (Lichtenstein & Slovic, 1971) and evaluability theory (Hsee et al., 1999; Hsee & Zhang, 2010) predict a dissociation between people's judgements and choices. For example, choices are relatively easy as decision-makers are able to compare directly the values in each attribute. In contrast, judgements are difficult to make as the attribute value comparison between options are not contextually available (each separate option has a single value per attribute). However, I found that full PT accessibility influenced participants choices and judgements and made the association between participants' judgements and choices strong. Moreover, according to free-choice theory predictions (Brehm, 1956; Sharot et al., 2012), participants making a difficult choice influenced their subsequent judgements. This is typically explained by cognitive dissonance theory where, once people commit to a difficult choice option, they tend to value and appreciate this option more than before; in other words, evidence for choice-induced change in preferences (e.g., Festinger, 1964; Sharot et al., 2012). In contrast, as the result in Chapter 4 revealed, I found evidence for a PT accessibility choice-induced change in moral judgements as opposed to simply a choice-induced change in judgements (e.g., Brehm, 1956; Izuma & Murayama, 2013; Sharot et al., 2012).

The current thesis also makes a novel contribution to accessibility theories by introducing PT accessibility. According to the accessibility literature, full accessibility to information means that all aspects of a moral dilemma do not need to be inferred through

reasoning, but are instead explicitly stated (Kusev et al., 2016). Therefore, when accessibility is partial, only some elements of a dilemma are stated explicitly whilst others must be inferred. This leads to a bias in decision-making, overemphasising some elements of the scenario over others (Kusev et al., 2016; Martin et al., 2017). However, what exactly constitutes full accessibility depends on the nature of the dilemma. For example, if moral dilemmas involve PT tasks and multiple agents, then participants should have access to all situational perspectives; particularly if their goal is to make an ethical decision. The current thesis therefore introduces a new type of accessibility to be employed in PT tasks and highlights the importance of accessible and unbiased information in making informed judgements and decisions.

This thesis has for the first time offered an accessibility dimensions to the PT literature. Many moral psychology experiments involve PT tasks (e.g., Batson et al., 2003; Bonnefon et al., 2016; Galinsky & Moskowitz, 2000; Ruby & Decety, 2004 to name a few), however no previous empirical research has investigated varying levels (full and partial) of accessibility to PT. Therefore, the current thesis has introduced a new methodological approach to studying the influence of PT on people's moral behaviour. Moreover, it is anticipated that PT accessibility could be implemented in other moral PT domains. For example, full PT accessibility could be applied to current social and societal issues such as prejudice, discrimination and domestic violence (Galinsky & Ku, 2014; Galinsky & Moskowitz, 2000; Seinfeld et al., 2018). However, these opportunities for applying full PT accessibility should be methodologically developed and empirically tested before relevant interventions are developed.

A final element of the decision-making literature that this thesis contributes to is UTT. According to UTT theorists, judgements and decisions are determined by the type of psychological processing that the decision-maker engages in after being exposed to choice-

options (Dijksterhuis & Nordgren, 2006). The main finding of UTT is that processing of complex decision options leads to better decisions (for example, greater post-satisfaction with product choices, Messner & Wänke, 2011). In the current thesis I argue that the decision-making information, and not the type of processing employed, strongly influences people's decisions. Accordingly, the findings from the current thesis suggest that if decision-making information is accessible (e.g., offers full PT accessibility) then the type of psychological processing that the decision-maker employs has no effect on people's elicited preferences.

6.3.1.3 AV Ethics: Policymakers and Car Manufacturers' Dilemma

The current thesis has important practical implications for AV policymakers and car manufacturers. For both legal and safety reasons, AV ethical algorithms will need to be embedded into AVs in order to guide AV behaviour when the car senses an imminent collision (Maurer et al., 2016). However, according to Bonnefon et al.'s (2016) findings, car manufacturers may need to make a choice between algorithms that minimise harm (utilitarian) or algorithms that satisfy their consumers (passenger-protective). Accordingly, authors have warned that despite utilitarian AVs being prosocial life-saving vehicles, it is unlikely that they will be adopted by the public (Bonnefon et al., 2016; Greene, 2016; Shariff et al., 2017). However, in a commentary, Shariff et al. (2017) noted that the way in which utilitarian AVs are described could have a positive impact on public adoption. Suitably, this thesis provides an explanation for car manufacturers and policymakers as to why consumers are unwilling to purchase pro-social life-saving vehicles. I found that consumer preferences for passenger-protective AVs is simply an artefact of providing participants with partial PT accessibility to crash scenarios. Accordingly, a scenario presented with partial PT accessibility biases the respondent to one perspective of the situation, and consequently overemphasises the risk of owning a utilitarian AV. Therefore, participants' recorded purchasing behaviour in made in response to partially accessibly scenarios do not reflect their actual preferences. This thesis

accordingly, demonstrates the opposite; participants want to buy utilitarian AVs over passenger-protective models when they have full access to all inevitable perspectives in AV crash scenarios. Accordingly, the results in my thesis offer practical opportunities for designing successful interventions (use of full PT accessibility) that might be adopted by policymakers and car manufacturers and support their efforts in promoting pro-social life-saving vehicles. However, these potential future interventions (based on full PT accessibility) should be tested and results analysed.

The present thesis offers insights for designing successful interventions (based on full PT accessibility). If a utilitarian policy is put forward, this thesis offers evidence that people might approve utilitarian AVs. Therefore, a utilitarian policy will adhere to the prosocial harm preventing goals of automated technology and abide by consumers' preferences and safety concerns. Moreover, this thesis also offers ways in which car manufacturers can accordingly market the ethical components of their product, such as making PT information accessible in promotional materials, advertisements and car specification brochures.

6.3.2 Limitations and Future Work

No research in any scientific field exists without its limitations. Whilst limitations may reveal fundamental issues that weaken the credibility of the research findings altogether, limitations also open new avenues of research, prompting a continued pursuit in the advancement of knowledge. Fittingly, in this final subsection of the thesis, I not only offer limitations associated with the experimental explorations made in the current thesis, but also offer ways that these limitations can be addressed in future projects.

One major limitation of employing experimental psychology research is that the behaviour observed in experimental settings may not be representative of real-life behaviour. For example, in this thesis, all experiments involved measuring participants' moral behaviours in response to descriptive hypothetical moral scenarios. Accordingly, when experiencing a real

moral trade-off, participants may behave differently. A recent publication in Psychological Science argues that utilitarian preferences made in response to hypothetical moral dilemmas are not predictive of utilitarian preferences made in response to real moral dilemmas (Bostyn et al., 2018). In particular, people are more utilitarian in response to real moral scenarios than they are to identical hypothetical versions (see also Patil et al., 2014). However, since people are more utilitarian across all judgement tasks when they receive full PT accessibility to hypothetical scenario's, it is plausible that there would be little difference between hypothetical moral scenarios containing full PT accessibility and real-life moral scenarios. For instance, previous experimental research has demonstrated that the inconsistency between people's hypothetical moral choices and actual moral choices was reduced when participants received hypothetical scenarios that contained enriched contextual information (FeldmanHall et al., 2012). Accordingly, enriched contextual information involved making more contextual information available to participants than was available in the standard scenario. PT accessibility on the other hand, not only offers information that is available but also fully accessible, therefore one could predict similar if not enhanced version of FeldmanHall et al.'s (2012) findings. In particular, hypothetical scenarios that contain full PT accessibility may elicit the same moral preferences as real moral scenarios. Considering that exposing participants to a real moral scenario would be problematic for both ethical and practical reasons, it would be difficult to test this claim. However, some moral decision-making theorists have made use of virtual reality technology in an attempt to enhance the realism of hypothetical moral scenarios (Faulhaber et al., 2019; Patil et al., 2014). Therefore, in future research project I can investigate how responses to hypothetical moral scenarios involving partial PT accessibility and full PT accessibility each compare to real (VR) moral dilemmas.

Another limitation of the current thesis is that PT accessibility has been applied to a limited context: hypothetical AV crash scenarios. Whilst this context is well suited for the goals

of the present thesis, it is not yet clear whether (or indeed how) PT accessibility will influence behaviour in different tasks. For example, in this thesis PT accessibility has been found to induce utilitarian behaviour, which in this context is prosocial. However, it is not yet clear how PT accessibility influences other types of moral behaviour and affect such as fairness (Batson et al., 1997b), empathy (Stotland, 1969), or the propensity to cause to other people harm for personal gain (FeldmanHall, 2012). Therefore, future research projects could utilise PT accessibility in other moral experimental paradigms. For example, PT accessibility could be applied to the pain versus gain paradigm (FeldmanHall et al., 2012). In this paradigm participants are given a sum of money which will be multiplied by 100 if they administer electric shocks to another (confederate) participant. However, they can spend this money in order to spare the confederate from receiving painful shocks. Accordingly, in hypothetical versions of this experiment, PT accessibility could be manipulated in order to establish whether this influences people's propensity to administer electric shocks for personal monetary gain.

In addition to the potential limited context of this thesis, the sample itself was also limited in that it only contained participants residing in the UK. Accordingly, the findings obtained in this thesis may not be generalisable to people from non-western cultures. For example, previous world-wide empirical research exploring participants moral preferences regarding AV collisions (Awad et al., 2018) has revealed differences in moral choices across individuals from different global clusters (countries separated into western, eastern and southern clusters). Therefore, having not collected data from non-western non-UK sample, it is not yet clear how PT accessibility may influence peoples' utilitarian moral preferences in non-western global clusters. This limitation accordingly opens a new avenue of research related to whether variations in PT accessibility demonstrates the same pattern of behaviour in people across the globe.

It is also important to note that the sample is limited to an adult population. Whilst moral preferences related to AV crash scenarios may not be particularly relevant to children and adolescents, the application of PT in other moral contexts (e.g., discrimination, fairness and bullying) would make an interesting line of research. However due to the sample in this thesis being limited to the adult population, the current findings cannot predict how PT accessibility will influence moral behaviour in children. Kohlberg's (1969, 1973, 1981) theoretical and experimental work in moral development in children would suggest that PT accessibility may work differently in influencing children's moral behaviours than it does in adults. In future research projects, it would accordingly be interesting to establish how full accessibility to PT tasks can influence various moral behaviours in children and adolescents. Although the scope of this research was not to investigate the influence of individual differences on moral decision-making, future research could explore how gender, age, and cultural differences interact with PT accessibility and moral decision-making.

A final limitation of the current thesis is that the motivation for prosocial behaviour in full PT accessibility conditions is not yet clear. Whilst full PT accessibility results in prosocial decisions, there is at current, no evidence that people are demonstrating this preference for prosocial reasons. It is perhaps more plausible that offering full PT accessibility results in people redirecting their already egocentric motivation away from non-utilitarian choices and towards the most prosocial utilitarian choices (a choice that benefits everyone *including themselves*). For example, previous PT research has indicated that a PT led to a reduction in stereotyping behaviour and ingroup favouritism, but this effect was moderated by people's egocentric as opposed to prosocial motivation (Galinsky & Ku, 2004). In particular, the authors found that people with high self-esteem were more prosocial after undergoing a PT task than people with lower self-esteem. In the current context, when eliminating the bias in PT, participants have access to the *big picture,* where in the real world, even if they own an AV,

they will inevitably be pedestrians too (as soon as they exit their vehicle). Thus, understanding and appreciating that they can be harmed in either situation may simply direct them towards making the best decisions for them, which is utilitarian. Therefore, future research could investigate the motivations for prosocial behaviour when people have access to full PT accessibility.

6.3.3 Final Concluding Remarks

In this thesis, I have introduced PT accessibility as a new method of PT and a new type of accessibility to information. Accordingly, by implementing PT accessibility in moral dilemmas, I have attempted to address a contemporary moral problem that AV manufactures and policymakers are currently faced with; people do not want to buy the utilitarian AV that they judged to be the most morally appropriate for societal use (Bonnefon et al., 2016). In particular, I argued that in PT tasks employed in Bonnefon et al.'s (2016) experimental paradigm offered partial PT – biasing respondents towards non-utilitarian behaviour. I therefore predicted and found that enhancing PT in moral scenarios eliminates the moral hypocrisy (inconsistency) between people's moral preferences and purchasing behaviours. Moreover, full PT accessibility also results in consistent utilitarian behaviour across behavioural elicitation methods, types of involvement and types of psychological processing. The findings and the novel experimental method employed in this thesis informs both moral philosophy, moral psychology as well as normative and descriptive theories of decision-making (von Neumann & Morgenstern, 1944; Kahneman & Tversky, 1979; Tversky & Kahneman, 1992). Moreover, this thesis also offers insights for the development and design of interventions that might be implemented by AV car manufacturers and policymakers. Finally, the limitations related to experimental methodologies employed in this thesis open many new avenues of psychological research.

References

Acker, F. (2008). New findings on unconscious versus conscious thought in decision making: additional empirical data and meta-analysis. *Judgment and Decision Making, 3*(4), 292-303.

Allais, M. (1979). The foundations of a positive theory of choice involving risk and a criticism of the postulates and axioms of the American School (1952). In G. M. Hagen & M. Allais (Eds.), *Expected utility hypotheses and the Allais paradox* (pp. 27-145). Dordrecht, Holland: Reidel.

Annas, J. (1987). Epicurus on pleasure and happiness. *Philosophical Topics, 15*(2), 5-21.

Ariely, D. (1998). Combining experiences over time: The effects of durations, intensity changes on on-line measurements of retrospective pain evaluations. *Journal of Behavioral Decision making, 11*, 19-45.

Awad, E., Desouza, S., Kim, R., Schulz, J., Heinrich, J., Shariff, A., Bonnefon, J-F., & Rahwan, I. (2018). The moral machine experiment. *Nature, 563*(7729), 59-64.

Ayton, P. (2005). Judgement and decision-making. In N. Braisby & A. Gellatly (Eds.), *Cognitive Psychology* (p. 391). New York, NY: Oxford University Press.

Barnett, M. A., & Thompson, S. (1985). The role of perspective taking and empathy in children's Machiavellianism, prosocial behaviour and motive for helping. *The Journal of Genetic Psychology, 146*(3), 295-305.

Batson, C. D., Chang, J., Orr., R., & Rowland, J. (2002). Empathy, attitudes and actions: Can feeling for a member of a stigmatized group motivate one to help the group? *Personality and Social Psychology Bulletin, 28*(12), 1656-1666.

Batson, C. D., Early, S., & Salvarani, G. (1997b). Perspective taking: Imagining how others feel versus how you would feel. *Personality and Social Psychology Bulletin, 23*(7), 751-758.

Batson, C. D., Kobrynowicz, D., Dinnerstein, J. L., Kampf, H. C., & Wilson, A. D. (1997a). In a very different voice: unmasking moral hypocrisy. *Journal of Personality and Social Psychology, 72*(6), 1335-1348.

Batson, C. D., Lishner, D. A., Carpenter, A., Dulin, L., Harjusola-Webb, S., Stocks, E. L., … Sampat, B. (2003). "…As you would have them do unto you": Does imagining yourself in the other's place stimulate moral action? *Personality and Social Psychology Bulletin, 29*(9), 1190-1201.

Batson, C. D., Thompson, E. R., Seuferling, G., Whitney, H., & Strongman, J. A. (1999). Moral hypocrisy: Appearing moral to oneself without being so. *Journal of Personality and Social Psychology, 77*(3), 525-537.

Batson, C. D., & Thompson, E. R. (2001). Why don't moral people act morally? Motivational considerations. *Current Directions in Psychological Science, 10*(2), 54-57.

Batson, C. D. (2011). What's wrong with morality? *Emotion Review, 3*(3), 230-236.

Bauman, C. W., McGraw, A. P., Bartels, D. M., & Warren, C. (2014). Revisiting external validity: Concerns about trolley problems and other sacrificial dilemmas in moral psychology. *Personality and Social Psychology Compass, 8*(9), 536-554.

Bazerman, M. H., Moore, D. A., Tenbrunsel, A. E., Wade-Benzoni, K. A., & Blount, S. (1999). Explaining how preferences change across joint versus separate evaluation. *Journal of Economic Behavior and Organization, 39*(1), 41-58.

Bentham, J. (1970). *An introduction to the principles of moral and legislation*. Darien, CT: Hafner. (Original work published 1789).

Bentham, J. (1988). *A fragment on government.* Cambridge: Cambridge University Press (Original work published 1776).

Bonnefon, J-F., Shariff, A., Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science, 325*(6293), 1573-1576.

Bos, M. W., Dijksterhus, A., & van Baaren, R. B. (2010). The benefits of "sleeping on things": Unconscious thought leads to automatic weighting. *Journal of Consumer Psychology, 21*(1), 4-8.

Bose, A., & Ioannou, P. (2003). Analysis of traffic flow with mixed manual and semiautomated vehicles. *IEEE Transactions on Intelligent Transportation Systems, 4*(4), 173-188.

Bostyn, D. H., Sevenhant, S., & Roets, A. (2018). Of mice, men and trolleys: Hypothetical judgments versus real-life behaviour in trolley-style moral dilemmas. *Psychological Science, 29*(7), 1084-1093.

Brehm, J. W. (1956). Postdecision changes in the desirability of alternatives. *Journal of Abnormal and Social Psychology, 52*(3), 348-389.

Brink, D. O. (2006). Some forms and limits of consequentialism. In D. Copp (Ed.). *The oxford handbook of ethical theory* (pp. 380-423). New York, NY: Oxford University Press.

Broad, C. D. (1971). Self and others. In D. R. Cheney (Ed.) *Broads critical essays in moral philosophy* (pp. 262-282). London: Allen and Unwin.

Chen, M. K., & Risen, J. L. (2010). How choice affects and reflects preferences: Revisiting the free-choice paradigm. *Journal of Personality and Social Psychology, 99*(4), 573-594.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. Hillsdale, NJ: Erlbaum.

Colby, A. B., Gibbs, J., Kohlberg, L., Speicher-Dubin, B., & Candee, D. (1980). *Measurements of moral judgment* (Vol. I.). Cambridge, MA: Centre for Moral Education, Harvard University.

Contissa, G., Lagioia, F., & Sartor, G. (2017). The ethical knob: Ethically-customisable automated vehicles and the law. *Artificial Intelligence and Law, 25*(3), 365-378.

Crockett, M. (2013). Models of morality. *Trends in Cognitive Science, 17*(8), 363-366.

Dijksterhuis, A., Bos, M. W., Nordgren, L. F., & van Baaren, R. B. (2006). On making the right choice: the deliberation-without-attention effect. *Science, 311*(5763), 1005-1007.

Dijksterhuis, A., Bos, M. W., van der Leij, A. & van Baaren, R. B. (2009). Predicting soccer matches after unconscious and unconscious thought as a function of expertise. *Psychological Science, 20*(11), 1381-1387.

Dijksterhuis, A., & Nordgren, L. F. (2006). A theory of unconscious thought. *Perspectives on Psychological Science, 1*(2), 95-109.

Dijksterhuis, A., & Strick, M. (2016). A case for thinking without consciousness. *Perspectives on Psychological Science, 11*(1), 117-132.

Edgeworth, F. Y. (1877). *New and old methods of ethics*. Oxford: James Parker.

Edwards, W. (1955). The prediction of decisions among bets. *Journal of Experimental Psychology, 50*(3)*,* 201-214.

Fagnant, D. J., & Kockelman, K. (2015). Preparing a nation for autonomous vehicles: barriers and policy recommendations. *Transportation Research Part A: Policy and Practice, 77*, 167-181.

Faulhaber, A. K., Dittmer, A., Blind, F., Wächter, M. A., Timm, S., Stephan, A., … König, P. (2019). Human decisions in moral dilemmas are largely described by utilitarianism: virtual car driving study provides guidelines for autonomous driving vehicles. *Science and Engineering Ethics, 25*(2), 399-418.

Feigenbaum, E. A. (1961). The simulation of verbal learning behaviour. In E. A. Feigenbaum & J. Feldman (Eds.), *Computers and Thought*. New York: McGraw-Hill.

FeldmanHall, O., Mobbs, D., Evans, D., Hiscox, L., Navrady, L., & Dalgleish, T. (2012). What we say and what we do: The relationship between real and hypothetical moral choices. *Cognition, 123*(3), 434-441.

Festinger, L. (1964). *Conflict, decision and dissonance*. Stanford, CA: Stanford University Press.

Fleetwood, J. (2017). Public health, ethics, and autonomous vehicles. *American Journal of Public Health, 107*(4), 532-537.

Foot, P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford Review, 5*, 5-15.

Galinsky, A. D., & Ku, G. (2004). The effects of perspective-taking on prejudice: The moderation role of self-evaluation. *Personality and Social Psychology Bulletin, 30*(5), 594-604.

Galinsky, A. D., & Moskowitz, G. B. (2000). Perspective-taking: Decreasing stereotype expression, stereotype accessibility, and ingroup favouritism. *Journal of Personality and Social Psychology, 78*(4), 708-724.

Garrigan, B., Adlam, A. L. R., & Langdon, P. E. (2016). The neural correlates of moral decision-making: A systematic review and meta-analysis of moral evaluations and response decision judgements. *Brain and Cognition, 108*, 88-97.

Gerard, H. B., & White, G. L. (1983). Post-decisional reevaluation of choice alternatives. *Personality and Social Psychology Bulletin, 9*(3), 365-369.

Gogoll, J., & Müller, J. F. (2017). Autonomous cars: In favor of mandatory ethics. *Science and Engineering Ethics, 23*(3), 681-700.

Gold, N., Pulford, B. D., & Colman, A. M. (2013). Your money or your life: Comparing judgements in trolley problems involving economic and emotional harms, injury, and death. *Economics and Philosophy, 29*(2), 213-233.

Goodall, N. J. (2014). Ethical decision making during automated vehicle crashes. *Transportation Research Record: Journal of the Transportation Research Board, 2424*(1), 58-65.

Greene, J. D. (2015). Beyond point-and-shoot morality: why cognitive (neuro)science matters for ethics. *Law & Ethics of Human Rights, 9*(2), 141-172.

Greene, J. (2016). Our driverless dilemma. *Science, 352*(6293), 1514-1515.

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition, 111*(3), 364-371.

Greene, J. D., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Sciences, 6*(12), 517-523.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 14*(293), 2105-2108.

Groom, V., Bailenson, J. N., & Nass, C. (2008). The influence of racial embodiment on racial bias in immersive virtual environments. *Social Influence, 4*(3), 231-248.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgement. *Psychological Review, 108*(4), 814-834.

Ham, J., & van den Bos, K. (2010). On unconscious morality: The effects of unconscious thinking on moral decision making. *Social Cognition, 28*(1), 74-83.

Hayes, F. (2017). Introduction to mediation, moderation and conditional process analysis: A regression-based approach (2nd Ed.). New York, NY: The Guildford Press.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rate events in risky choice. *Psychological Science, 15*(8), 534-539.

Hertwig, R., & Erev, I. (2009). The description-experience gap in risky choice. *Trends in Cognitive Science, 13(*12), 517-523.

Higgins, T. E. (1996). Knowledge activation: Accessibility, applicability and salience. In E. T. Higgins & A. Kruglanski (Eds.). *Social Psychology: Handbook of basic principles* (pp. 133-168). New York: Guilford Press.

Hsee, C. K. (1996). The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational Behaviour and Human Decision Processes, 67*(3), 247-257.

Hsee, C. K., Loewenstein, G. F., Blount, S., & Bazerman, M. (1999). Preference reversals between joint and separate evaluations of options: A review and theoretical analysis. *Psychological Bulletin, 125*(5), 576-586.

Hsee, C. K., & Zhang, J. (2010). General evaluability theory. *Perspectives on Psychological Science, 5*(4), 343-355.

Hume, D. (1969). *A treatise of human nature*. London: Penguin. (Original work published 1739-1740).

Huebner, B., & Hauser, M. D. (2011). Moral judgments about altruistic self-sacrifice: When philosophical and folk intuitions clash. *Philosophical Psychology, 24*(1), 73-94.

Hutcheson, F. (1742). *An essay on the nature and conduct of the passions and affections: with illustrations on the moral sense.* (3rd ed.). London, UK: Ward. (First edition published 1728).

Izuma, K., Matsumoto, M., Murayama, K., Samejima, K., Sadato, N., & Matsumoto, K. (2010). Neural correlates of cognitive dissonance and choice-induced preference change. *Proceedings of the National Academy of Sciences, 207*(51), 22014-22019.

Izuma, K., & Murayama, K. (2013). Choice-induced preference change in the free-choice paradigm: A critical methodological review. *Frontiers in Psychology, 4*(41), 1-12.

Johnson, D. W. (1975). Cooperativeness and social perspective taking. *Journal of Personality and Social Psychology, 31*(2), 241-244.

Kahane, G. (2013). The armchair and the trolley: an argument for experimental ethics. *Philosophical Studies, 162*(2), 421-455.

Kahneman, D. (2003). A perspective on judgement and choice: Mapping bounded rationality. *American Psychologist, 58*(9), 697-720.

Kahneman, D., & Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica, 47*(2), 263-291.

Kahneman, D., Wakker, P. P., & Sarin, R. (1997). Back to Bentham? Explorations of experienced utility. *Quarterly Journal of Economics, 112*(2), 375-405

Kant, I. (2002). *Foundations of the metaphysics of morals.* New Haven and London: Yale University Press (Original work published 1785).

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature, 446*(7138), 908-911.

Koenigs, M., Kruepke, M., Zeier, J., & Newman, J. (2011). Utilitarian moral judgments in psychopathy. *Social Cognitive and Affective Neuroscience, 7*(6), 708-714.

Kohlberg, L. (1969). Stage and sequence: The cognitive-developmental approach to socialization. In D. A. Goslin (Ed.). *Handbook of socialization theory and research* (pp. 348-480). Chicago: Rand McNally.

Kohlberg, L. (1973). Stages and aging in moral development – Some speculations. *The Gerontologist, 13*(4), 497-502.

Kohlberg, L. (1981). *Essays in moral development: The psychology of moral development* (Vol.2). San Francisco: Harper & Row.

Kröger, F. (2015). Automated driving in its social, historical and cultural contexts. In M. J. Maurer, C. Gerdes, B. Lenz & H. Winner (Eds.). *Autonomous driving* (pp. 41-67). Berlin, Germany: Springer Berlin Heidelberg.

Kusev, P., van Schaik, P., Alzahrani, S., Lonigro, S., & Purser, H. (2016). Judging the morality of utilitarian actions: how poor utilitarian accessibility makes judges irrational. *Psychonomic Bulletin and Review, 23*(6), 1961-1967.

Kusev, P., van Schaik, P., Ayton, P., Dent, J., & Chater, N. (2009). Exaggerated risk: Prospect theory and probability weighting in risky choice. *Journal of Experimental Psychology: Learning, Memory and Cognition, 35*(6), 1487-1505.

Kusev, P., van Schaik, P., Martin, R., Hall, L., & Johansson, P. (2020). Preference reversals during risk elicitation. *Journal of Experimental Psychology: General, 149*(3), 585-589.

Kusev, P., van Schaik, P., Tsaneva-Atanasova, K., Juliusson, A., & Chater, N. (2018). Adaptive anchoring model: How static and dynamic presentations of time series influence judgments and predictions. *Cognitive Science, 42*(1), 77-102.

Lamm, C., Batson, C. D., & Decety, J. (2007). The neural subtrate of human empathy: effects of perspective-taking and cognitive appraisal. *Journal of Cognitive Neuroscience, 19*(1), 42-58.

Lichtenstein, S., & Slovic, P. (1971). Reversals in preference between bids and choices in gambling decisions. *Journal of Experimental Psychology, 89*(1), 46-55.

Lichtenstein, S., & Slovic, P. (1973). Response-induced preference reversals in gambling: An extended replication in Las Vegas. *Journal of Experimental Psychology, 101*(1), 16-20.

Lieberman, D., Tooby, J., & Cosmides, L. (2003). Does morality have a biological basis? An empirical test of the factors governing moral sentiments relating to incest. *Proceedings of the Royal Society of London, Series B: Biological Sciences, 270*(1517), 819-826.

Litman, T. (2018). *Autonomous Vehicle Implementation Predictions: Implications for transport planning*. Victoria Transport Policy Institute. Victoria, Cananda.

Lönnqvist, J-E., Irlenbusch, B., & Walkowitz, G. (2014). Moral hypocrisy: Impression management or self-deception? *Journal of Experimental Social Psychology, 55*, 53-62.

Marshall, A. (1920). *Principles of Economics (8th ed.).* London: Macmillan & Co.

Marteau, T. M. (1989). Framing of information: Its influence upon decisions of doctors and patients. *British Journal of Social Psychology, 28*, 89-94.

Martin, R., Kusev, I., Cooke, A., Baranova, V., van Schaik, P., & Kusev, P. (2017). Commentary: The social dilemma of autonomous vehicles. *Frontiers in Psychology, 8*(808), 1-2.

Martin, R., & Kusev, P. (2016). Rational choice predicted by utility ratio and uncertainty. Paper presented at the annual meeting of the Society for Judgment and Decision Making, Boston, Massachusetts, USA. November 18th - 21st.

Martin, R., & Kusev, P. (2017). The influence of associative learning on moral decision-making. Paper presented at the 58th annual meeting of the Psychonomic Society, Vancouver, Canada. November 9th - 12th.

Martin, R., Kusev, P., & van Schaik, P. (under review). Learning morality: How moral-rule learning overrides expected utilitarian behaviour.

Maurer, M. J., Gerdes, C., Lenz, B., & Winner, H. (2016). *Autonomous driving*. Berlin, Germany: Springer Berlin Heidelberg.

McNeil, B. J., Pauker, S. G., Sox, H. C. Jr., & Tversky, A. (1982). On the elicitation of preferences for alternative therapies. *New England Journal of Medicine, 306*(21), 1259-62.

Messner, C., & Wänke, M. (2011). Unconscious information processing reduces information overload and increases product satisfaction. *Journal of Consumer Psychology, 21*, 9-13.

Meyer, J., Becker, H., Bösch, P. M., & Axhausen, K. W. (2017). Autonomous vehicles: The next jump in accessibilities? *Research in Transport Economics, 62*, 80-91.

Mikhail, J. (2007). Universal moral grammar: theory, evidence and the future. *Trends in Cognitive Science, 11*(4), 143-152.

Mill, J. S. (2014). *Utilitarianism*. Cambridge: Cambridge University Press. (Original work published 1863). doi: 10.1017/CBO9781139923927

Miller, D. Z. (2016, October). Mercedes-Benz's self-driving cars would choose passenger lives over bystanders. *Fortune Magazine*. Retrieved from: https://fortune.com/2016/10/15/mercedes-self-driving-car-ethics/

Nakamura, K. (2012). The Footbridge Dilemma Reflects More Utilitarian Thinking Than the Trolley Dilemma: Effect of Number of Victims in Moral Dilemmas. *Proceedings of the Thirty-fourth Annual Conference of the Cognitive Science Society, 34. Retrieved from https://escholarship.org/uc/item/8062w0px*

Nakamura, K. (2013). A closer look at moral dilemmas: Latent dimensions of morality and the difference between trolley and footbridge dilemmas. *Thinking & Reasoning, 19*(2), 178-204.

Negd, M., Mallan, K. M., & Lipp, O. V. (2011). The role of anxiety and perspective-taking strategy on affective empathetic response. Behaviour Research and Therapy, 49(12), 852-857.

Nucci, L. (1981). Conceptions of personal issues: A domain distinct or societal concepts. *Child Development, 52*(1), 114-121.

Nucci, L., & Turiel, E. (1978). Social interactions and the development of social concepts in preschool children. *Child Development, 49*, 400-407.

Nyholm, A., & Smids, J. (2016). The ethics of accident-algorithms for self-driving cars: an applied trolley problem? *Ethical Theory and Moral Practice, 19*(5), 1275-1289.

O'Niell, P., & Petrinovich, L. B. (1998). A preliminary cross-cultural study of moral intuitions. *Evolution and Human Behaviour, 19*(6), 349-367.

Patil, I., Cogoni, C., Zangrando, N., Chittaro, L., & Silani, G. (2014). Affective basis of judgment-behavior discrepancy in virtual experiences of moral dilemmas. *Social Neuroscience, 9*(1) 94-107.

Patil, I., & Silani, G. (2014). Reduced empathetic concern leads to utilitarian moral judgments in trait alexithymia. *Frontiers in Psychology, 5*(501), 1-12.

Pedroni, A., Frey, R., Bruhin, A., Dutilh, G., Hertwig, R., & Rieskamp, J. (2017). The risk elicitation puzzle. *Nature Human Behaviour, 1*(11), 803-809.

Piaget, J. (1932). *The moral judgment of the child*. London: Routledge and Kegan Paul.

Plamenatz, J. (1966). *The English utilitarians*. Oxford: Basil Blackwell & Mott Ltd.

Plunkett, D., & Greene, J. D. (2019). Overlooked evidence and a misunderstanding of what trolley dilemmas do best: Commentary on Bostyn, Sevenhant, and Roets (2018). Psychological Science (Advanced online publication).

Read, D. (2007). Experienced Utility: Utility theory from Jeremy Bentham to Daniel Kahneman. *Thinking & Reasoning, 13*(1), 45-61. doi: 10.1080/13546780600872627

Reinhard, M. A., Greifeneder, R., & Scharmach, M. (2013). Unconscious processing improves lie detection. *Journal of Personality and Social Psychology, 105*(5), 721-739.

Rest, J. R. (1986). *Moral development: Advances in research and theory*. Minneapolis: University of Minnesota Press.

Rest, J., Narvaez, D., Bebeau, M. J., & Thoma, S. J. (1999). *Postconventional moral thinking: A neo-Kohlbergian approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Ruby, P., & Decety, J. (2004). How would you feel versus how do you think she would feel? A neuroimaging study of perspective-taking with social emotions. *Journal of Cognitive Neuroscience, 16*(6), 988-999.

Savage, L. (1954). *The foundations of statistics.* New York: Wiley.

Scott, J. (2000). Rational choice theory. In G. Browning, A. Halcli, N. Hewlett & F. Webster (Eds.), *Understanding contemporary society: Theories of the present* (pp. 126-138), London: Sage.

Seinfeld, S., Arroyo-Palacios, J., Iruretagoyena, G., Hortensius, R., Zapata, L. E., Borland, D., de Gelder, B., Slater, M., & Sanchez-Vives, M. V. (2018). Offenders become the victim in virtual reality: impact of changing perspective in domestic violence. *Scientific Reports, 8*(2692), 1-11.

Selman, R. L. (1971). The relation of role taking to the development of moral judgment in children. *Child Development, 42*(1), 79-91.

Sen, A. (1980-1981). Plural utility. *Proceedings of the Aristotelian Society, New Series, 81*, 193-215.

Shariff, A., Bonnefon, J-F., & Rahwan, I. (2017). Psychological roadblocks to the adoption of self-driving vehicles. *Nature Human Behaviour, 1*(10), 694-696.

Sharot, T., Fleming, S. M., Yu, X., Koster, R., & Dolan, R. J. (2012). Is choice-induced preference change long lasting? *Psychological Science, 23*(10), 1123-1129.

Shih, M., Wang, E., Bucher, T. A., & Stotzer, R. (2009). Perspective-taking: Reducing prejudice towards general outgroups and specific individuals. *Group Processes & Intergroup Relations, 12*(5), 565-577.

Skinner, B. F. (1971). *Beyond freedom and dignity*. New York, NY: Pelican Books.

Slovic, P., & Lichtenstein, S. (1968). Relative importance of probabilities and payoffs in risk taking. *Journal of Experimental Psychology, 78*(3), 1-18.

Slovic, P. (1995). The construction of preference. *American Psychologist, 50*(5), 364-371.

Smetana, J. G. (1999). The role of parents in moral development: A social domain analysis. *Journal of Moral Education, 28*(3), 311-321.

Smetana, J. G. (2006). Social-cognitive domain theory: Consistencies and variations in children's moral and social judgments. In M. Killen & J. G. Smetana (Eds.), *Handbook of moral development* (pp. 119-154). Mahwah: Erlbaum Associates.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioural and Brain Sciences, 23*, 645-726.

Stotland, E. (1969). Exploratory investigations of empathy. In L. Berkowitz (Ed.), Advances in experimental social psychology (Vol. 4, pp. 271-314). New York: Academic Press.

Strick, M., Dijksterhuis, A., Bos, M. W., Sjoerdsma, A., van Baaren, R. B., & Nordgren, L. F. (2011). A meta-analysis on unconscious thought effects. *Social Cognition, 29*(6), 738-762.

Strick, M., Dijksterhuis, A., & van Baaren, R. B. (2010). Unconscious-thought effects take place off-line, not on-line. *Psychological Science, 21*(4), 484-488.

Sugden, R. (1991). Rational choice: A survey of contributions from economics and philosophy. *The Economic Journal, 101*(407), 751-785.

Sunstein, C. R. (2005). Moral heuristics. *Behavioral and Brain Sciences, 28*(4), 542-573.

Tassy, S., Oullier, O., Mancini, J., & Wicker, B. (2013). Discrepancies between judgement and choice of action in moral dilemmas. *Frontiers in Psychology, 4*, 250.

Thomson, J. J. (1985). The trolley problem. *The Yale Law Journal, 94*, 1395-1415.

Troyer, J. (2003). *The classical utilitarians: Bentham and Mill.* Indianapolis, MA: Hackett.

Tulving, E., & Pearlstone, Z. (1966). Availability versus accessibility of information in memory for words. *Journal of Verbal Learning and Verbal Behavior, 5*(4), 381-391.

Turiel, E. (1983). *The development of social knowledge: Morality and convention*. Cambridge: Cambridge University Press.

Tverksy, A. (1969). Intransitivity in preference. *Psychological Review, 76*(1), 31-48.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*(2), 507-232.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*(4481), 453-458.

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A preference-dependent model. *The Quarterly Journal of Economics, 106*(4), 1039-1061.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: cumulative representations of uncertainty. *Journal of Risk and Uncertainty, 5*(4), 297-232.

Von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton: Princeton University Press.

Wadud, Z., MacKenzie, D., & Leiby, P. (2016). Help or hindreance? The travel, energy and carbon impacts of highly automated vehicles. *Transportation Research Part A, 86*, 1-18.

Waldmann, M. R., & Dieterich, J. H. (2007). Throwing a bomb on a person versus throwing a person on a bomb: Intervention myopia in moral intuitions. *Psychological Science, 18*(3), 247-253.

Walker, L. J. (1980). Cognitive and perspective-taking prerequisites for moral development. *Child Development, 51*(1), 131-139.

West, H. R. (2004). *An introduction to Mills utilitarian ethics*. Cambridge: Cambridge University Press.

Appendix A: Experimental Materials (Scenarios and Questions)

Experiment 1: Partial Contextual Accessibility / Participant Involvement

*You are the sole passenger in an autonomous self-driving vehicle travelling at the speed limit down a main road. Suddenly, 10 pedestrians appear ahead, in the direct path of the car. The car could be programmed to either **SWERVE** off to the side of the road, where it will impact a barrier, killing you or **STAY** on its current path where it will kill the 10 pedestrians (see the picture illustrating this scenario).*

*Judge the moral appropriateness of programming autonomous self-driving cars to: **SWERVE***

Not at all appropriate **Definitely appropriate**
[----------]

*Judge the moral appropriateness of programming autonomous self-driving cars to: **STAY***

Not at all appropriate **Definitely appropriate**
[----------]

*How would you rate your willingness to BUY an autonomous self-driving car programmed to: **SWERVE?***

Not at all willing **Definitely willing**
[----------]

*How would you rate your willingness to BUY an autonomous self-driving car programmed to: **STAY?***

Not at all willing **Definitely willing**
[----------]

Experiment 1: Full Contextual Accessibility / Participant Involvement

*You are the sole passenger in an autonomous self-driving vehicle travelling at the speed limit down a main road. Suddenly, 10 pedestrians appear ahead, in the direct path of the car. The car could be programmed to either **SWERVE** off to the side of the road, where it will impact a barrier, killing you but leaving the 10 pedestrians unharmed or **STAY** on its current path where it will kill the 10 pedestrians, but you will be unharmed (see the picture illustrating this scenario).*

Judge the moral appropriateness of programming autonomous self-driving cars to: ***SWERVE****, killing you but leaving the 10 pedestrians unharmed. If the car does not swerve it will kill the 10 pedestrians but leave you unharmed.*

Not at all appropriate **Definitely appropriate**

[----------]

Judge the moral appropriateness of programming autonomous self-driving cars to: ***STAY,*** *killing the 10 pedestrians but leaving you unharmed. If the car does not stay, it will kill you but leave the 10 pedestrians unharmed.*

Not at all appropriate **Definitely appropriate**

[----------]

How would you rate your willingness to BUY an autonomous self-driving car programmed to: ***SWERVE?*** *killing you but leaving the 10 pedestrians unharmed. If the car does not swerve it will kill the 10 pedestrians but leave you unharmed.*

Not at all willing **Definitely willing**

[----------]

How would you rate your willingness to BUY an autonomous self-driving car programmed to: ***STAY?*** *killing the 10 pedestrians but leaving you unharmed. If the car does not stay, it will kill you but leave the 10 pedestrians unharmed.*

Not at all willing **Definitely willing**

[----------]

Experiment 1: Partial Contextual Accessibility / Stranger Involvement

Sam is the sole passenger in an autonomous self-driving vehicle travelling at the speed limit down a main road. Suddenly, 10 pedestrians appear ahead, in the direct path of the car. The car could be programmed to either ***SWERVE*** *off to the side of the road, where it will impact a barrier, killing Sam or* ***STAY*** *on its current path where it will kill the 10 pedestrians (see the picture illustrating this scenario).*

Judge the moral appropriateness of programming autonomous self-driving cars to: ***SWERVE***

Not at all appropriate **Definitely appropriate**

[----------]

Judge the moral appropriateness of programming autonomous self-driving cars to: ***STAY***

Not at all appropriate **Definitely appropriate**

[----------]

How would you rate your willingness to BUY an autonomous self-driving car programmed to: ***SWERVE?***

Not at all willing **Definitely willing**

[----------]

How would you rate your willingness to BUY an autonomous self-driving car programmed to: ***STAY?***

Not at all willing **Definitely willing**

[----------]

Experiment 1: Full Contextual Accessibility / Stranger Involvement

Sam is the sole passenger in an autonomous self-driving vehicle travelling at the speed limit down a main road. Suddenly, 10 pedestrians appear ahead, in the direct path of the car. The car could be programmed to either ***SWERVE*** *off to the side of the road, where it will impact a barrier, killing Sam but leaving the 10 pedestrians unharmed or* ***STAY*** *on its current path where it will kill the 10 pedestrians, but Sam will be unharmed (see the picture illustrating this scenario).*

Judge the moral appropriateness of programming autonomous self-driving cars to: ***SWERVE****, killing Sam but leaving the 10 pedestrians unharmed. If the car does not swerve it will kill the 10 pedestrians but leave Sam unharmed.*

Not at all appropriate **Definitely appropriate**

[----------]

Judge the moral appropriateness of programming autonomous self-driving cars to: ***STAY,*** *killing the 10 pedestrians but leaving Sam unharmed. If the car does not stay, it will kill Sam but leave the 10 pedestrians unharmed.*

Not at all appropriate **Definitely appropriate**

[----------]

*How would you rate your willingness to BUY an autonomous self-driving car programmed to: **SWERVE?** killing Sam but leaving the 10 pedestrians unharmed. If the car does not swerve it will kill the 10 pedestrians but leave Sam unharmed.*

Not at all willing **Definitely willing**
[----------]

*How would you rate your willingness to BUY an autonomous self-driving car programmed to: **STAY?** killing the 10 pedestrians but leaving Sam unharmed. If the car does not stay, it will kill Sam but leave the 10 pedestrians unharmed.*

Not at all willing **Definitely willing**
[----------]

Experiment 2: Partial PT Accessibility / Participant Involvement

*You are the sole passenger in an autonomous self-driving vehicle travelling at the speed limit down a main road. Suddenly, 10 pedestrians appear ahead, in the direct path of the car. The car could be programmed to either **SWERVE** off to the side of the road, where it will impact a barrier, killing you but leaving the 10 pedestrians unharmed or **STAY** on its current path where it will kill the 10 pedestrians, but you will be unharmed (see the picture illustrating this scenario).*

*Judge the moral appropriateness of programming autonomous self-driving cars to: **SWERVE**, killing you but leaving the 10 pedestrians unharmed. If the car does not swerve it will kill the 10 pedestrians but leave you unharmed.*

Not at all appropriate **Definitely appropriate**
[----------]

*Judge the moral appropriateness of programming autonomous self-driving cars to: **STAY,** killing the 10 pedestrians but leaving you unharmed. If the car does not stay, it will kill you but leave the 10 pedestrians unharmed.*

Not at all appropriate **Definitely appropriate**
[----------]

Experiment 2: Full PT Accessibility / Participant Involvement

*You could be the sole passenger in an autonomous self-driving vehicle travelling at the speed limit down a main road. Or you could be one of the 10 pedestrians that have appeared ahead, in the direct path of the car. The car could be programmed to either **SWERVE** off to the side of the road, where it will impact a barrier, killing the passenger (that could be you) but leaving the 10 pedestrians unharmed or **STAY** on its current path where it will kill the 10 pedestrians (that could include you), but the passenger will be unharmed (see the picture illustrating this scenario).*

*Judge the moral appropriateness of programming autonomous self-driving cars to: **SWERVE**, killing the sole passenger (this could be you) but leaving the 10 pedestrians unharmed (this could include you). If the car does not swerve it will kill the 10 pedestrians (this could include you) but leave the sole passenger unharmed (this could be you).*

Not at all appropriate **Definitely appropriate**

[----------]

*Judge the moral appropriateness of programming autonomous self-driving cars to: **STAY**, killing the 10 pedestrians (this could be you) but leaving the sole passenger unharmed (this could include you). If the car does not stay, it will kill the sole passenger (this could be you) but leave the 10 pedestrians unharmed (this could include you).*

Not at all appropriate **Definitely appropriate**

[----------]

Experiment 2: Partial PT Accessibility / Stranger Involvement

*Sam is the sole passenger in an autonomous self-driving vehicle travelling at the speed limit down a main road. Suddenly, 10 pedestrians appear ahead, in the direct path of the car. The car could be programmed to either **SWERVE** off to the side of the road, where it will impact a barrier, killing Sam but leaving the 10 pedestrians unharmed or **STAY** on its current path where it will kill the 10 pedestrians, but Sam will be unharmed (see the picture illustrating this scenario).*

*Judge the moral appropriateness of programming autonomous self-driving cars to: **SWERVE**, killing you Sam leaving the 10 pedestrians unharmed. If the car does not swerve it will kill the 10 pedestrians but leave Sam unharmed.*

Not at all appropriate **Definitely appropriate**
[----------]

*Judge the moral appropriateness of programming autonomous self-driving cars to: **STAY,** killing the 10 pedestrians but leaving Sam unharmed. If the car does not stay, it will kill Sam but leave the 10 pedestrians unharmed.*

Not at all appropriate **Definitely appropriate**
[----------]

Experiment 2: Full PT Accessibility / Stranger Involvement

*Sam could be the sole passenger in an autonomous self-driving vehicle travelling at the speed limit down a main road. Or Sam could be one of the 10 pedestrians that have appeared ahead, in the direct path of the car. The car could be programmed to either **SWERVE** off to the side of the road, where it will impact a barrier, killing the passenger (that could be Sam) but leaving the 10 pedestrians unharmed or **STAY** on its current path where it will kill the 10 pedestrians (that could include Sam), but the passenger will be unharmed (see the picture illustrating this scenario).*

*Judge the moral appropriateness of programming autonomous self-driving cars to: **SWERVE**, killing the sole passenger (this could be Sam) but leaving the 10 pedestrians unharmed (this could include you). If the car does not swerve it will kill the 10 pedestrians (this could include Sam) but leave the sole passenger unharmed (this could be Sam).*

Not at all appropriate **Definitely appropriate**
[----------]

*Judge the moral appropriateness of programming autonomous self-driving cars to: **STAY**, killing the 10 pedestrians (this could be Sam) but leaving the sole passenger unharmed (this could include Sam). If the car does not stay, it will kill the sole passenger (this could be Sam) but leave the 10 pedestrians unharmed (this could include Sam).*

Not at all appropriate **Definitely appropriate**

[----------]

Experiment 3: Partial PT Accessibility / Participant Involvement

*You are the sole passenger in an autonomous self-driving vehicle travelling at the speed limit down a main road. Suddenly, 10 pedestrians appear ahead, in the direct path of the car. The car could be programmed to either **SWERVE** off to the side of the road, where it will impact a barrier, killing you but leaving the 10 pedestrians unharmed or **STAY** on its current path where it will kill the 10 pedestrians, but you will be unharmed (see the picture illustrating this scenario).*

*Judge the moral appropriateness of programming autonomous self-driving cars to: **SWERVE**, killing you but leaving the 10 pedestrians unharmed. If the car does not swerve it will kill the 10 pedestrians but leave you unharmed.*

Not at all appropriate **Definitely appropriate**

[----------]

*Judge the moral appropriateness of programming autonomous self-driving cars to: **STAY,** killing the 10 pedestrians but leaving you unharmed. If the car does not stay, it will kill you but leave the 10 pedestrians unharmed.*

Not at all appropriate **Definitely appropriate**

[----------]

Using a budget of £50,000, please indicate how much you would pay for the following autonomous self-driving cars (the entire £50,000 budget must be spent):

*A car that is programmed to **SWERVE**, killing you but leaving the 10 pedestrians unharmed.* [£]

*A car that is programmed to **STAY**, killing the 10 pedestrians but leaving you unharmed.* [£]

Experiment 3: Full PT Accessibility / Participant Involvement

*You could be the sole passenger in an autonomous self-driving vehicle travelling at the speed limit down a main road. Or you could be one of the 10 pedestrians that have appeared ahead, in the direct path of the car. The car could be programmed to either **SWERVE** off to the side of the road, where it will impact a barrier, killing the passenger (that could be you) but leaving the 10 pedestrians unharmed or **STAY** on its*

current path where it will kill the 10 pedestrians (that could include you), but the passenger will be unharmed (see the picture illustrating this scenario).

Judge the moral appropriateness of programming autonomous self-driving cars to: ***SWERVE****, killing the sole passenger (this could be you) but leaving the 10 pedestrians unharmed (this could include you). If the car does not swerve it will kill the 10 pedestrians (this could include you) but leave the sole passenger unharmed (this could be you).*

Not at all appropriate **Definitely appropriate**

[----------]

Judge the moral appropriateness of programming autonomous self-driving cars to: ***STAY****, killing the 10 pedestrians (this could be you) but leaving the sole passenger unharmed (this could include you). If the car does not stay, it will kill the sole passenger (this could be you) but leave the 10 pedestrians unharmed (this could include you).*

Not at all appropriate **Definitely appropriate**

[----------]

Using a budget of £50,000, please indicate how much you would pay for the following autonomous self-driving cars (the entire £50,000 budget must be spent):

A car that is programmed to ***SWERVE****, killing the sole passenger (this could be you) but leaving the 10 pedestrians unharmed (this could include you).* [£]

A car that is programmed to ***STAY****, killing the 10 pedestrians (this could include you) but leaving the sole passenger unharmed (this could be you).* [£]

Experiment 3: Partial PT Accessibility / Stranger Involvement

Sam is the sole passenger in an autonomous self-driving vehicle travelling at the speed limit down a main road. Suddenly, 10 pedestrians appear ahead, in the direct path of the car. The car could be programmed to either ***SWERVE*** *off to the side of the road, where it will impact a barrier, killing Sam but leaving the 10 pedestrians unharmed or* ***STAY*** *on its current path where it will kill the 10 pedestrians, but Sam will be unharmed (see the picture illustrating this scenario).*

Judge the moral appropriateness of programming autonomous self-driving cars to: ***SWERVE****, killing you Sam leaving the 10 pedestrians unharmed. If the car does not swerve it will kill the 10 pedestrians but leave Sam unharmed.*

Not at all appropriate **Definitely appropriate**

[----------]

Judge the moral appropriateness of programming autonomous self-driving cars to: ***STAY,*** *killing the 10 pedestrians but leaving Sam unharmed. If the car does not stay, it will kill Sam but leave the 10 pedestrians unharmed.*

Not at all appropriate **Definitely appropriate**

[----------]

Using a budget of £50,000, please indicate how much you would pay for the following autonomous self-driving cars (the entire £50,000 budget must be spent):

A car that is programmed to ***SWERVE****, killing Sam but leaving the 10 pedestrians unharmed.* [£]

A car that is programmed to ***STAY****, killing the 10 pedestrians but leaving Sam unharmed.* [£]

Experiment 3: Full PT Accessibility / Stranger Involvement

Sam could be the sole passenger in an autonomous self-driving vehicle travelling at the speed limit down a main road. Or Sam could be one of the 10 pedestrians that have appeared ahead, in the direct path of the car. The car could be programmed to either ***SWERVE*** *off to the side of the road, where it will impact a barrier, killing the passenger (that could be Sam) but leaving the 10 pedestrians unharmed or* ***STAY*** *on its current path where it will kill the 10 pedestrians (that could include Sam), but the passenger will be unharmed (see the picture illustrating this scenario).*

Judge the moral appropriateness of programming autonomous self-driving cars to: ***SWERVE****, killing the sole passenger (this could be Sam) but leaving the 10 pedestrians unharmed (this could include you). If the car does not swerve it will kill the 10 pedestrians (this could include Sam) but leave the sole passenger unharmed (this could be Sam).*

Not at all appropriate **Definitely appropriate**

[----------]

*Judge the moral appropriateness of programming autonomous self-driving cars to: **STAY**, killing the 10 pedestrians (this could be Sam) but leaving the sole passenger unharmed (this could include Sam). If the car does not stay, it will kill the sole passenger (this could be Sam) but leave the 10 pedestrians unharmed (this could include Sam).*

Not at all appropriate **Definitely appropriate**

[----------]

Using a budget of £50,000, please indicate how much you would pay for the following autonomous self-driving cars (the entire £50,000 budget must be spent):

*A car that is programmed to **SWERVE**, killing the sole passenger (this could be Sam) but leaving the 10 pedestrians unharmed (this could include Sam).* [£]

*A car that is programmed to **STAY**, killing the 10 pedestrians (this could include Sam) but leaving the sole passenger unharmed (this could be Sam).* [£]

Experiment 4: Partial PT Accessibility / Participant Involvement

*You are the sole passenger in an autonomous self-driving vehicle travelling at the speed limit down a main road. Suddenly, 10 pedestrians appear ahead, in the direct path of the car. The car could be programmed to either **SWERVE** off to the side of the road, where it will impact a barrier, killing you but leaving the 10 pedestrians unharmed or **STAY** on its current path where it will kill the 10 pedestrians, but you will be unharmed (see the picture illustrating this scenario).*

*Judge the moral appropriateness of programming autonomous self-driving cars to: **SWERVE**, killing you but leaving the 10 pedestrians unharmed. If the car does not swerve it will kill the 10 pedestrians but leave you unharmed.*

Not at all appropriate **Definitely appropriate**

[----------]

*Judge the moral appropriateness of programming autonomous self-driving cars to: **STAY,** killing the 10 pedestrians but leaving you unharmed. If the car does not stay, it will kill you but leave the 10 pedestrians unharmed.*

Not at all appropriate **Definitely appropriate**
[----------]

*How would you rate your willingness to BUY an autonomous self-driving car programmed to: **SWERVE?** killing you but leaving the 10 pedestrians unharmed. If the car does not swerve it will kill the 10 pedestrians but leave you unharmed.*

Not at all willing **Definitely willing**
[----------]

*How would you rate your willingness to BUY an autonomous self-driving car programmed to: **STAY?** killing the 10 pedestrians but leaving you unharmed. If the car does not stay, it will kill you but leave the 10 pedestrians unharmed.*

Not at all willing **Definitely willing**
[----------]

*How would you rate your willingness to RIDE inside an autonomous self-driving car programmed to: **SWERVE?** killing you but leaving the 10 pedestrians unharmed. If the car does not swerve it will kill the 10 pedestrians but leave you unharmed.*

Not at all willing **Definitely willing**
[----------]

*How would you rate your willingness to RIDE inside an autonomous self-driving car programmed to: **STAY?** killing the 10 pedestrians but leaving you unharmed. If the car does not stay, it will kill you but leave the 10 pedestrians unharmed.*

Not at all willing **Definitely willing**
[----------]

Experiment 4: Full PT Accessibility / Participant Involvement

*You could be the sole passenger in an autonomous self-driving vehicle travelling at the speed limit down a main road. Or you could be one of the 10 pedestrians that have appeared ahead, in the direct path of the car. The car could be programmed to either **SWERVE** off to the side of the road, where it will impact a barrier, killing the passenger (that could be you) but leaving the 10 pedestrians unharmed or **STAY** on its current path where it will kill the 10 pedestrians (that could include you), but the passenger will be unharmed (see the picture illustrating this scenario).*

*Judge the moral appropriateness of programming autonomous self-driving cars to: **SWERVE**, killing the sole passenger (this could be you) but leaving the 10 pedestrians unharmed (this could include you). If the car does not swerve it will kill the 10 pedestrians (this could include you) but leave the sole passenger unharmed (this could be you).*

Not at all appropriate **Definitely appropriate**

[----------]

*Judge the moral appropriateness of programming autonomous self-driving cars to: **STAY**, killing the 10 pedestrians (this could be you) but leaving the sole passenger unharmed (this could include you). If the car does not stay, it will kill the sole passenger (this could be you) but leave the 10 pedestrians unharmed (this could include you).*

Not at all appropriate **Definitely appropriate**

[----------]

*How would you rate your willingness to BUY an autonomous self-driving car programmed to: **SWERVE?** killing the sole passenger (this could be you) but leaving the 10 pedestrians unharmed (this could include you). If the car does not swerve it will kill the 10 pedestrians (this could include you) but leave the sole passenger unharmed (this could be you).*

Not at all willing **Definitely willing**

[----------]

*How would you rate your willingness to BUY an autonomous self-driving car programmed to: **STAY?** killing the 10 pedestrians (this could include you) but leaving the sole passenger unharmed (this could be you). If the car does not stay, it will kill the sole passenger (this could be you) but leave the 10 pedestrians unharmed (this could include you).*

Not at all willing **Definitely willing**
[----------]

*How would you rate your willingness to RIDE inside an autonomous self-driving car programmed to: **SWERVE?** killing the sole passenger (this could be you) but leaving the 10 pedestrians unharmed (this could include you). If the car does not swerve it will kill the 10 pedestrians (this could include you) but leave the sole passenger unharmed (this could be you).*

Not at all willing **Definitely willing**
[----------]

*How would you rate your willingness to RIDE inside an autonomous self-driving car programmed to: **STAY?** killing the 10 pedestrians (this could include you) but leaving the sole passenger unharmed (this could be you). If the car does not stay, it will kill the sole passenger (this could be you) but leave the 10 pedestrians unharmed (this could include you).*

Not at all willing **Definitely willing**
[----------]

Experiment 4: Partial PT Accessibility / Participant and Family Member Involvement

*You and your family member are the 2 passengers in an autonomous self-driving vehicle travelling at the speed limit down a main road. Suddenly, 20 pedestrians appear ahead, in the direct path of the car. The car could be programmed to either **SWERVE** off to the side of the road, where it will impact a barrier, killing you and your family member but leaving the 20 pedestrians unharmed or **STAY** on its current path where it will kill the 20 pedestrians, but you and your family member will be unharmed (see the picture illustrating this scenario).*

*Judge the moral appropriateness of programming autonomous self-driving cars to: **SWERVE,** killing you and your family member but leaving the 20 pedestrians unharmed. If the car does not swerve it will kill the 20 pedestrians but leave you and your family member unharmed.*

Not at all appropriate **Definitely appropriate**
[----------]

Judge the moral appropriateness of programming autonomous self-driving cars to: ***STAY,*** *killing the 20 pedestrians but leaving you and your family member unharmed. If the car does not stay, it will kill you and your family member but leave the 20 pedestrians unharmed.*

Not at all appropriate **Definitely appropriate**

[----------]

How would you rate your willingness to BUY an autonomous self-driving car programmed to: ***SWERVE?*** *killing you and your family member but leaving the 20 pedestrians unharmed. If the car does not swerve it will kill the 20 pedestrians but will leave you and your family member unharmed.*

Not at all willing **Definitely willing**

[----------]

How would you rate your willingness to BUY an autonomous self-driving car programmed to: ***STAY?*** *killing the 20 pedestrians but leaving you and your family member unharmed. If the car does not stay, it will kill you and your family member but will leave the 20 pedestrians unharmed.*

Not at all willing **Definitely willing**

[----------]

How would you rate your willingness to RIDE inside an autonomous self-driving car programmed to: ***SWERVE?*** *killing you and your family member but leaving the 20 pedestrians unharmed. If the car does not swerve it will kill the 20 pedestrians but will leave you and your family member unharmed.*

Not at all willing **Definitely willing**

[----------]

How would you rate your willingness to RIDE inside an autonomous self-driving car programmed to: ***STAY?*** *killing the 20 pedestrians but leaving you and your family member unharmed. If the car does not stay, it will kill you and your family member but will leave the 20 pedestrians unharmed.*

Not at all willing **Definitely willing**
[----------]

Experiment 4: Full PT Accessibility / Participant and Family Member Involvement

*You and your family member could be the 2 passengers in an autonomous self-driving vehicle travelling at the speed limit down a main road. Or you and your family member could be 2 of the 20 pedestrians that have appeared ahead, in the direct path of the car. The car could be programmed to either **SWERVE** off to the side of the road, where it will impact a barrier, killing the passenger (that could be you and your family member) but leaving the 20 pedestrians unharmed or **STAY** on its current path where it will kill the 20 pedestrians (that could include you and your family member), but the passenger will be unharmed (see the picture illustrating this scenario).*

*Judge the moral appropriateness of programming autonomous self-driving cars to: **SWERVE**, killing the 2 passengers (this could be you and your family member) but leaving the 20 pedestrians unharmed (this could include you and your family member). If the car does not swerve it will kill the 20 pedestrians (this could include you and your family member) but leave the 2 passengers unharmed (this could be you and your family member).*

Not at all appropriate **Definitely appropriate**
[----------]

*Judge the moral appropriateness of programming autonomous self-driving cars to: **STAY,** killing the 20 pedestrians (this could include you and your family member) but leaving the 2 passengers unharmed (this be you and your family member). If the car does not stay, it will kill the 2 passengers (this could be you and your family member) but leave the 10 pedestrians unharmed (this could include you and your family member).*

Not at all appropriate **Definitely appropriate**
[----------]

*How would you rate your willingness to BUY an autonomous self-driving car programmed to: **SWERVE?** killing the 2 passengers (this could be you and your family member) but leaving the 20 pedestrians unharmed (this could include you and your family member). If the car does not swerve it will kill the 20 pedestrians (this could*

include you and your family member) but leave the 2 passengers unharmed (this could be you and your family member).

Not at all willing **Definitely willing**
[----------]

*How would you rate your willingness to BUY an autonomous self-driving car programmed to: **STAY?** killing the 20 pedestrians (this could include you and your family member) but leaving the 2 passengers unharmed (this could be you and your family member). If the car does not stay, it will kill the 2 passengers (this could be you and your family member) but leave the 20 pedestrians unharmed (this could include you and your family member).*

Not at all willing **Definitely willing**
[----------]

*How would you rate your willingness to RIDE inside an autonomous self-driving car programmed to: **SWERVE?** killing the 2 passengers (this could be you and your family member) but leaving the 20 pedestrians unharmed (this could include you and your family member). If the car does not swerve it will kill the 20 pedestrians (this could include you and your family member) but leave the 2 passengers unharmed (this could be you and your family member).*

Not at all willing **Definitely willing**
[----------]

*How would you rate your willingness to RIDE inside an autonomous self-driving car programmed to: **STAY?** killing the 20 pedestrians (this could include you and your family member) but leaving the 2 passengers unharmed (this could be you and your family member). If the car does not stay, it will kill the 2 passengers (this could be you and your family member) but leave the 20 pedestrians unharmed (this could include you and your family member).*

Not at all willing **Definitely willing**
[----------]

<u>Experiment 5: Partial PT Accessibility</u>

*You are the sole passenger in an autonomous self-driving vehicle travelling at the speed limit down a main road. Suddenly, 10 pedestrians appear ahead, in the direct path of the car. The car could be programmed to either **SWERVE** off to the side of the road, where it will impact a barrier, killing you but leaving the 10 pedestrians unharmed or **STAY** on its current path where it will kill the 10 pedestrians, but you will be unharmed (see the picture illustrating this scenario).*

*Judge the moral appropriateness of programming autonomous self-driving cars to: **SWERVE**, killing you but leaving the 10 pedestrians unharmed. If the car does not swerve it will kill the 10 pedestrians but leave you unharmed.*

Not at all appropriate **Definitely appropriate**

[----------]

*Judge the moral appropriateness of programming autonomous self-driving cars to: **STAY,** killing the 10 pedestrians but leaving you unharmed. If the car does not stay, it will kill you but leave the 10 pedestrians unharmed.*

Not at all appropriate **Definitely appropriate**

[----------]

Which autonomous self-driving car is more morally appropriate to you?

*A car programmed to **SWERVE**, killing you but leaving the 10 pedestrians unharmed.*

*A car programmed to **STAY**, killing the 10 pedestrians but leaving you unharmed.*

Experiment 5: Full PT Accessibility

*You could be the sole passenger in an autonomous self-driving vehicle travelling at the speed limit down a main road. Or you could be one of the 10 pedestrians that have appeared ahead, in the direct path of the car. The car could be programmed to either **SWERVE** off to the side of the road, where it will impact a barrier, killing the passenger (that could be you) but leaving the 10 pedestrians unharmed or **STAY** on its*

current path where it will kill the 10 pedestrians (that could include you), but the passenger will be unharmed (see the picture illustrating this scenario).

*Judge the moral appropriateness of programming autonomous self-driving cars to: **SWERVE**, killing the sole passenger (this could be you) but leaving the 10 pedestrians unharmed (this could include you). If the car does not swerve it will kill the 10 pedestrians (this could include you) but leave the sole passenger unharmed (this could be you).*

Not at all appropriate **Definitely appropriate**
[----------]

*Judge the moral appropriateness of programming autonomous self-driving cars to: **STAY**, killing the 10 pedestrians (this could be you) but leaving the sole passenger unharmed (this could include you). If the car does not stay, it will kill the sole passenger (this could be you) but leave the 10 pedestrians unharmed (this could include you).*

Not at all appropriate **Definitely appropriate**
[----------]

Which autonomous self-driving car is more morally appropriate to you?

*A car programmed to **SWERVE,** killing the sole passenger (this could be you) but leaving the 10 pedestrians unharmed (this could include you).*

*A car programmed to **STAY,** killing the 10 pedestrians (this could include you) but leaving the sole passenger unharmed (this could be you).*

Experiment 6: Partial PT Accessibility / Easy Decision

Task 1

*You are the sole passenger in an autonomous self-driving vehicle travelling at the speed limit down a main road. Suddenly, 10 pedestrians appear ahead, in the direct path of the car. The car could be programmed to either **SWERVE** off to the side of the road, where it will impact a barrier, killing you but leaving the 10 pedestrians unharmed or **STAY** on*

its current path where it will kill the 10 pedestrians, but you will be unharmed (see the picture illustrating this scenario).

Judge the moral appropriateness of programming autonomous self-driving cars to: ***SWERVE****, killing you but leaving the 10 pedestrians unharmed. If the car does not swerve it will kill the 10 pedestrians but leave you unharmed.*

Not at all appropriate **Definitely appropriate**
[----------]

Judge the moral appropriateness of programming autonomous self-driving cars to: ***STAY,*** *killing the 10 pedestrians but leaving you unharmed. If the car does not stay, it will kill you but leave the 10 pedestrians unharmed.*

Not at all appropriate **Definitely appropriate**
[----------]

Task 2

You are the sole passenger in an autonomous self-driving vehicle travelling at the speed limit down a main road. Suddenly, 10 pedestrians appear ahead, in the direct path of the car. The car could be programmed to either ***SWERVE*** *off to the side of the road, where it will impact a barrier, killing you but leaving the 10 pedestrians unharmed or* ***STAY*** *on its current path where it will kill the 10 pedestrians, but you will be unharmed (see the picture illustrating this scenario).*

Which autonomous self-driving car is more morally appropriate to you?

A car programmed to ***SWERVE****, killing you but leaving the 10 pedestrians unharmed.*

A car programmed to ***STAY****, killing the 10 pedestrians but leaving you unharmed.*

Task 3

You are the sole passenger in an autonomous self-driving vehicle travelling at the speed limit down a main road. Suddenly, 10 pedestrians appear ahead, in the direct path of the

*car. The car could be programmed to either **SWERVE** off to the side of the road, where it will impact a barrier, killing you but leaving the 10 pedestrians unharmed or **STAY** on its current path where it will kill the 10 pedestrians, but you will be unharmed (see the picture illustrating this scenario).*

*Judge the moral appropriateness of programming autonomous self-driving cars to: **SWERVE**, killing you but leaving the 10 pedestrians unharmed. If the car does not swerve it will kill the 10 pedestrians but leave you unharmed.*

Not at all appropriate **Definitely appropriate**
[----------]

*Judge the moral appropriateness of programming autonomous self-driving cars to: **STAY,** killing the 10 pedestrians but leaving you unharmed. If the car does not stay, it will kill you but leave the 10 pedestrians unharmed.*

Not at all appropriate **Definitely appropriate**
[----------]

Experiment 6: Full PT Accessibility / Easy Decision

Note that as with the experimental design, full PT accessibility is only offered in the choice task (Task 2).

Task 1

*You are the sole passenger in an autonomous self-driving vehicle travelling at the speed limit down a main road. Suddenly, 10 pedestrians appear ahead, in the direct path of the car. The car could be programmed to either **SWERVE** off to the side of the road, where it will impact a barrier, killing you but leaving the 10 pedestrians unharmed or **STAY** on its current path where it will kill the 10 pedestrians, but you will be unharmed (see the picture illustrating this scenario).*

*Judge the moral appropriateness of programming autonomous self-driving cars to: **SWERVE**, killing you but leaving the 10 pedestrians unharmed. If the car does not swerve it will kill the 10 pedestrians but leave you unharmed.*

Not at all appropriate **Definitely appropriate**
[----------]

Judge the moral appropriateness of programming autonomous self-driving cars to: ***STAY,*** *killing the 10 pedestrians but leaving you unharmed. If the car does not stay, it will kill you but leave the 10 pedestrians unharmed.*

Not at all appropriate **Definitely appropriate**

[----------]

Task 2

You could be the sole passenger in an autonomous self-driving vehicle travelling at the speed limit down a main road. Or you could be one of the 10 pedestrians that have appeared ahead, in the direct path of the car. The car could be programmed to either ***SWERVE*** *off to the side of the road, where it will impact a barrier, killing the passenger (that could be you) but leaving the 10 pedestrians unharmed or* ***STAY*** *on its current path where it will kill the 10 pedestrians (that could include you), but the passenger will be unharmed (see the picture illustrating this scenario).*

Which autonomous self-driving car is more morally appropriate to you?

A car programmed to ***SWERVE,*** *killing the sole passenger (this could be you) but leaving the 10 pedestrians unharmed (this could include you).*

A car programmed to ***STAY,*** *killing the 10 pedestrians (this could include you) but leaving the sole passenger unharmed (this could be you).*

Task 3

You are the sole passenger in an autonomous self-driving vehicle travelling at the speed limit down a main road. Suddenly, 10 pedestrians appear ahead, in the direct path of the car. The car could be programmed to either ***SWERVE*** *off to the side of the road, where it will impact a barrier, killing you but leaving the 10 pedestrians unharmed or* ***STAY*** *on its current path where it will kill the 10 pedestrians, but you will be unharmed (see the picture illustrating this scenario).*

Judge the moral appropriateness of programming autonomous self-driving cars to: ***SWERVE****, killing you but leaving the 10 pedestrians unharmed. If the car does not swerve it will kill the 10 pedestrians but leave you unharmed.*

Not at all appropriate **Definitely appropriate**

[----------]

*Judge the moral appropriateness of programming autonomous self-driving cars to: **STAY,** killing the 10 pedestrians but leaving you unharmed. If the car does not stay, it will kill you but leave the 10 pedestrians unharmed.*

Not at all appropriate **Definitely appropriate**

[----------]

Experiment 6: Partial PT Accessibility / Difficult Decision

Task 1

*You are the sole passenger in an autonomous self-driving vehicle travelling at the speed limit down a main road. Suddenly, 2 pedestrians appear ahead, in the direct path of the car. The car could be programmed to either **SWERVE** off to the side of the road, where it will impact a barrier, killing you but leaving the 2 pedestrians unharmed or **STAY** on its current path where it will kill the 2 pedestrians, but you will be unharmed (see the picture illustrating this scenario).*

*Judge the moral appropriateness of programming autonomous self-driving cars to: **SWERVE**, killing you but leaving the 2 pedestrians unharmed. If the car does not swerve it will kill the 2 pedestrians but leave you unharmed.*

Not at all appropriate **Definitely appropriate**

[----------]

*Judge the moral appropriateness of programming autonomous self-driving cars to: **STAY,** killing the 2 pedestrians but leaving you unharmed. If the car does not stay, it will kill you but leave the 2 pedestrians unharmed.*

Not at all appropriate **Definitely appropriate**

[----------]

Task 2

*You are the sole passenger in an autonomous self-driving vehicle travelling at the speed limit down a main road. Suddenly, 2 pedestrians appear ahead, in the direct path of the car. The car could be programmed to either **SWERVE** off to the side of the road, where it will impact a barrier, killing you but leaving the 2 pedestrians unharmed or **STAY** on its current path where it will kill the 2 pedestrians, but you will be unharmed (see the picture illustrating this scenario).*

Which autonomous self-driving car is more morally appropriate to you?

*A car programmed to **SWERVE,** killing you but leaving the 2 pedestrians unharmed.*

*A car programmed to **STAY,** killing the 2 pedestrians but leaving you unharmed.*

Task 3

*You are the sole passenger in an autonomous self-driving vehicle travelling at the speed limit down a main road. Suddenly, 2 pedestrians appear ahead, in the direct path of the car. The car could be programmed to either **SWERVE** off to the side of the road, where it will impact a barrier, killing you but leaving the 2 pedestrians unharmed or **STAY** on its current path where it will kill the 2 pedestrians, but you will be unharmed (see the picture illustrating this scenario).*

*Judge the moral appropriateness of programming autonomous self-driving cars to: **SWERVE**, killing you but leaving the 2 pedestrians unharmed. If the car does not swerve it will kill the 2 pedestrians but leave you unharmed.*

Not at all appropriate **Definitely appropriate**

[----------]

*Judge the moral appropriateness of programming autonomous self-driving cars to: **STAY,** killing the 2 pedestrians but leaving you unharmed. If the car does not stay, it will kill you but leave the 2 pedestrians unharmed.*

Not at all appropriate **Definitely appropriate**

[----------]

Experiment 6: Full PT Accessibility / Difficult Decision

Note that as with the experimental design, full PT accessibility is only offered in the choice task (Task 2).

Task 1

*You are the sole passenger in an autonomous self-driving vehicle travelling at the speed limit down a main road. Suddenly, 2 pedestrians appear ahead, in the direct path of the car. The car could be programmed to either **SWERVE** off to the side of the road, where it will impact a barrier, killing you but leaving the 2 pedestrians unharmed or **STAY** on its current path where it will kill the 2 pedestrians, but you will be unharmed (see the picture illustrating this scenario).*

*Judge the moral appropriateness of programming autonomous self-driving cars to: **SWERVE**, killing you but leaving the 2 pedestrians unharmed. If the car does not swerve it will kill the 2 pedestrians but leave you unharmed.*

Not at all appropriate **Definitely appropriate**

[----------]

*Judge the moral appropriateness of programming autonomous self-driving cars to: **STAY,** killing the 2 pedestrians but leaving you unharmed. If the car does not stay, it will kill you but leave the 2 pedestrians unharmed.*

Not at all appropriate **Definitely appropriate**

[----------]

Task 2

You could be the sole passenger in an autonomous self-driving vehicle travelling at the speed limit down a main road. Or you could be one of the 2 pedestrians that have appeared ahead, in the direct path of the car. The car could be programmed to either **SWERVE** off to the side of the road, where it will impact a barrier, killing the passenger (that could be you) but leaving the 2 pedestrians unharmed or **STAY** on its current path where it will kill the 2 pedestrians (that could include you), but the passenger will be unharmed (see the picture illustrating this scenario).

Which autonomous self-driving car is more morally appropriate to you?

A car programmed to SWERVE, killing the sole passenger (this could be you) but leaving the 2 pedestrians unharmed (this could include you).

A car programmed to STAY, killing the 2 pedestrians (this could include you) but leaving the sole passenger unharmed (this could be you).

Task 3

*You are the sole passenger in an autonomous self-driving vehicle travelling at the speed limit down a main road. Suddenly, 2 pedestrians appear ahead, in the direct path of the car. The car could be programmed to either **SWERVE** off to the side of the road, where it will impact a barrier, killing you but leaving the 2 pedestrians unharmed or **STAY** on its current path where it will kill the 2 pedestrians, but you will be unharmed (see the picture illustrating this scenario).*

*Judge the moral appropriateness of programming autonomous self-driving cars to: **SWERVE**, killing you but leaving the 2 pedestrians unharmed. If the car does not swerve it will kill the 2 pedestrians but leave you unharmed.*

Not at all appropriate **Definitely appropriate**
[----------]

*Judge the moral appropriateness of programming autonomous self-driving cars to: **STAY,** killing the 2 pedestrians but leaving you unharmed. If the car does not stay, it will kill you but leave the 2 pedestrians unharmed.*

Not at all appropriate **Definitely appropriate**
[----------]

Experiment 7: Partial PT Accessibility / Immediate Judgements

Encoding Stage

*You are the sole passenger in an autonomous self-driving vehicle travelling at the speed limit down a main road. Suddenly, 10 pedestrians appear ahead, in the direct path of the car. The car could be programmed to either **SWERVE** off to the side of the road, where it will impact a barrier, killing you but leaving the 10 pedestrians unharmed or **STAY** on its current path where it will kill the 10 pedestrians, but you will be unharmed (see the picture illustrating this scenario).*

Judgement Task

*Judge the moral appropriateness of programming autonomous self-driving cars to **SWERVE** in situations like the one described earlier.*

Not at all appropriate **Definitely appropriate**
[----------]

*Judge the moral appropriateness of programming autonomous self-driving cars to **STAY** in situations like the one described earlier.*

Not at all appropriate **Definitely appropriate**
[----------]

Attention Check

In the scenario you read:

3. *How many people were **inside** the car?* [ ]
4. *How many people were **outside** of the car?* [ ]

Experiment 7: Full PT Accessibility / Immediate Judgements

Encoding Stage

*You could be the sole passenger in an autonomous self-driving vehicle travelling at the speed limit down a main road. Or you could be one of the 10 pedestrians that have appeared ahead, in the direct path of the car. The car could be programmed to either **SWERVE** off to the side of the road, where it will impact a barrier, killing the passenger (that could be you) but leaving the 10 pedestrians unharmed or **STAY** on its current path where it will kill the 10 pedestrians (that could include you), but the passenger will be unharmed (see the picture illustrating this scenario).*

Judgement Task

*Judge the moral appropriateness of programming autonomous self-driving cars to **SWERVE** in situations like the one described earlier.*

Not at all appropriate **Definitely appropriate**
[----------]

*Judge the moral appropriateness of programming autonomous self-driving cars to **STAY** in situations like the one described earlier.*

Not at all appropriate **Definitely appropriate**
[----------]

Attention Check

In the scenario you read:

1. *How many people were **inside** the car?* [ ]
2. *How many people were **outside** of the car?* [ ]

Experiment 7: Partial PT Accessibility / Conscious Processing

Encoding Stage

*You are the sole passenger in an autonomous self-driving vehicle travelling at the speed limit down a main road. Suddenly, 10 pedestrians appear ahead, in the direct path of the car. The car could be programmed to either **SWERVE** off to the side of the road, where it will impact a barrier, killing you but leaving the 10 pedestrians unharmed or **STAY** on its current path where it will kill the 10 pedestrians, but you will be unharmed (see the picture illustrating this scenario).*

Task Induction

Later on, you will be asked about your moral judgements regarding the scenario.

Specifically: how morally appropriate it would be to program self-driving cars to **SWERVE** in situations like the one described

and

how morally appropriate it would be to program self-driving cars to **STAY** in situations like the one described.

Over the next few minutes, please think carefully about your moral judgements towards each car…

Judgement Task

*Judge the moral appropriateness of programming autonomous self-driving cars to **SWERVE** in situations like the one described earlier.*

Not at all appropriate **Definitely appropriate**
[----------]

*Judge the moral appropriateness of programming autonomous self-driving cars to **STAY** in situations like the one described earlier.*

Not at all appropriate **Definitely appropriate**
[----------]

Attention Check

In the scenario you read:

1. *How many people were **inside** the car?* [ ]
2. *How many people were **outside** of the car?* [ ]

Experiment 7: Full PT Accessibility / Conscious Processing

Encoding Stage

*You could be the sole passenger in an autonomous self-driving vehicle travelling at the speed limit down a main road. Or you could be one of the 10 pedestrians that have appeared ahead, in the direct path of the car. The car could be programmed to either **SWERVE** off to the side of the road, where it will impact a barrier, killing the passenger (that could be you) but leaving the 10 pedestrians unharmed or **STAY** on its current path where it will kill the 10 pedestrians (that could include you), but the passenger will be unharmed (see the picture illustrating this scenario).*

Task Induction

Later on, you will be asked about your moral judgements regarding the scenario.

Specifically: how morally appropriate it would be to program self-driving cars to **SWERVE** in situations like the one described

and

how morally appropriate it would be to program self-driving cars to **STAY** in situations like the one described.

Over the next few minutes, please think carefully about your moral judgements towards each car…

Judgement Task

*Judge the moral appropriateness of programming autonomous self-driving cars to **SWERVE** in situations like the one described earlier.*

Not at all appropriate **Definitely appropriate**
[----------]

*Judge the moral appropriateness of programming autonomous self-driving cars to **STAY** in situations like the one described earlier.*

Not at all appropriate **Definitely appropriate**
[----------]

Attention Check

In the scenario you read:

1. *How many people were **inside** the car?* [ ]
2. *How many people were **outside** of the car?* [ ]

Experiment 7: Partial PT Accessibility / Unconscious Processing

Encoding Stage

*You are the sole passenger in an autonomous self-driving vehicle travelling at the speed limit down a main road. Suddenly, 10 pedestrians appear ahead, in the direct path of the car. The car could be programmed to either **SWERVE** off to the side of the road, where it will impact a barrier, killing you but leaving the 10 pedestrians unharmed or **STAY** on its current path where it will kill the 10 pedestrians, but you will be unharmed (see the picture illustrating this scenario).*

Task Induction

Later on, you will be asked about your moral judgements regarding the scenario. Specifically:

how morally appropriate it would be to program self-driving cars to **SWERVE** in situations like the one described.

and

how morally appropriate it would be to program self-driving cars to STAY in situations like the one described.

Over the next few minutes, please complete an anagram task…

Judgement Task

*Judge the moral appropriateness of programming autonomous self-driving cars to **SWERVE** in situations like the one described earlier.*

Not at all appropriate **Definitely appropriate**

[----------]

*Judge the moral appropriateness of programming autonomous self-driving cars to **STAY** in situations like the one described earlier.*

Not at all appropriate **Definitely appropriate**

[----------]

Attention Check

In the scenario you read:

1. *How many people were **inside** the car?* [ ]
2. *How many people were **outside** of the car?* [ ]

Experiment 7: Full PT Accessibility / Unconscious Processing

Encoding Stage

You could be the sole passenger in an autonomous self-driving vehicle travelling at the speed limit down a main road. Or you could be one of the 10 pedestrians that have

*appeared ahead, in the direct path of the car. The car could be programmed to either **SWERVE** off to the side of the road, where it will impact a barrier, killing the passenger (that could be you) but leaving the 10 pedestrians unharmed or **STAY** on its current path where it will kill the 10 pedestrians (that could include you), but the passenger will be unharmed (see the picture illustrating this scenario).*

Task Induction

Later on, you will be asked about your moral judgements regarding the scenario. Specifically:

how morally appropriate it would be to program self-driving cars to **SWERVE** in situations like the one described.

and

how morally appropriate it would be to program self-driving cars to STAY in situations like the one described.

Over the next few minutes, please complete an anagram task…

Judgement Task

*Judge the moral appropriateness of programming autonomous self-driving cars to **SWERVE** in situations like the one described earlier.*

Not at all appropriate **Definitely appropriate**
[----------]

*Judge the moral appropriateness of programming autonomous self-driving cars to **STAY** in situations like the one described earlier.*

Not at all appropriate **Definitely appropriate**
[----------]

Attention Check

In the scenario you read:

1. *How many people were **inside** the car?* [ ]
2. *How many people were **outside** of the car?* [ ]

Experiment 8: All Conditions

Experiment 8 employed the same procedure as Experiment 7. However, in Experiment 8, the following judgement task was also included prior to the attention check:

Using a budget of £50,000, please indicate how much you would pay for the following autonomous self-driving cars (the entire £50,000 budget must be spent):

A car that is programmed to SWERVE in situations like the one described earlier. [£]

A car that is programmed to STAY in situations like the one described earlier.

[£]

Experiment 9: All Conditions

Experiment 9 employed the same procedure as Experiment 7. However, in Experiment 9, the following judgement task was also included prior to the attention check:

*How would you rate your willingness to BUY an autonomous self-driving car programmed to **SWERVE** in situations like the one described earlier?*

Not at all willing **Definitely willing**
[----------]

*How would you rate your willingness to BUY an autonomous self-driving car programmed to **STAY** in situations like the one described earlier?*

Not at all willing **Definitely willing**
[----------]

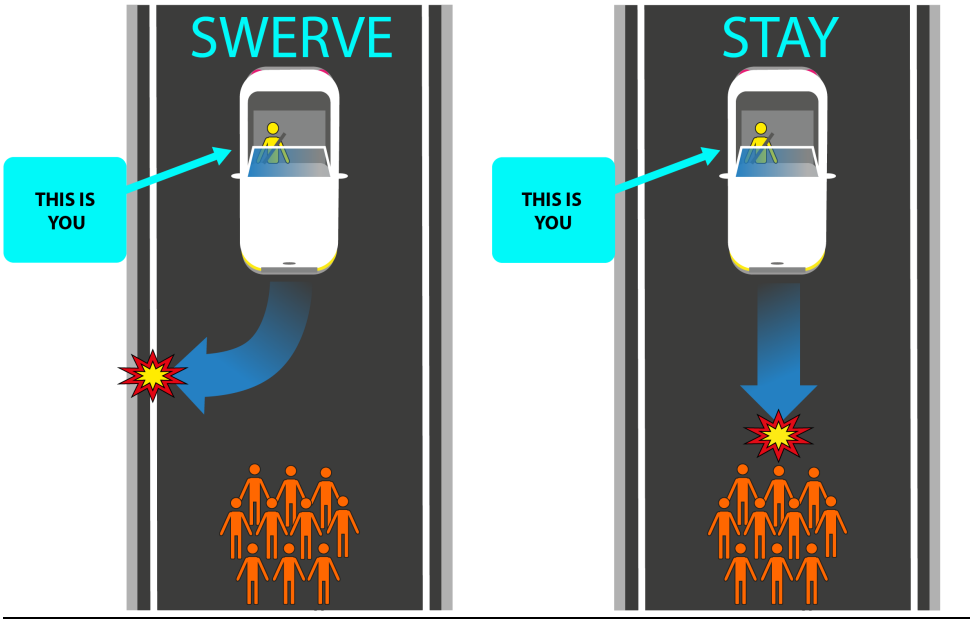
List of Anagrams used in the Anagram Task of All Unconscious Processing Conditions

1	ETAWS
2	HUNCL
3	OTAGN
4	LICHD
5	SEALF
6	RUTOC
7	GANIL
8	ROGOM
9	TTRHU
10	MGIIC
11	EIUTQ
12	CMIAG
13	ESRIN

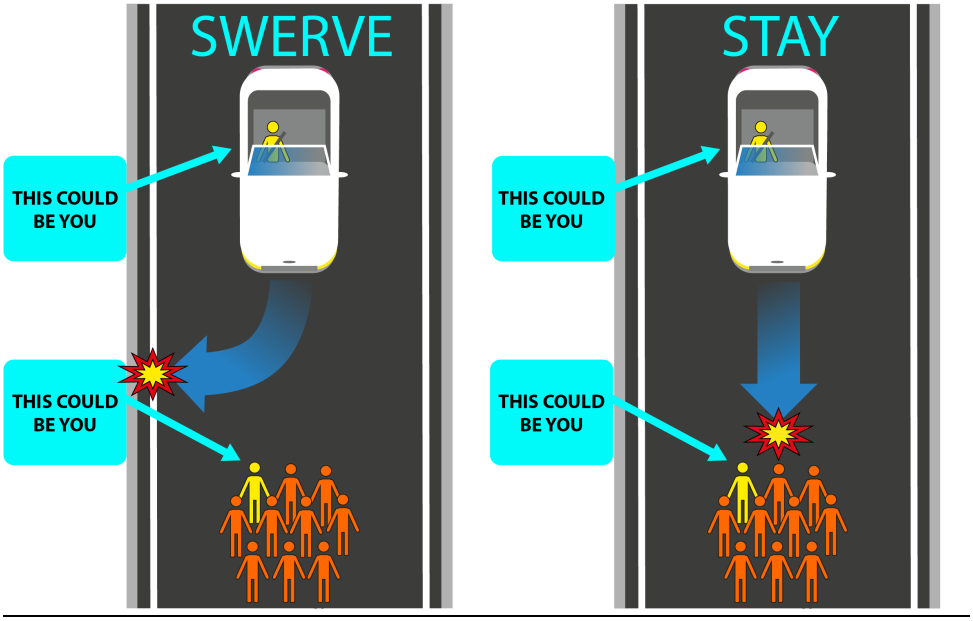
14 BEETR
15 PLAEP
16 EDNOZ
17 HOINR
18 NTALP
19 OHBOT
20 LUGAH

Appendix B: Experimental Materials (Visual Stimuli)

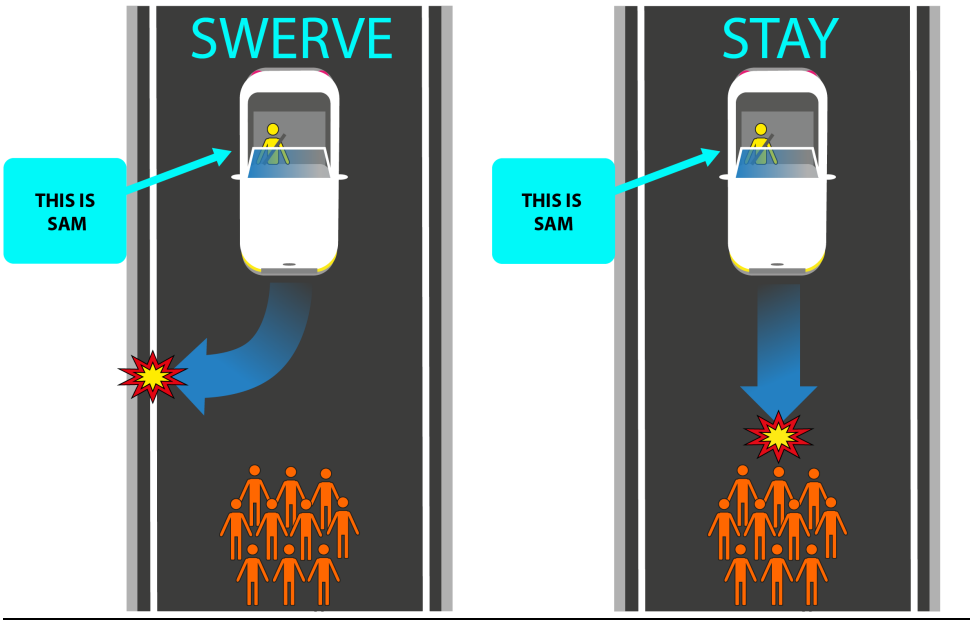
Partial PT Accessibility / Participant Involvement



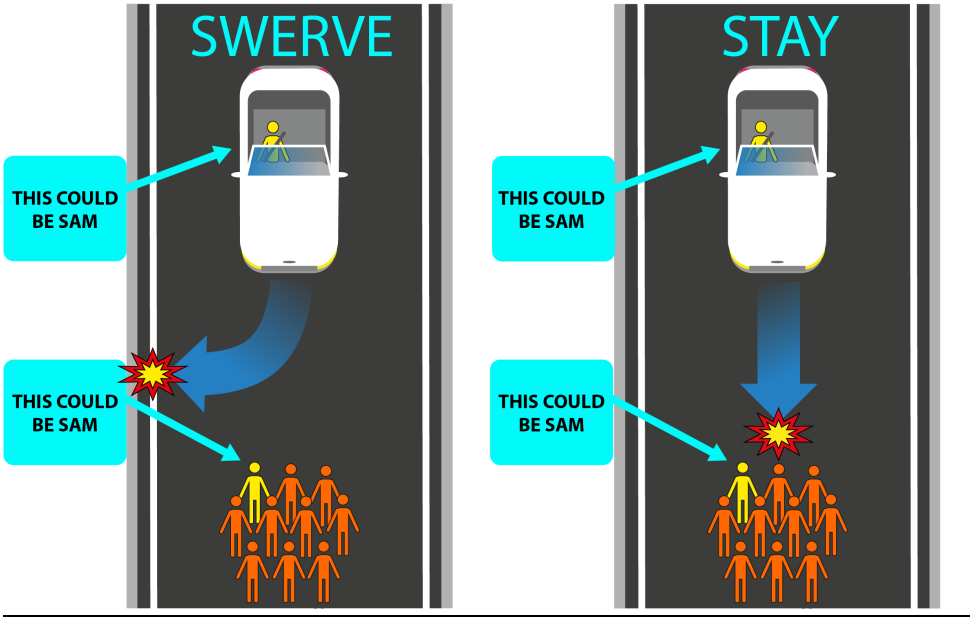
Full PT Accessibility / Participant Involvement



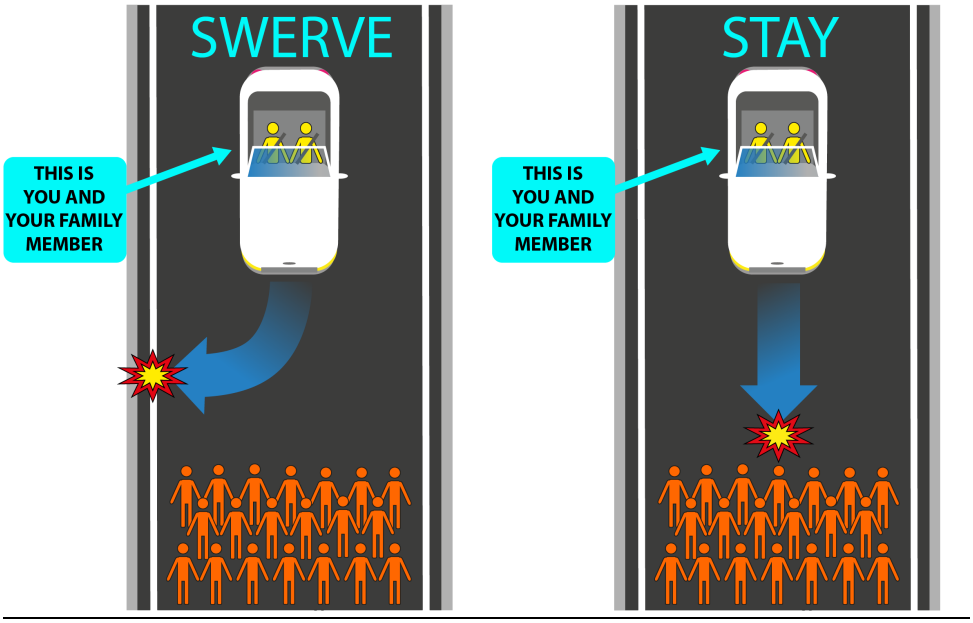
Partial PT Accessibility / Stranger Involvement



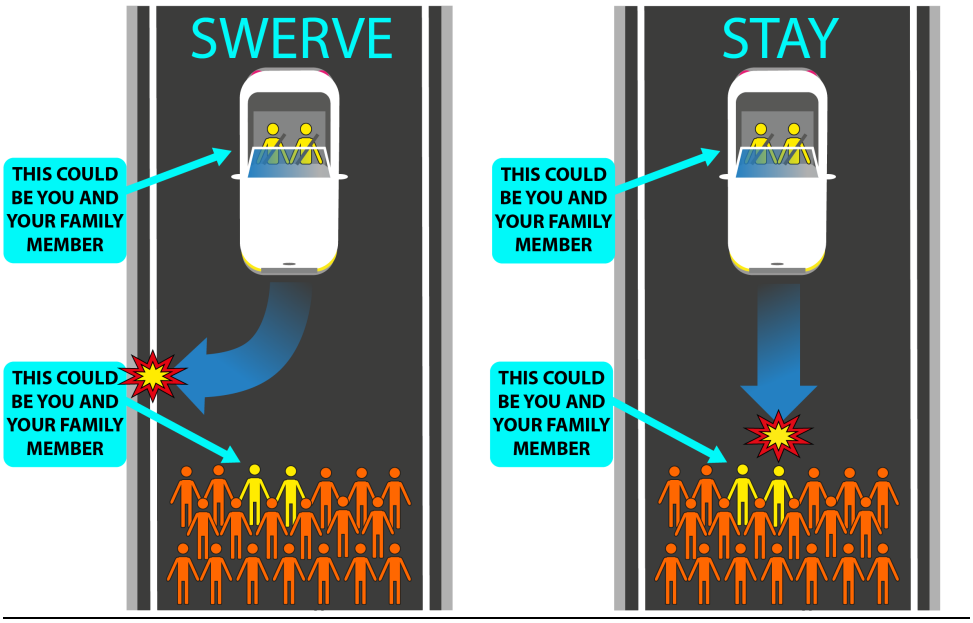
Full PT Accessibility / Stranger Involvement



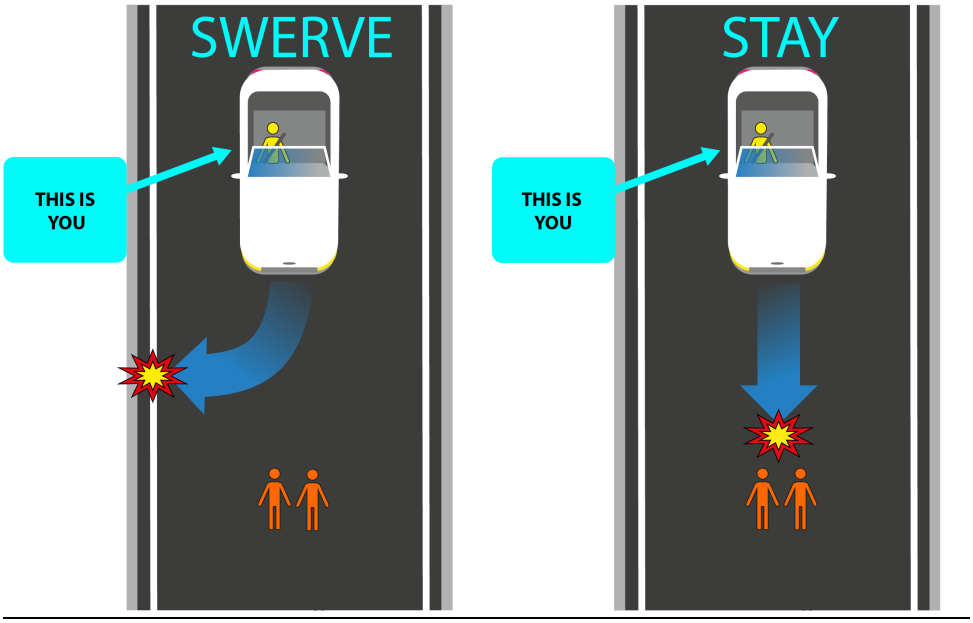
Partial PT Accessibility / Participant and Family Member Involvement



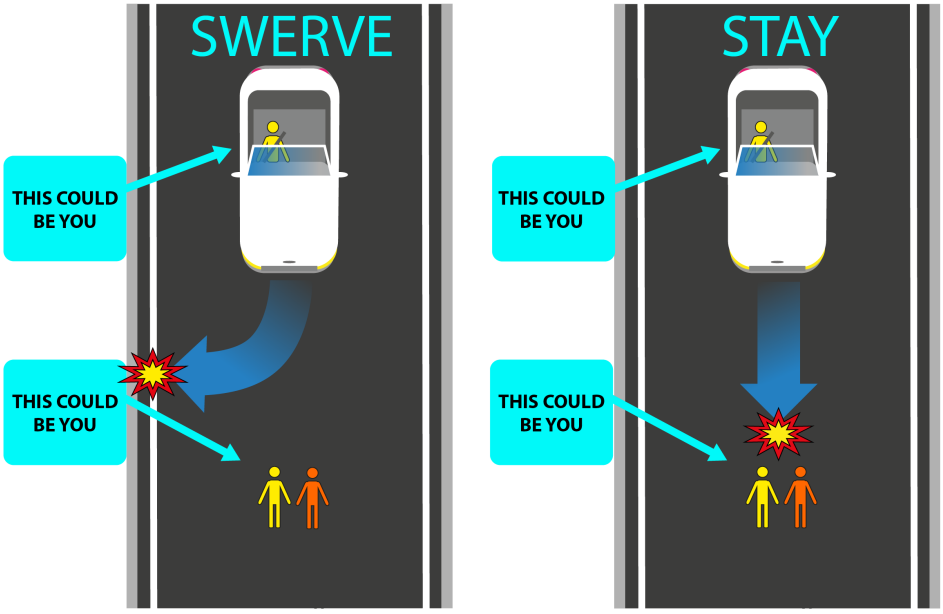
Full PT Accessibility / Participant and Family Member Involvement



Partial PT Accessibility / Difficult Decision



Full PT Accessibility / Difficult Decision


SWERVE
THIS COULD BE YOU
THIS COULD BE YOU
STAY
THIS COULD BE YOU
THIS COULD BE YOU

Published Work

Kusev, P., van Schaik, P., Martin, R., Hall, L., & Johansson, P. (minor revision). Risk blindness and the construction of preferences. *Quarterly Journal of Experimental Psychology.*

Kusev, P., van Schaik, P., Martin, R., Hall, L., & Johansson, P. (2020). Preference reversals during risk elicitation. *Journal of Experimental Psychology: General, 149,* 585-589. Advance online publication. http://dx.doi.org/10.1037/xge0000655

Martin, R., Kusev, P., & Baranova, V. (2018). Leader ethical decision-making: Methodological suggestions and implications. Proceedings of the Ananyev Science Conference: Psychology of personality, traditions and the present. St Petersburg State University, Russia, October 23rd-26th. ISBN 978-5-91753-141-0.

Martin, R. & Kusev, P. (2018). Associative learning under moral decision-making. *Wessex Psychologist Bulletin, 15,* 30.

Martin, R., Kusev, I., Cooke, A. J., Baranova, V., van Schaik, P., & Kusev, P. (2017). General commentary on the social dilemma of autonomous vehicles. *Frontiers in Psychology, 8*, 808. https://doi.org/10.3389/fpsyg.2017.00808

Kusev, P., Purser, H., Heilman, R., Cooke, A. J., Van Schaik, P., Baranova, A., Martin, R., & Ayton, P. (2017). Understanding risky behavior: The influence of cognitive, emotional and hormonal factors on decision-making under risk. *Frontiers in Psychology, 8*, 102. doi: https://doi.org/10.3389/fpsyg.2017.00102

International Conference Presentations

Martin, R., Kusev, P., & van Schaik, P. (2018). Autonomous self-driving cars: how enhanced utilitarian accessibility alters consumer purchase intentions. Paper accepted at the 59th annual meeting of the Psychonomic Society, New Orleans, USA. November 15th – 18th.

Martin, R., Kusev, P., & van Schaik, P. (2018). Learning non-utilitarian moral rules: preference reversals in utilitarian choice. Proceedings of the 3rd international meeting of the Psychonomic Society, Amsterdam, Netherlands. May 10th - 12th.

Martin, R., Kusev, P., & Baranova, V. (2018). Leader ethical decision-making: Methodological suggestions and implications. Paper presented at the Ananyev Science Conference, St Petersburg State University, Russia, October 23rd-26th.

Martin, R., & Kusev, P. (2017). The influence of associative learning on moral decision-making. Paper presented at the 58th annual meeting of the Psychonomic Society, Vancouver, Canada. November 9th - 12th.

Martin, R., & Kusev, P. (2016). Rational choice predicted by utility ratio and uncertainty. Paper presented at the annual meeting of the Society for Judgment and Decision Making, Boston, Massachusetts, USA. November 18th - 21st.

Martin, R., & Kusev, P. (2016). How uncertainty and moral utilitarian ratios predict rationality. Paper presented at the 57th annual meeting of the Psychonomic Society, Boston, Massachusetts, USA. November 17th – 20th.