A Confucian Algorithm for Autonomous Vehicles

Tingting Sui[1] · Sebastian Sunday Grève[1,2]

Abstract

Any moral algorithm for autonomous vehicles must provide a practical solution to moral problems of the trolley type, in which all possible courses of action will result in damage, injury, or death. This article discusses a hitherto neglected variety of this type of problem, based on a recent psychological study whose results are reported here. It argues that the most adequate solution to this problem will be achieved by a moral algorithm that is based on Confucian ethics. In addition to this philosophical and psychological discussion, the article outlines the mathematics, engineering, and legal implementation of a possible Confucian algorithm. The proposed Confucian algorithm is based on the idea of making it possible to set an autonomous vehicle to allow an increased level of protection for selected people. It is shown that the proposed algorithm can be implemented alongside other moral algorithms, using either the framework of personal ethics settings or that of mandatory ethics settings.

Introduction

The era of autonomous vehicles (AV) is dawning, and it brings with it a variety of difficult problems that urgently require solutions.[1] From an engineering perspective, these problems include how to prevent hardware failures as well as the avoidance of implementing and how to avoid introducing software bugs, perceptual errors, or errors of reasoning errors.[2] Another possible type of error is of an ethical nature. For example, AVs may face choices where every feasible action leads to injury or loss of

---

[1] Two useful recent surveys are Feng et al., 2021 and Parekh et al., 2022.

[2] See, for example, Waldrop 2015, Fraichard and Kuffner 2012.

✉ Tingting Sui
     suitngting@pku.edu.cn

[1]   Chinese Institue of Foreign Philosophy, Peking University, 5 Yiheyuan Road, Beijing, China

[2]   Department of Philosophy, Peking University, Beijing, China

Ⓐ Springer

life. In general, in these kinds of situations an agent's course of action requires ethical evaluation (Evans, 2020Vallor, 2021; Sullins, 2021). Specifically, how should AVs be programmed to act in such contexts? Should they alter their path to shift the risk of harm from multiple individuals onto a single person? The facet of programming that dictates AV behaviour is commonly referred to as a "moral algorithm", which is typically implemented as a top-down approach (Himmelreich, 2022; Lucifora et al., 2020).[3] From the perspective of AV architecture, it influences decisions related to route optimization, hazard avoidance, and risk distribution which related to the tactical planning and AD functions.

In virtue of its simplicity and quantifiability, the ethical theory of utilitarianism—which reduces what is morally right to maximising the sum of positive outcomes overall—has been widely adopted by theorists as a starting point for the development of moral algorithms, including in the case of autonomous vehicles.[4] From a general philosophical perspective, the most fundamental problem is knowing what is morally right in the first place. Moreover, AVs are engineered to operate globally, serving users from diverse cultural backgrounds. This highlights the importance of incorporating intercultural approaches into their design. By doing so, designers can create AVs that offer a more diverse set of choices, aligning with the values and preferences of users from various cultural contexts (Hongladarom & Bandasak, 2023; Segun, 2021).

In this context, practical ethics—embracing moral algorithm development—offers an experimental route, readily applying established ethical theories to specific, well-defined scenarios. This method helps us test the acceptance of ethical rules toward the trolley problem more easily by using real data and examples (Sui, 2023).

A seminal study on AVs (Bonnefon et al., 2016) in the U.S. revealed that while a majority deemed utilitarianism morally acceptable in typical trolley-type dilemmas for AVs, a significant proportion within the same cohort expressed reluctance to purchase AVs employing this principle.[5] Some also disapproved of laws permitting AVs to act utilitarianly. This paradox mainly stems from individuals' aversion to personal harm, even when ethical principles prioritize the greater good. Interestingly, the study consistently observed that participants displayed concern for the well-being of their family members. This observation implies that people may anticipate a moral norm that places an emphasis on familial concern. From an intercultural approach, this expectation aligns with the principles of Confucianism, an Eastern philosophy that prioritizes family values and relationships.

In what follows, we first (in Sect. "Experimental Survey") report the results of a psychological study that we conducted, which supports the hypothesis of partiality towards family members and its ethical significance for the development of moral

---

[3] For additional details on the connection between top-down and moral algorithm, see Evans 2021.

[4] See, for example, Anderson et al., 2004, Anderson & Anderson, 2011, Bonnefon et al., 2016, and Lucifora et al., 2020. A Rawlsian algorithm for autonomous vehicles can be found in Leben 2017. For other moral algorithms, which are not designed specifically for autonomous vehicles, see Powers 2006 (Kantianism), Wallach and Allen 2010 (virtue ethics), and Anderson and Anderson 2011 (*prima facie* duties).

[5] This partial result is in line with Hauser et al., 2007.

algorithms in autonomous vehicles. In Sect. "Confucian Ethics", we describe relevant features of Confucian ethics, which, we argue, offers the most adequate solution to this problem. In Sect. "The Confucian Algorithm", we propose how an algorithm could be formulated in line with this ethics. In Sect. "Discussion", we offer further discussion of a number of technical, ethical, and legal issues.

The main reason why none of the major ethical systems in the Western tradition can easily account for partiality towards family members is that, unlike Confucianism, they do not treat this partiality as a fundamental element of morality, but tend to treat it instead as merely a problem of human moral psychology.[6]

Experimental Survey

In this section, we report our empirical findings concerning partiality towards family members in trolley-type dilemmas in relation to autonomous vehicles. This methodology is rooted in the tradition of experimental philosophy, which advocates for the empirical examination of individuals' moral judgments when asserting them as premises.

A previous study (Bonnefon et al., 2016) reported similar findings based on a survey of exclusively US respondents and scenarios in which family members only played the role of passengers within autonomous vehicle. We believe that these previous results are robust. Our study complements these results. We exclusively surveyed people in China, introducing moral dilemmas in both trolley scenarios and autonomous vehicle scenarios. This approach was undertaken to minimize the likelihood of divergent preferences among individuals in the context of the traditional trolley problem and scenarios involving autonomous vehicles. To address potential order effects, we administered two distinct surveys, each comprising 200 participants, organized into separate groups.

The experimental data was collected through Credamo, an online third-party professional research platform. With the approval of participants, we looked up their IP address location in order to guarantee ensure that the data coverage was sufficiently spread out across adequately covered the target region, China.

Design

Trolley Dilemma

In "*Trolley*" group, each survey participant was presented with the same task. The task was to answer the same two questions regarding each of three scenarios, one after the other, in the sequence given below, with the latter two scenarios being variations of the first. The first scenario is a standard trolley dilemma:

---

[6] A notable exception in modern Western ethics is the feminist approach known as care ethics.

> The brake of a runaway trolley barrelling down the railway track suddenly fails. Ahead, there are five workers working on the track. If nothing is done, the five workers will be crushed and killed. You are standing near the railway track, next to a switch. The only way to save the five workers is to pull the switch and divert the trolley onto a side track. However, a worker on the side track would be crushed and killed.

After reading the vignette, participants were asked to choose whether they would pull the switch ("yes" or "no"). In addition, they were asked to rate the moral acceptability of the "yes" option (i.e. to kill the one and save the five) on a seven-point scale ranging from "totally unacceptable" to "totally acceptable". The other two scenarios replicated the first except that in the last sentence "worker on the side track" was replaced by "you, who are standing on the side track," and "one of your family members, who is standing on the side track," respectively.[7]

A total of 200 participants were surveyed, of which 122 identified as female and 78 as male. Originally, 209 respondents completed the survey, but nine of these responses were excluded from the data set, because four of them gave incorrect answers to an easy mathematical question for control $(20+40=?)$, whilst another five completed their responses in a suspiciously fast manner.

Autonomous Vehicle Dilemma

In "*AV*" group, the survey procedure mirrored that of the "Trolley" group. Each participant in this group was assigned an analogous task, requiring them to respond to the same two questions pertaining to each of three scenarios consecutively. The sequence followed for these scenarios was as follows, with the latter two scenarios representing variations of the initial one. The initial scenario presented was the trolley dilemma within the context of an autonomous vehicle scenario:

> The brake of an autonomous vehicle running at a high speed on the highway suddenly fails. Ahead, there are five pedestrians. If nothing is done, the five pedestrians will be crushed and killed. The only way to save the five pedestrians is to swerve the vehicle. However, this redirection will result in the death of one pedestrian on the turnoff.

After reading the vignette, participants were asked to choose whether they would buy the vehicle that choose to kill one and save the five ("yes" or "no"). In addition, they were asked to rate the moral acceptability of this option (i.e. to sacrifice the one and save the five) on a seven-point scale ranging from "totally unacceptable" to "totally acceptable". The other two scenarios replicated the first except that in the last sentence "will result in the death of one pedestrian on the turnoff" was replaced by "will lead the vehicle off the cliff, resulting in the death of yourself as a passenger in the vehicle," and "one of your family members, who is a passenger in the vehicle," respectively.[8]

---

[7] The full version of the survey is given in an appendix to this article. The appendix and research data can be found at https://osf.io/nku4w/.

[8] The full version of the survey is given in an appendix to this article. The appendix and research data can be found at https://osf.io/nku4w/.

A comprehensive survey was administered to a total of 200 participants, with 126 self-identifying as female and 74 as male. Initially, 212 respondents completed the survey; however, seventeen of these responses were omitted from the dataset. This exclusion criteria comprised seven participants who provided inaccurate responses to a basic mathematical question $(20 + 40 = ?)$, and an additional five participants who exhibited unusually rapid completion times, suggesting a potential lack of engagement with the survey content.

Analysis

For reasons of brevity, we will call the three different scenarios (and associated tasks) *Family*, *Self*, and *Other* respectively.[9] The most striking result is perhaps that participants in "*Trolley*" and "*AV*" groups both displayed a relatively strong tendency against killing a family member, which was even stronger than that against killing themselves.

In "*Trolley*" group, 66.5% of people chose to kill the person on the side track and save the five in "*Other*" scenario. In "*Self*" scenario, this percentage dropped to 54%. In "*Family*", it dropped to 33.5%; hence, less than half the figure for *Other*. In fact, *Family* is the exact converse of *Other* in this respect. In "*AV*" group, 58.00% of people choose to buy the AV that sacrifice one pedestrian to save the five. However, in "*Self*" scenario, this percentage decreased to 42%. In "*Family*", the percentage further dropped to 37%. Despite the differences of percentage, the analysis revealed no significant differences between the corresponding scenarios in the two groups (see Fig. 1).Applying Pearson's chi-squared test confirms that the results observed in both groups are statistically significant (*Trolley* group: *Other* vs *Self*, χ^2 (1, 200) $= 42.47$, $p < 0.01$; *Self* vs *Family*, χ^2 (1, 200) $= 83.06$, $p < 0.01$; *Family* vs *Other*, χ^2 (1, 200) $= 16.88$, $p < 0.01$; *AV* group: *Other* vs *Self*, χ^2 (1, 200) $= 58.19$, $p < 0.01$; *Self* vs *Family*, χ^2 (1, 200) $= 133.38$, $p < 0.01$; *Family* vs *Other*, χ^2 (1, 200) $= 59.89$, $p < 0.01$). The data indicates that individuals' choices consistently demonstrate their reluctance to sacrifice family members.

Answers to the question of moral acceptability show that participants tended to consider killing a family member to be the least morally acceptable thing to do in these types of scenarios, as compared with killing oneself or killing one worker/pedestrian. In fact, participants in both groups tended to consider killing a family member to be the only morally unacceptable option, as indicated by responses falling below 4, a threshold of neutral on the moral acceptance scale (*Trolley* group: $M = 3.31$, $SD = 2.07$; *AV* group: $M = 3.96$, $SD = 1.92$). Notably, the *AV* group exhibited a higher level of moral acceptance toward sacrificing both pedestrians and family members compared to the trolley group. The data implies that there may be a nuanced divergence in moral judgments, with people displaying a more permissive attitude towards the moral implications of AV decisions while maintaining a

---

[9] Notice that the name "Other" here should not be taken to imply that the worker on the side track in that scenario is *not* a family member. It is possible that the worker could be a family member. However, it seems acceptable to disregard this possibility in the present context.

steadfast moral boundary against harming their own family members (see Figs. 2a and b).[10]

Applying ANOVA and paired t-tests confirms that the results we claimed are statistically significant between groups (*Other*: $F(1, 199) = 21.37$, $p < 0.001$; *Self*: $F(1, 199) = 0.09$, $p = 0.776$; *Family*: $F(1, 199) = 10.55$, $p < 0.001$) and within groups (*Trolley* group: *Other* vs *Self*, $t(1, 199) = -4.84$, $p < 0.001$; *Self* vs *Family*, $t(1, 199) = 9.14$, $p < 0.001$; *Family* vs *Other*, $t(1, 199) = -3.75$, $p < 0.001$; *AV* group: *Other* vs *Self*, $t(1, 199) = 1.03$, $p = 0.305$; *Self* vs *Family*, $t(1, 199) = 4.52$, $p < 0.001$; *Family* vs *Other*, $t(1, 199) = 4.97$, $p < 0.001$).

Although respondents exhibited a stronger tendency to consider swerving the trolley or AV morally acceptable in *Self* as compared with that of in *Other*, fewer people answered they would actually do it in *Self* than in *Other*. By contrast, no such disparity between reported moral ideals and practical dispositions can be observed for *Family*. On the contrary, swerving the trolley or AV in *Family* was both considered the least morally acceptable thing to do and was what the smallest number of people said that they would actually do.

Notably, our surveys do not aim to prove the superiority of moral partiality over impartiality in ethics. Instead, its objective is to illuminate the practical necessity for partiality, especially within the context of family relationships. This approach serves to mitigate the risk of proposing a morally justified theory that may encounter practical rejection, akin to challenges faced by utilitarianism. This precaution is taken to avoid the potential discrepancy between a theoretically sound ethical stance and its real-world applicability, as mentioned in Keeling's work from 2019. While some philosophers may reject the moral significance of partiality from an ethical perspective, many others find it plausible to argue that partiality towards certain individuals is not only permissible but, at times, morally commendable and obligatory (Nagel, 1991; Scanlon, 1998; Young, 2023). Philosophers such as Bernard Williams and T.M. Scanlon contribute to the defence of moral partiality by endorsing relationship-based views. Williams argues that "there are certain respects in which creatures are treated in one way rather than another simply because they belong to a certain category, the human species" (2008). Scanlon, in his Contractualism, justifies the significance of ties based on shared humanity, rejecting the characterization of this stance as mere speciesism. According to Scanlon, our relational ties with beings 'of human born' provide a robust reason to accord them the same moral status as other humans, recognizing that this reason may not apply to beings with comparable capacities outside the human species (1978; 1998). These discussions may not establish the optimization of moral partiality over impartiality. However, they strongly suggest the viability of conceptualizing moral partiality as a practical and morally defensible moral algorithm. The recognition that moral partiality can be morally justified and, crucially, practically accepted, underscores its potential as a guiding framework for ethical decision-making.

---

[10] Figure 2a presents the mean and standard deviation for moral acceptance, while Fig. 2b illustrates the distribution of participants' choices.
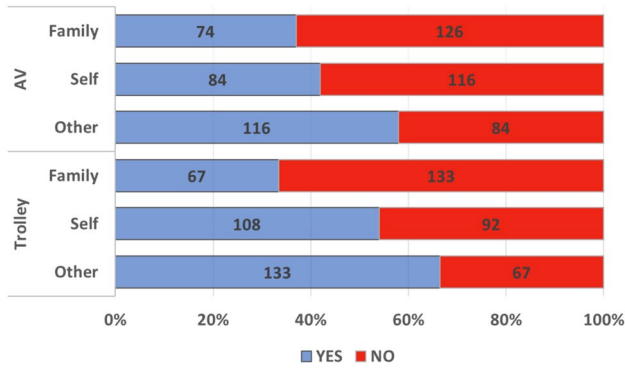
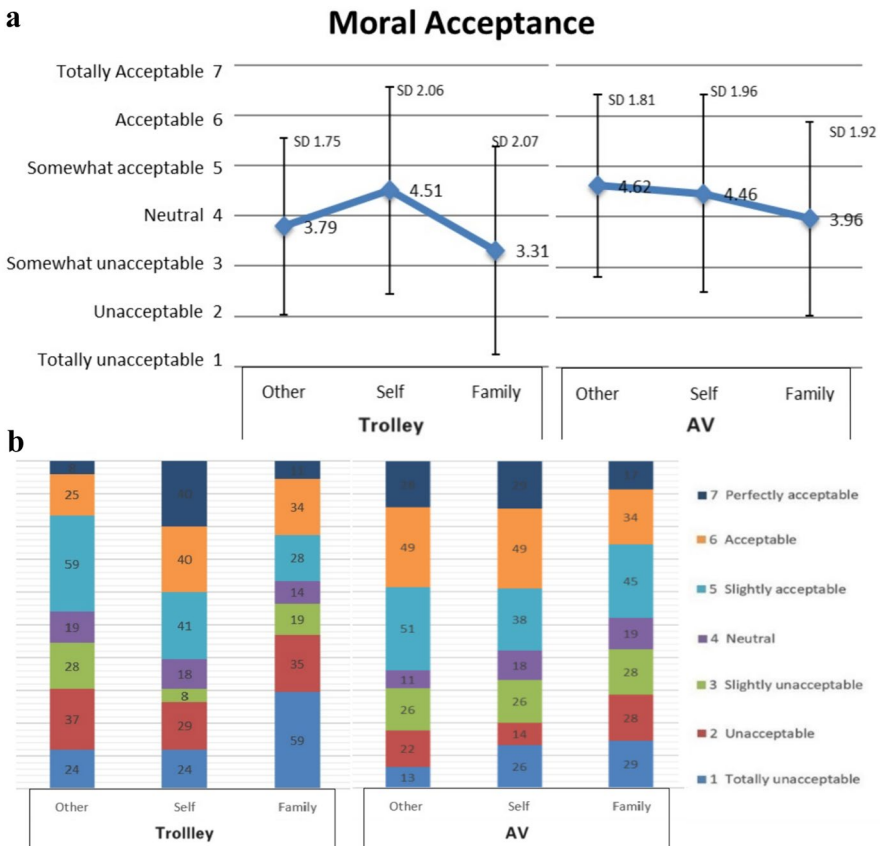**Fig. 1** "Would you pull the switch?"/ "Would you buy the AV?"

Fig. 2 Means (M) and standard deviations (SD). **b** "Is pulling the switch morally acceptable?"/ "Is the AV morally acceptable?".

Confucian Ethics

We believe that Confucianism offers the most adequate ethical framework to account for the reported partiality towards family members and the reported tendency to consider this partiality to be morally acceptable (in particular, the reported tendency to consider acting against this partiality, in our *Family* scenario, to be morally unacceptable, which corroborates previous findings by Bonnefon et al., 2016, as explained in Sects. "Introduction" and "Experimental Survey" above).[11]

The concept that perhaps best describes the key feature of Confucianism in this connection can be translated as "familial love" (*qīn qīn zhī ài* 亲亲之爱 or, simply, *qīn qīn* 亲亲). The Confucian concept of familial love may have something of a Western counterpart in the ancient Greek *storgḗ* στοργή, which has sometimes been translated the same way into English, but as far as we are aware the Confucian theory that surrounds the Chinese term is quite different from any theoretical account in the Western tradition.

Comparing with Utilitarianism, which offers a decision calculus based on quantifiable factors, Confucianism places a robust emphasis on roles, relationships, and virtue. It revolves around the fulfilment of various roles such as parent, child, and friend, emphasizing the importance of promoting harmony within social structures. Confucianism conceives of familial love as a natural relation between individuals, an effective mode of socialisation, and a fundamental element of morality. Both the claimed naturalness and the claimed strength of this relation between individuals find a powerful and intuitive expression in the *Mencius*—one of the Sishu (Four Books) of Confucianism—when Mencius asks his interlocutor: "Does Adept Yi really believe we can love [*qīn* 亲] a neighbour's newborn child the way we love our own brother's child?" (*Mencius* 3A, 5).[12] Mencius thought that, in scenarios of this kind, the norms of morality should not be against human nature, and that it was more plausible to expect people to conform to moral norms if these encouraged them first to admit their natural love of, and partiality towards, their own family, and then, on this basis, also to extend that love towards the rest of society.[13]

The renowned Chinese sociologist Fei Xiaotong called the corresponding kinship-based mode of socialisation, which has arguably been dominant in Chinese culture for thousands of years, "the differential mode of association" (*chà xù gé jú* 差序格局).[14] In this mode of association, the constitution and growth of an individual's social relations is an extension of the child's natural human relationships, which will typically begin with immediate family members (mother, father, siblings) and then extended family members (grandparents, aunts, uncles, cousins, etc.). These natural human relationships, which have their root in the immediate family, are what Fei calls "kinship". As he writes:

---

[11] One useful general introduction to Confucianism is Yao 2000.

[12] The Sishu is the series of four ancient Confucian texts that formed the official subject matter for civil service examinations in China between 1313 and 1905. They are, in order: *Daxue* ("Great Learning"), *Zhongyong* ("Doctrine of the Mean"), *Lunyu* ("Conversations" or "Analects"), and *Mencius* (named after the Confucian scholar Mencius).

[13] See also Yao 1995.

[14] See Fei 1948/92, chapter 4.

In Chinese society, the most important relationship—kinship—is similar to the concentric circles formed when a stone is thrown into a lake. Kinship is a social relationship formed through marriage and reproduction. The networks woven by marriage and reproduction can be extended to embrace countless numbers of people—in the past, present, and future. The same meaning is implied in our saying "Everyone has a cousin three thousand miles away," with three thousand miles indicating the vastness of kinship networks. (Fei, 1948/92, p. 63)

Fei illustrates the further extension of an individual's social relations, beyond kinship so defined, as follows:

Every family regards its own household as the center and draws a circle around it. This circle is the neighborhood, which is established to facilitate reciprocation in daily life. A family invites the neighbors to its weddings, sends them red-dyed eggs when a new baby is born, and asks for their help in lifting its dead into coffins and carrying the coffins to the cemetery. (Fei, 1948/92, p. 64)

In this way, then, familial love is the basic natural relation, whose differentiation when extended beyond kinship is what constitutes a healthy society on the Confucian model. Consequently, familial love is conceived as the fundamental principle of morality in Confucian ethics. The *Zhongyong* (Doctrine of the Mean)—another of the Sishu (Four Books)—famously says: "To become consummate in your conduct is to become a person, where devotion to your family [*qīn qīn* 亲亲] is the most important factor" (*Zhongyong* 20).[15] Roger Ames gives the following useful summary of the system of values that results from this directive:

Kinship as the root of human relations is defined by the moral imperatives of "family reverence" (*xiào* 孝) and "fraternal deference" (*tì* 悌). And friendship as the way of extending this pattern of kinship relations to include non-relatives is pursued through an ethic of "putting oneself in the place of others" (*shù* 恕), "commitment and resolve" (*chéng* 誠), "doing one's utmost" (*zhōng* 忠), and "making good on one's word" (*xìn* 信). All such ethical values are aspired to as the way of reconciling the tensions among and promoting the accommodations made within the specific personal relationships of family members and community. (Ames, 2022, p. 361)[16]

Thus, Confucianism takes partiality towards family members as a natural given and, at the same time, as inherently morally acceptable in virtue of its being the very foundation of social cohesion and morality.[17]

---

[15] The original reads: *Rén zhě rén yě, qīn qīn wéi dà* 仁者人也, 亲亲为大.

[16] See also, for example, *Lunyu* (Analects) 1.4 and 1.8. For a recent discussion of the place of values such as justice, fairness, and equality in Confucianism, see Lambert 2020.

[17] Confucian family partiality is fundamentally centered on familial relationships, virtues, and duties. Its specific considerations of gender, age, and social standpoint can vary significantly based on individual interpretations and the cultural context in which Confucian values are applied. Due to the inherent controversy and diversity in these interpretations, we refrain from incorporating them into our moral algo-

Notably, Confucian principles generally discourage causing harm to others. In a non-dilemma scenario, Confucianism could advocate for AVs to enhance their understanding of serving the familial role of owners. This could involve adapting to the family's preferences, such as opting for a scenic route instead of the shortest one, thereby acknowledging and accommodating the family's distinctive preferences and values. However, in moral dilemma such as trolley scenario that presents a challenging situation where harm appears unavoidable, the Confucian perspective resonates with the Doctrine of Double Effect, recognizing that harm to others is a foreseen but unintended consequence of prioritizing the well-being of one's family members.[18] Meanwhile, although Confucianism shows devotion to family, it also advocates the extension of love beyond familial bonds, suggesting that individuals can cultivate a sense of care and consideration for people beyond their family. In this context, the application of Confucian principles also implies that individuals, while acknowledging the importance of family relationships, are not bound to offer special protection exclusively to family members. This is why we provide people with choices regarding the extent to which they would like to protect their family members in our design. From the perspective of societal implication, this design can preserve people's driving preferences to a large extent, that is, autonomous vehicles remain largely consistent with human-driven vehicles. The driving model's continuity minimizes the psychological gap between human-driven and autonomous vehicles, contributing to a more seamless integration of this transformative technology into daily life. However, this approach may raise concerns about privacy, specifically the need for universal face recognition to distinguish family members from non-family members. We believe this issue can be addressed through technical means, such as claiming the presence of family members inside the car or utilizing unique Bluetooth devices paired with AVs (see more detail in Sect. "The Confucian algorithm"). Hence, we posit that incorporating Confucian partiality towards family into a moral algorithm is a commendable approach.

At this point, it may be objected that our proposed solution (namely, that Confucianism offers the most adequate ethical framework within which to resolve the issue of the development of moral algorithms for autonomous vehicles) appears to require substantive agreement, perhaps even on the part of prospective vehicle users, not merely with some observations about what people tend to consider morally right or wrong but with an entire theory of what could possibly constitute what really is morally right or wrong, or even with an entire culture that may be different from our own. However, it would be mistaken to believe that our proposed solution requires any such thing. Rather, the sole intention behind the preceding introduction of Confucianism was to offer an ethical framework whose theoretical application account for the reported partiality towards family members in trolley-type scenarios and the reported tendency to consider this partiality to be morally acceptable. We

---

Footnote 17 (Continued)

rithm. In other words, our design is exclusively based on family bonds rather than considerations of gender, age, or social standpoint.

[18] Doctrine of Double Effect (DDE) is an ethical principle that distinguishes between intended and foreseen consequences in moral decision-making, see Foot 1967.

are maintaining only that relevant parts of Confucian theory as we have presented it, including relevant bits of moral psychology, can be used for this purpose without anyone having to subscribe to Confucianism. The relevant ideas could equally have been formulated within the framework of utilitarianism, albeit perhaps not at quite such a fundamental level. In this sense, it is a mere historical fact that these ideas must be credited to Confucianism. But it is important to us that we should give a historically accurate account here.[19]

The Confucian Algorithm

We now turn to the matter of the technical implementation of the psychological results reported earlier, specifically the tendencies to protect family members more than other people and to judge this to be morally acceptable. The chief task of this section will be to formulate a corresponding decision procedure. Since we found that Confucianism offers the most adequate ethical framework within which to accommodate the psychological results, we have chosen to call this a "Confucian" algorithm. In what follows, we briefly outline the basic mathematical formulation and principles of engineering for a Confucian algorithm for autonomous vehicles that encounter trolley-type scenarios of the sort discussed above.

We propose that such a vehicle could be set to allow an increased level of protection for selected people.[20] Mathematically speaking, we can formulate a decision procedure based on comparing the values of a utility function Uti $(…)$—taking decisions D_i and objects $O_{m,n}$ as arguments—and a distinction between two different types of objects $W_j a_j$ and $W_k b_k$, that is, selected people a_j and non-selected people b_k, each together with an assignment of a weight ranging from 0 to 1, W_j and W_k respectively, corresponding to different levels of protection. The utility function would thus read as follows:

$$Uti\left(D_i, O_{m,n}\right) = \sum_{j=1}^{m} W_J a_j + \sum_{k=1}^{n} W_k b_k$$

We assume that starting from the set of possible decisions D_i gives the relevant sets of objects $O_{m,n}$. This is to assume the availability of data that would be analogous to our knowing in the *Family* scenario described above that pulling the switch will result in the death of exactly one family member and not pulling the switch will result in the deaths of exactly five non-family members.

Consider an example. Suppose that, in analogy with our *Family* scenario, there are only two possible decisions for an autonomous vehicle: D_1 (changing course)

---

[19] It may be thought that the ideal of fraternal love in the Western religious tradition could have served essentially the same purpose. However, this is not the case. For example, the most important expression of that ideal in the Bible reads, in English translation, "Thou shalt love thy neighbour as thyself" (Leviticus 19:18, Matthew 19:19, 22:39, Mark 12:31, Romans 13:9, Galatians 5:14, James 2:8). One important difference is that this relational structure does not seem to assume as its starting point partiality towards one's family but, instead, partiality towards oneself.

[20] We address the debate over mandatory ethics settings (MES) vs personal ethics settings (PES) in Sect. "Discussion".

and D_2 (not changing course). D_1 will result in the death of exactly one selected person a_1, whilst D_2 will result in the deaths of exactly five non-selected people (b_1–b_5). Suppose further that the weight of a_1 is set at 0.9, while the weight of non-selected people is set at 0.1. This means that the objects in $O_{[D1]}$ are the five 0.1-weighted non-selected people ($0.1*b_1$, $0.1*b_2$, etc.), and the only object in $O_{[D2]}$ is the 0.9-weighted selected person ($0.9*a_1$). Now the utility function will calculate the corresponding values for D_1 and D_2 as follows:

$$Uti(D_1, O_{[D1]}) = 0.9 * 0 + 0.1 * 5 = 0.5$$

$$Uti(D_2, O_{[D2]}) = 0.9 * 1 + 0.1 * 0 = 0.9$$

Given its higher utility value, D_2 will be the output of the algorithm in this case.

The flowchart below (Fig. 3) illustrates the rest of the decision procedure, which follows the calculation of utility values. If the scenario involves the comparison of a large number of possible decisions D_i, then a sorting algorithm can be used to determine the decision with the highest utility value. If more than one possible decision receives the highest utility value, then the output will be determined randomly. Additionally, if there is no family member involved in the scenario, the AV would treat passengers and pedestrians equally. In this case, it worked similar to utilitarian scheme.

The software interface used to configure protection settings for selected people need not be complicated. A simple method is to declare the presence of family members inside the car directly and set the desired protection level. This streamlined approach could address potential privacy concerns and enhance user-friendliness. However, considering the possibility that individuals may falsely declare the presence of a family member when none is actually inside the car, an alternative method involves requesting users to wear some kind of Bluetooth devices that enable the car to identify the person. It's crucial to note that this approach may raise concerns about the exposure of privacy information. To address these challenges, we strongly recommend the implementation of advanced algorithms with robust privacy protection features. These algorithms should incorporate various techniques, such as encryption to secure data access only to authorized parties, anonymization to obscure personally identifiable information (PII) in datasets, and the introduction of controlled noise or randomness to prevent the extraction of individual contributions from datasets. Additionally, users should have the ability to turn off the Bluetooth device at any time. While we offer these suggestions, we acknowledge the complexity of the technological aspects involved and trust that engineers will utilize their expertise to develop more effective and appropriate solutions towards the potential privacy problems. Figure 4 below gives a schematic illustration. A number of associated technical, ethical, and legal issues are discussed in the next section (Sect. "Discussion").
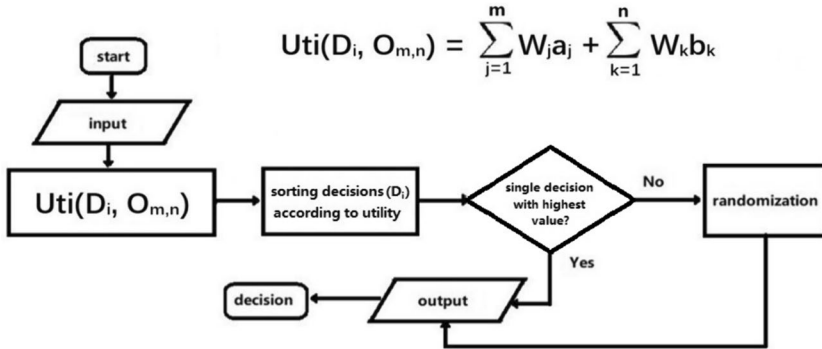
Fig. 3 Flowchart for the Confucian algorithm and utility function

Discussion

We noted earlier, at the end of Sect. "Experimental Survey", the strong disparity between reported moral ideals and reported practical dispositions with regard to the *Self* scenario in our empirical study. This result is in line with an earlier study (Bonnefon et al., 2016), which found that people have a tendency to consider the decision by a utilitarian algorithm in an autonomous vehicle that corresponds to pulling the switch in *Self* (i.e. killing the passenger where that passenger is the respondent themselves) to be morally acceptable but, at the same time, would be unwilling to buy an autonomous vehicle that applies utilitarianism in this way. The same work demonstrated a similar problem for the case that corresponds to pulling the switch in our *Family* scenario (i.e. killing a family member), which our study has corroborated. But we further noted that, for the *Family* scenario, there was no such disparity between reported moral ideals and practical dispositions. Rather, we found that the reported high level of partiality towards family members was matched by a strong tendency to consider this partiality morally acceptable.[21] To reiterate, the main reason why none of the major ethical systems in the Western tradition can easily account for this partiality towards family members and its apparent moral acceptability is that, unlike Confucianism, they do not treat this partiality as a fundamental element of morality. On the contrary, the Western tradition has tended to treat partiality towards family members as merely a psychological problem that stands in the

---

[21] Some may argue that sacrificing the lives of five individuals could disrupt the relational dynamics of five families and their broader networks of friends. In other words, family inclination might advocate for the protection of these individuals from a broader perspective. However, we are under the premise that the moral algorithm must be accepted by car owners, as demonstrated by previous surveys on utilitarianism. Therefore, we focus our discussion on moral partiality from individual moral intuition, specifically considering the views of car owners.
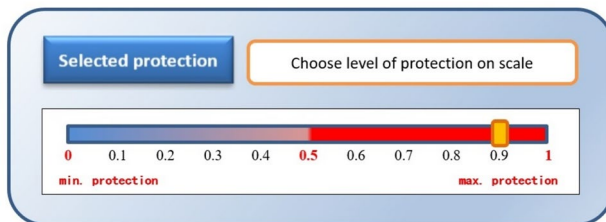
Fig. 4 Sample software interface for configuring protection settings

way of certain ethical ideals, most obviously perhaps the conception of equality as general impartiality.[22]

It is therefore not surprising that existing moral algorithms for autonomous vehicles do not tend to distinguish between different types of people, except perhaps along the lines of passengers vs. non-passengers. Nor is it surprising that, while Derek Leben's Rawlsian algorithm introduces a number of additional differential parameters, and the inventors of the "ethical knob" very briefly mention the protection of family members as a possible part of their "Egoist mode", neither of them even consider the possibility that partiality towards family members might not only be a desirable differential parameter to add to the algorithm of an autonomous vehicle but indeed an integral part of its moral algorithm.[23]

However, this does not mean that utilitarian, Kantian, Rawlsian, or other algorithmic solutions cannot in principle incorporate the relevant element of Confucian ethics. When it comes to issues of practical ethics, such as the one being considered here, it is often reasonable to work on a pluralist assumption. One does not first have to integrate the theoretical frameworks in their entirety. As long as their mixed application is practically consistent, significant progress may be achieved. As far as we can see, our Confucian algorithm will lend itself to easy integration into larger systems modelled on any major ethical framework.

The Confucian algorithm can be implemented as part of personal ethics settings or as part of mandatory ethics settings.[24] Given the psychological results reported in this article, and considering the reasons we have presented for taking partiality towards family members to be of fundamental moral significance, it follows that anyone advocating for personal ethics settings and against mandatory ones should seriously consider implementing the Confucian algorithm. Moreover, anyone advocating for the contrary position should equally consider implementing it. An

---

[22] Again, a notable exception in modern Western ethics is the feminist approach known as care ethics. Although there are many scholars such as Nagel or Scanlon who would like to justify moral partiality, they do not put forward moral norms mainly based on moral partiality.

[23] See Leben 2017 and Contissa et al., 2017, 370. Notably, Contissa et al. do not provide any details in this connection, apart from the revealing fact that they would regard protecting family members as belonging on the side of what they call "egoism"; moreover, their "egoistic mode" is in fact defined as an exclusive preference for passengers, so it cannot possibly give adequate protection to family members, who could also figure in these scenarios as non-passengers.

[24] On the debate about these different types of settings, see esp. Contissa et al., 2017, Gogoll and Müller 2017, and Himmelreich 2022.

important reason in favour of implementing the Confucian algorithm in any type of ethics setting is that, at least as long as autonomous vehicles are privately owned, users should probably retain some of the moral and legal rights and responsibilities that are attached to driving non-autonomous vehicles, including the moral right of protecting selected people, such as family members, more than others from harm caused by the vehicle they control (even if their control is limited to setting the vehicle's destination on a digital map).[25]

Notably, there will probably have to be limits to the number and types of people that can be added for extra protection in order to prevent abuse of the system. It is not obvious what the exact criteria should be, but, from both a moral and a legal perspective, the selection of people for extra protection will probably require their official consent. This requirement may be used to effectively impose a certain kind of limit. For example, if the Confucian algorithm is implemented as part of a personal ethics setting, a feasible solution may be a principle of mutual consent and protection, so that anyone giving special protection to someone else must also receive special protection from that person in return. If the algorithm is implemented as part of a mandatory ethics setting, giving special consent may not be necessary.

There are, of course, many other possible complications and necessary refinements that we cannot consider here, but we see no reason to think that the development and implementation of the Confucian algorithm is somehow in principle less desirable, less feasible, or less justified than that of any other proposed moral algorithms for autonomous vehicles.[26]

Conclusion

The advent of autonomous vehicles appears virtually inevitable. The problem of developing a moral algorithm for autonomous vehicles is therefore an urgent one. In particular, any moral algorithm for autonomous vehicles must provide a practical solution to trolley-type scenarios, in which all possible courses of action will result in damage, injury, or death. In this paper, we have reported psychological results that point to a hitherto neglected variety of this type of problem. Specifically, we have reported the findings of a new empirical study, which complement and corroborate earlier findings that similarly suggested widespread partiality towards family members and a widespread tendency to consider this partiality morally acceptable. We have argued that the resulting variety of trolley-type scenarios is important with regard to the development of moral algorithms for autonomous vehicles, and we have argued that the most adequate solution to this problem will be achieved by a moral algorithm that is based on Confucian ethics. We have outlined one such algorithm based on the principle that it should be possible to set an autonomous vehicle

---

[25] See also Millar 2015.

[26] It must be acknowledged that several ethical principles, such as Buddhism and Islamism, have not yet been incorporated into moral algorithms. Currently, we are unable to assess them or assume whether moral algorithms possess regional or national characteristics. These are questions beyond the scope of this article. However, we believe that these aspects merit further discussion in future research.

to allow an increased level of protection for selected people. And we have outlined how this algorithm might be implemented.

Appendix 1: Vignettes for Survey (English Translation)

"Trolley" Group

"Other" Scenario

The brake of a runaway trolley barrelling down the railway track suddenly fails. Ahead, there are five workers working on the track. If nothing is done, the five workers will be crushed and killed. You are standing near the railway track, next to a switch. The only way to save the five workers is to pull the switch and divert the trolley onto a side track. However, a worker on the side track would be crushed and killed.

In this scenario, would you pull the switch?

Yes   No.

In this scenario, would it be morally acceptable if you pulled the switch?

(1)   Totally acceptable
(2)   Acceptable
(3)   Somewhat acceptable
(4)   Neutral
(5)   Somewhat unacceptable
(6)   Unacceptable
(7)   Totally unacceptable

"Self" Scenario

The brake of a runaway trolley barrelling down the railway track suddenly fails. Ahead, there are five workers working on the track. If nothing is done, the five workers will be crushed and killed. You are standing near the railway track, next to a switch. The only way to save the five workers is to pull the switch and divert the trolley onto a side track. However, you, who are standing on the side track, would be crushed and killed.

In this scenario, would you pull the switch?

Yes   No.

In this scenario, would it be morally acceptable if you pulled the switch?

(1)   Totally acceptable
(2)   Acceptable
(3)   Somewhat acceptable
(4)   Neutral
(5)   Somewhat unacceptable

(6) Unacceptable
(7) Totally unacceptable

"Family" Scenario

The brake of a runaway trolley barrelling down the railway track suddenly fails. Ahead, there are five workers working on the track. If nothing is done, the five workers will be crushed and killed. You are standing near the railway track, next to a switch. The only way to save the five workers is to pull the switch and divert the trolley onto a side track. However, one of your family members, who is standing on the side track, would be crushed and killed.

In this scenario, would you pull the switch?

Yes   No.

In this scenario, would it be morally acceptable if you pulled the switch?

(1) Totally acceptable
(2) Acceptable
(3) Somewhat acceptable
(4) Neutral
(5) Somewhat unacceptable
(6) Unacceptable
(7) Totally unacceptable

"AV" Group

"Other" Scenario

The brake of an autonomous vehicle running at a high speed on the highway suddenly fails. Ahead, there are five pedestrians. If nothing is done, the five pedestrians will be crushed and killed. The only way to save the five pedestrians is to swerve the vehicle. However, this redirection will result in the death of one pedestrian on the turnoff.

In this scenario, would you buy this AV that choose to kill one and save the five?

Yes   No.

In this scenario, would it be morally acceptable if the AV did so?

(1) Totally acceptable
(2) Acceptable
(3) Somewhat acceptable
(4) Neutral
(5) Somewhat unacceptable
(6) Unacceptable
(7) Totally unacceptable

"Self" Scenario

The brake of an autonomous vehicle running at a high speed on the highway suddenly fails. Ahead, there are five pedestrians. If nothing is done, the five pedestrians will be crushed and killed. The only way to save the five pedestrians is to swerve the vehicle. However, this redirection will lead the vehicle off the cliff, resulting in the death of yourself as a passenger in the vehicle.

In this scenario, would you buy this AV that choose to kill one and save the five? Yes   No.

In this scenario, would it be morally acceptable if the AV did so?

(1) Totally acceptable
(2) Acceptable
(3) Somewhat acceptable
(4) Neutral
(5) Somewhat unacceptable
(6) Unacceptable
(7) Totally unacceptable

"Family" Scenario

The brake of an autonomous vehicle running at a high speed on the highway suddenly fails. Ahead, there are five pedestrians. If nothing is done, the five pedestrians will be crushed and killed. The only way to save the five pedestrians is to swerve the vehicle. However, this redirection will lead the vehicle off the cliff, resulting in the death of one of your family members, who is a passenger in the vehicle.

In this scenario, would you buy this AV that choose to kill one and save the five? Yes   No.

In this scenario, would it be morally acceptable if the AV did so?

(1) Totally acceptable
(2) Acceptable
(3) Somewhat acceptable
(4) Neutral
(5) Somewhat unacceptable
(6) Unacceptable
(7) Totally unacceptable

Vignettes for Survey (Chinese Original)

电车难题组
    "他人"情境
    一辆正在行使的有轨电车刹车突然失灵，轨道前方有五名正在作业的工人。如果不采取措施，5名工人将会丧命。在你的旁边有一个制动装置，救这5名工

人唯一的办法就是扳动制动装置, 让这辆失灵的电车驶向另一条轨道。但这样做, 在那条轨道上工作的另1名工人会被撞死。

你是否会扳动制动装置?

是　否

如果你扳动制动装置, 让电车驶向1人的轨道, 你觉得道德上可以接受吗?

(1)　非常能接受
(2)　能接受
(3)　比较能接受
(4)　中立
(5)　比较不能接受
(6)　不能接受
(7)　非常不能接受

"自我"情境

一辆正在行使的有轨电车刹车突然失灵, 轨道前方有五名正在作业的工人。如果不采取措施, 5名工人将会丧命。在你的旁边有一个制动装置, 救这5名工人唯一的办法就是扳动制动装置, 让这辆失灵的电车驶向你所在的轨道, 但如果电车变向, 你会被撞死。

你是否会扳动制动装置?

是　否

如果你扳动制动装置, 让电车驶向自己的轨道, 你觉得在道德上可以接受吗?

(1)　非常能接受
(2)　能接受
(3)　比较能接受
(4)　中立
(5)　比较不能接受
(6)　不能接受
(7)　非常不能接受

"亲人"情境

一辆正在行使的有轨电车刹车突然失灵, 轨道前方有五名正在作业的工人。如果不采取措施, 5名工人将会丧命。在你的旁边有一个制动装置, 救这5名工人唯一的办法就是扳动制动装置, 让这辆失灵的电车驶向另一条轨道。但你的亲人恰好在那一条轨道上, 如果电车变向, 你的亲人会被撞死。

你是否会扳动制动装置?

是　否

如果你扳动制动装置驶向1个人的轨道, 你觉得在道德上可以接受吗?

(1)　非常能接受
(2)　能接受
(3)　比较能接受
(4)　中立
(5)　比较不能接受

(6)   不能接受
(7)   非常不能接受

自动驾驶组
"他人"情境
一辆在高速公路上高速行驶的自动驾驶汽车刹车突然失灵。车辆前方有五个行人。如果不采取措施，五个行人会被撞死。拯救五名行人的唯一方法是让车辆转弯。然而,转弯将车辆掉下悬崖,车上那一名乘客 (你自己) 会因此死亡。

你是否会购买这种牺牲一人救五人的自动驾驶车?
是   否
如果自动驾驶车牺牲那一个人,你觉得在道德上可以接受吗?

(1)   非常能接受
(2)   能接受
(3)   比较能接受
(4)   中立
(5)   比较不能接受
(6)   不能接受
(7)   非常不能接受

"自我"情境
一辆在高速公路上高速行驶的自动驾驶汽车刹车突然失灵。车辆前方有五个行人。如果不采取措施，五个行人会被撞死。拯救五名行人的唯一方法是让车辆转弯。转弯将车辆掉下悬崖, 作为车上那一名乘客 (你的亲人) 会因此死亡。

你是否会购买这种牺牲一人救五人的自动驾驶车?
是   否
如果自动驾驶车牺牲那一个人,你觉得在道德上可以接受吗?

(1)   非常能接受
(2)   能接受
(3)   比较能接受
(4)   中立
(5)   比较不能接受
(6)   不能接受
(7)   非常不能接受

"亲人"情境
一辆正在行使的有轨电车刹车突然失灵,   轨道前方有五名正在作业的工人。如果不采取措施, 5名工人将会丧命。在你的旁边有一个制动装置, 救这5名工人唯一的办法就是扳动制动装置, 让这辆失灵的电车驶向另一条轨道。但你的亲人恰好在那一条轨道上,如果电车变向,你的亲人会被撞死。

你是否会购买这种牺牲一人救五人的自动驾驶车?
是   否

如果自动驾驶车牺牲那一个人, 你觉得在道德上可以接受吗?

(1) 非常能接受
(2) 能接受
(3) 比较能接受
(4) 中立
(5) 比较不能接受
(6) 不能接受
(7) 非常不能接受

Declarations

Conflict of interest The authors declare that neither of them have any conflict of interest.

References

Ames, R. T. (2022). *A conceptual lexicon for classical Confucian philosophy*. SUNY Press.

Anderson, M., Anderson, S. L., & Armen, C. (2004). Towards machine ethics. AAAI-04 Workshop on Agent Orientations: Theory and Practice.

Anderson, S. L., & Anderson, M. (2011). A prima facie duty approach to machine ethics and its application to elder care. Human–Robot Interaction in Elder Care: Papers from the 2011 AAAI Workshop (WS-11–12). https://doi.org/10.5555/2908724.2908725

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science, 352*(6293), 1573–1576.

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. https://doi.org/10.1007/s10506-017-9211-z

Evans, K. (2021). The implementation of ethical decision procedures in autonomous systems: The case of the autonomous vehicle [DOCTEUR DE L'UNIVERSITÉ SORBONNE UNIVERSITÉ, Sorbonne Université].

Evans, K., de Moura, N., Chauvier, S., Chatila, R., & Dogan, E. (2020). Ethical decision making in autonomous vehicles: The AV ethics project. *Science and Engineering Ethics, 26*, 3285–3312.

Fei, X. (1948/92). *From the soil: The foundations of Chinese society* (trans. G. G. Hamilton and W. Zheng), (Originally published 1948). University of California Press.

Feng, S., Yan, X., Sun, H., Feng, Y., & Liu, H. X. (2021). Intelligent driving intelligence test for autonomous vehicles with naturalistic and adversarial environment. *Nature Communications, 12*(1), 1–14. https://doi.org/10.1038/s41467-021-21007-8

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

Fraichard, T., & Kuffner, J. J. (2012). Guaranteeing motion safety for robots. *Autonomous Robots, 32*(3), 173–175. https://doi.org/10.1007/s10514-012-9278-z

Gogoll, J., & Müller, J. F. (2017). Autonomous cars: In favor of a mandatory ethics setting. *Science and Engineering Ethics., 23*, 681–700. https://doi.org/10.1007/s11948-016-9806-x

Hauser, M., Cushman, F., Young, L., Kang-Xing Jin, R., & Mikhail, J. (2007). A dissociation between moral judgments and justifications. *Mind & Language, 22*(1), 1–21. https://doi.org/10.1111/j.1468-0017.2006.00297.x

Himmelreich, J. (2022). No wheel but a dial: Why and how passengers in self-driving cars should decide how their car drives. *Ethics and Information Technology, 24*(4), 1–12. https://doi.org/10.1007/s10676-022-09668-5

Hongladarom, S., & Bandasak, J. (2023). Non-western AI ethics guidelines: Implications for intercultural ethics of technology. *Ai & Society*. https://doi.org/10.1007/s00146-023-01665-6

Keeling, G., Evans, K., Thornton, S. M., Mecacci, G., & Santoni de Sio, F. (2019). Four perspectives on what matters for the ethics of automated vehicles. In *Road vehicle automation 6 6* (pp. 49–60). Springer.

Lambert, A. (2020). Love's extension: Confucian familial love and the challenge of impartiality. In R. Fedock, M. Kühler, & R. Rosenhagen (Eds.), *Love, justice, and autonomy* (pp. 259–288). Routledge.

Lau, D. C., & (trans.),. (1979). *Confucius: The analects (Lunyu)*. Penguin Books.

Leben, D. (2017). A Rawlsian algorithm for autonomous vehicles. *Ethics and Information Technology, 19*(2), 107–115. https://doi.org/10.1007/s10676-017-9419-3

Lucifora, C., Grasso, G. M., Perconti, P., & Plebe, A. (2020). Moral dilemmas in self-driving cars. *Rivista Internazionale di Filosofia e Psicologia, 11*(2), 238–250.

Millar, J. (2015). Technology as moral proxy: Autonomy and paternalism by design. *IEEE Technology and Society Magazine, 34*(2), 47–55. https://doi.org/10.1109/MTS.2015.2425612

Nagel, T. (1991). *Equality and partiality* (p. 1991). Oxford University Press.

Parekh, D., Poddar, N., Rajpurkar, A., Chahal, M., Kumar, N., Joshi, G. P., & Cho, W. (2022). A review on autonomous vehicles: Progress methods and challenges. *Electronics, 11*(14), 2162. https://doi.org/10.3390/electronics11142162

Powers, T. (2006). Prospects for a Kantian machine. *IEEE Intelligent Systems, 21*, 46–51. https://doi.org/10.1109/MIS.2006.77

Scanlon, T. M. (1978). Rights, goals, and fairness. In S. Hampshire (Ed.), *Public and private morality*. Cambridge University Press.

Scanlon, T. M. (1998). *What we owe to each other*. Belknap/Harvard University Press.

Segun, S. T. (2021). Critically engaging the ethics of AI for a global audience. *Ethics and Information Technology, 23*(2), 99–105. https://doi.org/10.1007/s10676-020-09570-y

Sui, T. (2023). Exploring moral algorithm preferences in autonomous vehicle dilemmas: An empirical study. *Frontiers in Psychology, 14*, 1–12. https://doi.org/10.3389/fpsyg.2023.1229245

Sullins, J. (2021). Artificial phronesis: What it is and what it is not. In E. Ratti & T. A. Stapleford (Eds.), *Science, technology & virtues*. Oxford University Press.

Vallor, S. (2021). Twenty- first- century virtue: Living well with emerging technologies. In E. Ratti & T. A. Stapleford (Eds.), *Science, technology & virtues*. Oxford University Press.

Waldrop, M. M. (2015). Autonomous vehicles: No drivers required. *Nature, 518*(7537), 20–23. https://doi.org/10.1038/518020a

Wallach, W., & Allen, C. (2010). *Moral machines: Teaching robots right from wrong*. Oxford University Press.

Williams, B. (2008). The human prejudice. In *Philosophy as a humanistic discipline* (pp. 135–152). Princeton.

Yao, X. (1995). Jen, love and universality—three arguments concerning Jen in confucianism. *Asian Philosophy, 5*(2), 181–195. https://doi.org/10.1080/09552369508575419

Yao, X. (2000). *An introduction to confucianism*. Cambridge University Press.

Yong, L. (2023). *Moral Partiality*. Routledge.