Computer Fraud and Security
ISSN (online): 1873-7056

Epistemic Side-effect Effect of Moral Decision-making in Autonomous Driving

Xiangyu Chen

College of Social Sciences, Shenzhen University, Shenzhen 518060, Guangdong, China.

Corresponding Email: cxybeing@gmail.com

Abstract: The *Epistemic Side-Effect Effect* (ESEE) influences how people perceive the reasonableness of ethical decisions made by autonomous driving and the extent to which the driving system is informed of adverse outcomes. When autonomous driving faces complex adverse outcomes, it must deal with challenges related to knowledge and responsibility attribution. In order to minimise the impact of ESEE, more advanced probabilistic assessment models of adverse outcomes need to be built into the algorithmic system to overcome the knowledge attribution problem, and transparency requirements and responsibility tracking mechanisms need to be established to ensure that responsibilities are clearly defined after an accident to address the responsibility attribution problem.

Keywords: autonomous driving; moral decision-making; epistemic side-effect effect; knowledge attribution; responsibility attribution

I. Introduction

Along with the rapid development of autonomous driving technology, the challenges of moral decision-making that it faces have gradually emerged, such as making appropriate choices when harm cannot be avoided, the conflict between utilitarianism and deontology, fairness among different groups, ethical diversity and cultural differences, algorithmic transparency, and legal responsibility. These issues are not only technical challenges but also involve multiple dimensions, such as psychology, ethics, law, and society, and require us to rethink the role of humans in moral decision-making for autonomous driving, especially the attribution of knowledge and responsibility for such decision-making processes.

II. ESEE theory

The theoretical origin of ESEE can be traced back to a series of empirical papers by experimental philosopher Joshua Knobe (Knobe, 2003a; Knobe, 2003b; Knobe, 2004), in which Knobe argued for the *Side-Effect Effect* (SEE) in moral judgement through thought experiments. This effect is also known as the *Knobe Effect*. Beebe & Buckwalter, 2010 expanded SEE from moral judgement to epistemology and explored the intrinsic nature of ESEE through a series of moral and immoral experimental situations, i.e., the side-effects of behaviour when assessing knowledge attribution. When assessing knowledge attribution, the side effects of a behaviour can influence their judgement, especially in the case of a harmful side effect, where people are more inclined to believe that the actor knew about the side effect. Beebe & Jensen, 2012 verified the generality of this effect in a cross-cultural experiment and found that the phenomenon was reproduced across cultures and languages. Rakoczy et al., 2015 found that the ESEE effect applied to adults and showed a similar trend in a group of children, suggesting that the phenomenon has a cognitive basis across ages. Why is this such a widespread and robust phenomenon? There is currently some controversy in the academic community. One explanation suggests that ESEE occurs because adverse outcomes are more causally related or probabilistic, i.e., people are more likely to perceive behaviours that lead to significant adverse outcomes as 'known' (Dalbauer & Hergovich, 2013; Paprzycka- Hausman, 2020). Kneer, 2018 highlights that moral judgements play a key role in ESEE, i.e. the moral nature of action directly influences people's knowledge attributions. Maćkiewicz et al., 2024 found through meta-analysis that ESEE is more pronounced when an action relates to a moral code, which is in line with the findings in the business context study by Robinson et al. 2013 in a business context study, where participants were more likely to perceive the decision maker as 'knowing' about the negative ethical consequences of a business decision.

Knowledge Attribution (KA) is a central issue in modern analytic philosophy, which involves inferring whether a person possesses knowledge about a situation in the face of uncertainty. KA is a question of who 'knows' what and an in-depth exploration of how individuals, environments, cognitive processes, and social norms affect knowledge acquisition and attribution. There are two different positions on KA: *contextualism* and *invariantism*. Contextualist epistemology asserts that KA is context-sensitive, i.e., the truth value of 'knowing' itself often depends on the linguistic context in which it is used. For example, the classic 'I know that I have hands' case is a false statement in specific contexts (e.g., in a dream or a philosophical discussion). However, in everyday life, such a statement is a true proposition.

In contrast, epistemic invariance asserts that the truth value of a statement about knowledge does not change depending on the context. In response to the difficulty of attributing knowledge, Schaffer & Knobe, 2012 proposed the notion of *Contrastive Knowledge*, which argues that knowledge is not just simple propositions but a state of contrast relative to other possibilities and that people are assessing whether or not someone 'knows' when they assess whether or not they 'know'. When assessing whether someone 'knows', they compare multiple possible alternative answers.

Despite the universality of ESEE, there is more pronounced variability in its strength, which several factors, such as the experimental design, the subject of knowledge attribution, and the type of norm, may influence. For example, the more complex the situation described in the experiment, or the more pronounced the moral outcome, the stronger the ESEE. This suggests that the details of the experimental design had a significant effect on participants' knowledge attribution decisions. In legal contexts, the tendency to attribute knowledge may also be confounded by different norms. For example, the assessment of criminal behaviour in a court of law may involve whether a defendant 'knows' the consequences of an act, and this judgement may be influenced by moral bias, leading to a subjective assessment of the knowledge level of the perpetrator of the act. In the theory of knowledge, ESEE also challenges the classical theory of knowledge in the context of the *Gettier Problem*. According to the traditional theory of knowledge, knowledge needs to satisfy the three conditions of 'belief, truth, and evidence', but ESEE theories validated through experimental philosophical tools have shown that KA is more likely to be influenced by behavioural outcomes and that there is a more complex relationship between knowledge and action (Turri, 2014), which triggers a debate on the definition of knowledge and the definition of action knowledge.

Overall, ESEE is a complex and interesting cognitive phenomenon that reveals a strong link between KA and action outcomes. The above studies show that adverse outcomes significantly increase the tendency to attribute knowledge and that this phenomenon is robust across cultures and age groups. Although there is still some controversy in the academic community about the causes and influences of ESEE, the comparative analysis of different theoretical explanations and experimental results shows that ESEE is of significant research significance at both the theoretical and applied levels. It is widely used for several problem domains, such as AI moral decision-making, legal responsibility attribution, and emerging technologies ethics.

III. Moral decision-making in autonomous driving

As a new growth area in AI, the rapid development of autonomous driving technology has led to more and more traffic decisions being shifted from humans to intelligent algorithms. The emergence of autonomous driving technology is more than just a technological revolution; it has also triggered a wide-ranging discussion about moral decision-making. How should autonomous driving make ethical decisions in emergencies? How do we ensure that every decision meets ethical standards when faced with an unavoidable accident? This question has sparked a great deal of controversy and research.

i. From the Trolley Problem to ethical framework for risk

The moral decision-making problem of autonomous driving originates from the classic *Trolley Problem*. The moral dilemma shows that people are often influenced by intuition, rationality, emotion, and logic in their moral decisions. The conflict between deontology and consequentialism reveals the complexity of this moral judgement—people are concerned with the conformity of actions to moral rules and the ultimate consequences of actions. While this dichotomous moral choice provides room for theoretical exploration of early moral decision-making on autopilot, real-world ethical dilemmas are far more complex than trolley conundrums.

In recent years, research on moral decision-making in autonomous driving has gradually shifted from the fundamental trolley dilemma to a more practical risk ethics framework (Geisslinger et al., 2021). This framework argues that autonomous driving needs to weigh in binary choices when making decisions and manage the risks of different stakeholders in a complex, dynamic driving environment. The ethics of risk framework is more suitable for solving ethical dilemmas in real-life traffic scenarios because it considers managing multiple risks, not just simple life-and-death choices. Other scholars have chosen a naturalistic path of progression distinct from traditional ethical theories. It is argued that traditional ethical theories, such as utilitarianism, are usually based on algorithmic models to maximise overall welfare. However, this approach often appears rigid and impractical in complex and changing real-world scenarios. The naturalistic approach, on the other hand, focuses on the generalisation of ethical principles from actual behaviours, emphasising the situational nature of ethical judgements and the importance of human intuition, which is more suitable for the dynamic and unpredictable real-world contexts faced by autonomous driving (Arfini et al., 2022).

ii. Individualisation and uniformity

In the autonomous driving moral decision-making field, there is also an apparent theoretical controversy regarding the standard issue of ethics, i.e., whether individualised moral decision-making should be allowed or mandatory societal standards should be established. At the heart of this issue is whether autonomous driving systems should provide individualised moral decision-making options for each user or whether society or government should set a uniform ethical standard for all autonomous driving systems.

Contessa et al., 2017 proposed the concept of *Ethical Knob*, a setting through which users can adjust their ethical preferences, choosing whether to prioritise the protection of themselves or others in emergencies. The knob also adjusts the car's balance between protecting passengers and protecting pedestrians to reflect different ethical principles (e.g. utilitarianism, deontology, etc.) (Evans et al., 2020; Vakili et al., 2024). Building on this, other scholars have suggested that religious beliefs significantly influence people's preferences when using ethical knobs. For example, subjects from specific religious backgrounds were more likely to protect pedestrians outside than passengers inside the vehicle, which aligns with their beliefs' altruism principle. In contrast, subjects from other religious backgrounds may be more inclined to protect the passengers inside the vehicle, believing their safety is their most immediate responsibility (Stephen, 2019). These ways of personalising moral decision-making aim to provide users with greater autonomy and meet different ethical needs. However, this design also poses additional ethical challenges. For example, how can we ensure that such personalised choices do not negatively impact the overall safety of society?

In contrast, another group of researchers advocates that autonomous driving systems' moral decision-making should be guided by uniform societal standards rather than individualised choices. They argue that mandatory uniform ethical standards can ensure that autonomous driving technologies achieve higher levels of fairness and safety in society as a whole (Gogoll & Müller, 2017; Hansson et al., 2021), while the four dimensions of ethicality, safety, transparency, and consistency have to be considered in the process of developing uniform standards of behaviour (Papadimitriou et al., 2022). The assumption behind this view is that personalised choices may lead to social divisions and even serious traffic accidents due to inequality. Therefore, the establishment of uniform and mandatory ethical standards can prevent this from happening. However, it is undeniable that there is currently no framework of ethical standards that can be universally applied to all autonomous driving. Different ethical theories offer different solutions, but they face significant challenges when embedded in autonomous driving decision-making systems. Furthermore, the diversity of cultural, legal and social norms in different regions complicates the development of global standards (Wang et al., 2020).

Transparency and interpretability are key to moral decision-making for autonomous driving, whether using individualised or uniform ethical standards. Choi & Ji, 2015 have shown that public acceptance of autonomous driving systems depends on their ability to understand and trust the system's decision-making process and that transparent algorithms can significantly increase public trust in autonomous driving technologies (Kizilcec, 2016; Dignum, 2019; Krügel & Uhl, 2024). Public trust and acceptance of a system will be significantly increased if the system can explain its rationale for making ethical decisions to the users. Therefore, developing autonomous driving systems with transparent moral decision-making processes is currently a key focus of the field. However, the design of algorithms to achieve interpretability also faces a series of technical challenges. On the one hand, the complexity of an algorithm makes its decision-making process challenging to explain to the average user. On the other hand, oversimplified explanations may not cover the completeness of algorithmic decisions. Therefore, future research also needs to explore how to improve the interpretability of algorithms without sacrificing decision accuracy.

Undeniably, there are still more problems in the current research on moral decision-making in autonomous driving. Firstly, automatic driving systems must make fast decisions in changing driving environments. This requires further exploration of how to optimise moral decision-making in dynamic environments through algorithms, especially how to make fast ethical decisions in unexpected situations. Secondly, most existing studies are still in the theoretical and simulation stages, lacking data support for large-scale practical applications. It should be increased to verify the performance of moral decision-making of autonomous driving systems in real road scenarios and to ensure their wide adaptability in complex social environments. Finally, the development of autonomous driving technology is inevitably accompanied by legal and policy adjustments. This also requires further exploration of how to integrate moral decision-making into existing legal frameworks to ensure that autonomous driving technologies can be implemented by the law while meeting ethical requirements.

iii. Algorithm model

Before determining which algorithm to adopt, it is widely recognised in academia that different ethical frameworks (e.g. utilitarianism, virtue ethics, deontology, etc.) may suggest different decisions for the same scenario in the real world. Therefore, algorithms must incorporate the principles of different ethical frameworks through appropriate weighting methods to make decision-making more flexible, inclusive and rational. Sui, 2023 has shown that hybrid strategies incorporating various ethical principles are the most favoured public acceptance of moral decision-making. In fact, at the technical level, researchers have begun to explore how to embed relevant ethical theories into automated driving algorithms to achieve complex moral decision-making. Various machine learning models have been widely used in moral decision-making in automated driving.

For example, scholars have proposed a *Deep Q-Network* (DQN) algorithmic framework based on *Deep Reinforcement Learning* (DRL) (Cui et al., 2023; Hoel et al., 2023; Tammewar et al., 2023; Wang et al., 2024). DQN is a common algorithm in deep reinforcement learning that combines Q-learning and deep neural networks. It is mainly used to solve problems related to high-dimensional state spaces, such as complex moral decision-making in autonomous driving. The core of Q-learning is the Q-function, which represents the expected cumulative reward that can be obtained by taking action in states. The updated formula of Q-learning is:

$$Q(s,a) = Q(s,a) + \alpha[r + \gamma max Q(s',a') - Q(s,a)]$$

Where α is the learning rate, γ is the discount factor, r is the immediate reward, s' is the next state, and a' is the optimal action in the next state. However, traditional Q-learning cannot effectively store and update the Q-value table when the state space is very large or continuous. This is where deep neural networks are used to approximate the Q-function. In autonomous driving moral decision-making, the DQN uses a deep neural network to parameterise the Q-function：

$$Q(s,a,\theta) \approx Q^{*}(s,a)$$

Where s is the current state (e.g. the state of the autonomous vehicle or the local real-time traffic situation), a is the action taken (e.g. braking, steering, accelerating, emergency avoidance, etc.). θ is the weight parameter of the neural network.

An autonomous driving system’s core task is making real-time decisions, such as accelerating, braking, changing lanes, etc. DQN models the autonomous driving environment as a reinforcement learning problem, where the state can be the current position and speed of the vehicle and the perception of the surrounding environment (such as the position of other vehicles, the state of traffic lights, etc.), while the action is the control action (such as accelerating, braking, steering, etc.) that the vehicle can take. The goal of the DQN is to learn a Q-value function that evaluates the expected reward of each action given the current state to help the vehicle choose the optimal control action at each step. The DQN can gradually optimise its strategy through repeated interactions with the environment, enabling the vehicle to make the best decisions in various complex situations. Despite the variety of algorithmic models in autonomous driving, the core problem remains unsolved: that is, how to transform complex and abstract ethical concepts into concrete quantitative reward functions, especially when multi-dimensional ethical issues are involved, and how to find a suitable balance between different ethical standards still requires further research.

IV. Knowledge and responsibility attribution for adverse outcomes

Currently, ESEE has become a core problem in the field of AI decision-making. Alizadeh Alamdari et al., 2022 have shown that ESEE can negatively affect human-computer interaction by changing human beliefs or expectations about AI through, so the problem of ESEE effects in AI should be solved in the same way as overcoming the physical side effects (Klassen et al., 2023). ESEE in the moral decision-making of autonomous driving is also compelling, especially regarding how the vehicle evaluates and weighs different moral choices and the need to determine whether the system is ‘aware’ of the potential consequences of the specific ‘blame’ for a crash. Some researchers have found that people associate responsibility with KA in situations involving adverse outcomes, and some even overestimate the probability of their occurrence. For example, when an autonomous driving system makes a decision resulting in a pedestrian being injured, the public may be inclined to assume that the system ‘knows’ about the consequence and is morally responsible. This implies that transparency of ethics and predictability of consequences in the design and use of automated driving systems are key factors influencing public trust.

i. Knowledge attribution: probabilistic assessment of adverse outcomes

Adverse outcomes (traffic accidents) that autonomous driving systems can cause are essentially Low-Probability-High-Impact (LPHI) events, which are usually not easy to predict and manage, and traditional risk assessment tools (e.g., probability-impact matrices) are challenging to apply to this type of risk because they often fail to represent the severity of potential consequences and cascade effects adequately (Acebes et al., 2024). In the cascade effect, more advanced modelling techniques are needed to enable the probabilistic assessment of adverse outcomes, clarify the level of knowledge of the intelligent system about the potential consequences, and thus overcome ESEE in autonomous driving systems as much as possible. Krakovna et al., 2020 propose a *Task-Focused Reward* (TFR) based approach. The core idea is to reduce the side effects of the current task by designing appropriate rewards for future tasks so that the AI agent will not only pay attention to the success rate of the current task but will also proactively consider the potential impacts on the future task when performing the current task. In the case of individuals, it is often easy to overestimate or underestimate the probability of a negative outcome based on emotional states, leading to many cognitive biases in the assessment of knowledge, which can lead to systematic errors in assessing the potential risks of different courses of action. The challenge for AI systems lies in identifying and compensating for such human cognitive limitations to prevent ESEE, which requires counteracting these biases by providing accurate probabilistic assessments and facilitating rational decision-making. While relevant studies accurately portray that individuals are biased in their probabilistic assessments of adverse outcomes, they ignore the complexity of collective or organisational decision-making, where individual-level cognitive processes differ significantly from the broader dynamics of group decision-making. For example, simulated environment studies involving reinforcement learning have shown promising results in reducing ESEE, but their applicability in more complex and dynamic real-world environments remains elusive.

In the current development of autonomous driving, the public has become increasingly accepting of the potential adverse outcomes of autonomous driving. However, when designing algorithms for moral decision-making, it is important to consider balancing obstacle avoidance with protecting passengers in the vehicle and ensure that AI systems are transparent and predictable in their behaviour (Baisero & Amato, 2021). After all, in public perception, if a negative outcome (e.g., hitting a pedestrian) occurs, there may be a tendency to assume that the algorithmic designer is aware of this consequence and consequently creates a higher ethical demand on the system. This attribution may affect the acceptance of autonomous driving technologies and the definition of legal responsibilities. To counteract this probabilistic cognitive bias, more transparent decision-making models and visualisation tools for moral trade-offs need to be introduced into autonomous driving systems so that the public can see more clearly how the system makes moral decisions in different contexts, thus understanding the system's decision-making logic and reducing the negative attributional bias due to cognitive misunderstandings.

ii. Responsibility attribution: the legal determination of moral decision-making

Decision-making in autonomous driving is driven by a complex array of algorithms and sensors, challenging traditional models of attributing moral responsibility. For example, if the autonomous driving system fails to avoid all risks in a complex environment and ends up injuring or killing a pedestrian. How should liability be assigned?

Should the algorithm developer, the car owner, or the car itself be held responsible? While human drivers tend to make decisions based on intuition or emotion when faced with ethical dilemmas, autonomous driving is pre-programmed with well-calculated decisions. Therefore, should autonomous driving be held to a higher ethical standard than human drivers (Bigman & Gray, 2020)?

It has been shown that people are more likely to attribute knowledge to actors when it comes to decisions that violate ethical norms. Under the influence of ESEE, people are more inclined to attribute responsibility to the actor. In contrast, in the case of autonomous driving, this 'actor' is no longer transparent, and this uncertainty, in turn, increases the public's scepticism about autonomous driving systems. In the case of autonomous driving systems, where the system represents to some extent the intentions of the programmers and designers and where the ethical justification and informed attribution of such 'intentions' is unclear, the moral responsibility of the designers of the algorithms and the vehicles themselves for these adverse outcomes remains controversial, inducing the phenomenon of *Diffusion of Responsibility* (Wallach et al., 1964; Whyte, 1991), which means that the attribution of responsibility may become more ambiguous in contexts involving automated systems. For example, developing an automated driving system involves multiple stakeholders, including software developers, hardware manufacturers, and car companies, which complicates the attribution of liability after an accident. Inappropriate attribution of responsibility will likely lead to severe psychological consequences, including unfair liability assumption (drivers), breakdown of trust (the public), and increased anxiety (manufacturers and developers) (Liu et al., 2021). This diffusion of responsibility is also bound to exacerbate ESEE further, as the public has difficulty clarifying who is informed of and responsible for adverse outcomes.

Therefore, when designing an autonomous driving system, transparently spelling out its decision-making logic, liability-taking mechanism, and how to minimise the occurrence of negative outcomes is an important strategy to enhance public trust. The allocation of responsibilities should be clarified within a legal framework, for example, by establishing systematic transparency requirements and tracking mechanisms to ensure that the responsibilities of all parties can be clearly defined after an accident. Legal clarity of responsibilities can help alleviate public distrust of autonomous driving systems due to unclear responsibilities.

V. Discussion

The moral decision-making dilemma of autonomous driving is not only a technical challenge but also a profound test of public perception and social systems. This article reveals the following core contradictions that can be resolved by analysing the impact of ESEE on the attribution of knowledge and the determination of responsibility.

First, there is an imbalance between public perception bias and technical interpretive efficacy. ESEE shows that the public tends to attribute adverse outcomes to the informativeness of algorithms rather than the uncertainty of complex environments. This bias is particularly pronounced in accident scenarios – even if intelligent systems make optimal choices based on probabilistic models, the public may still question their moral indifference. Although existing transparency tools can partially alleviate the trust crisis, oversimplified ethical trade-offs may obscure the multi-objective optimisation process in dynamic decision-making (Sahoo & Goswami, 2023; Khan et al., 2024; Weerts et al., 2024; Shafik, 2025). For example, in dense traffic, algorithms must consider the collision probability, traffic regulations, and passengers' emotional stress simultaneously, while the public often only focuses on the traffic results. Therefore, future transparent designs must introduce a multi-dimensional interpretation framework to transform moral decision-making from a 'black box output' to an 'engaging narrative'. Although current individualised solutions such as the 'ethical knob' gives users the right to make moral choices, the potential risk is that it may trigger a *Race to the Bottom*. Based on the basic idea of game theory, if most users choose to maximise their safety, it may lead to an imbalance in the risk distribution of the overall traffic system. Conversely, while mandating a uniform standard can ensure fairness, it may neglect the individuality and rationality of specific scenarios.

Second, there is the problem of the adaptability of moral quantification to dynamic environments. Existing moral algorithms (such as DQN) still rely on manually defined reward functions, but there are fundamental challenges in translating abstract values such as 'freedom', 'fairness' and 'dignity' into mathematical parameters (Green, 2022; Prem, 2023). For example, how can the differences in the definition of 'vulnerable road users' in different cultures be quantified? Another key issue is that existing ethical algorithms are mostly based on Western individualistic values, ignoring the differences in priorities in collectivist cultures. In East Asia, influenced by traditional Confucian culture, protecting dense pedestrian flows may be more morally urgent than the safety of passengers in cars. Of course, simple regional parameter adjustments may lead to fragmentation of algorithms, hindering the global deployment of technology. Solving this contradiction also requires constructing an 'ethics ontology library'—extracting minimally agreed principles through cross-cultural empirical research, allowing regional modules to expand specific rules.

Finally, there is the fragmentation of legal responsibility allocation and the black-box nature of technology. Accidents involving autonomous driving involve multiple parties, such as developers, manufacturers, and software suppliers, and the 'all or nothing' liability model of the traditional legal framework is difficult to apply. What makes the situation more complicated is that the unexplainability of deep learning models may lead to a 'no-blame vacuum' in the event of an accident. For example, when a system misjudges a path due to an adversarial attack, should the responsibility be attributed to the algorithm defect, the sensor supplier, or the cyber attacker? The 'technology-law' interface needs to be re-engineered to address this issue. On the one hand, a 'digital traceability chain' for algorithmic decisions should be established to record evidence of the entire process from training data to real-time inference; on the other hand, a hybrid system of 'algorithmic strict liability + proportionate fault' should be introduced to distinguish the weight of responsibility between core ethical defects and secondary technical faults.

VI. Conclusion

A breakthrough in autonomous driving moral decision-making requires a three-dimensional synergy of technological innovation, institutional restructuring and cultural tolerance. First, a cognitive-friendly algorithm interpretation system needs to be constructed at the technological level to make the moral decision-making process transparent. The system can display core ethical principles, real-time trade-off logic, and data verification information through a hierarchical interpretation architecture to the user. For example, a decision flow chart can be dynamically displayed on the in-vehicle interface so that the public can trace the complete reasoning chain from environmental perception to action execution. At the same time, a flexible ethical framework is developed that allows for personalised choices within safety boundaries while setting a rigid moral bottom line, but extreme preferences must be dynamically constrained through risk prediction models.

Second, the collaborative evolution of law and technology is the key to clarifying the attribution of responsibility. This requires the establishment of a trinity responsibility framework: using blockchain technology to achieve full-cycle traceability of algorithms to ensure that model versions and decision-making logic can be accurately located after an accident; constructing a multi-party insurance pool to cover long-tail risks; and implementing a hierarchical accountability system to clarify the primary responsibility of developers for the ethical framework and the secondary responsibility of vehicle owners for abuse. This system design can solve the dilemma of the failure of traditional laws in the context of autonomous driving and alleviate public anxiety about ambiguous responsibilities.

Finally, global deployment requires that the technology be both culturally compatible and ethically consistent. The system can integrate a minimal ethical consensus verified across cultures at the core layer through modular architecture design and support the loading of regionally customised rules at the extension layer. A dynamic and evolving ethical knowledge base can be gradually formed on this basis. At the same time, a global ethical conflict arbitration platform should be established to coordinate differences in standards in different regions and avoid being trapped in the dilemma of ‘technological colonisation’ or ‘ethical separatism’. This model of ‘universal core-flexible extension’ provides a principled and flexible solution for the global implementation of autonomous driving technology.

In the future, the moral evolution of autonomous driving needs to move towards human-machine symbiosis. On the one hand, a neural symbolic system that incorporates moral intuition should be developed so that AI can calculate the probability of risk and understand abstract values such as ‘dignity of life’. On the other hand, it promotes public participatory design and allows human consensus to shape algorithmic rules through citizen juries or democratic voting mechanisms.

Competing interests

The authors have no relevant financial or non-financial interests to disclose.

Acknowledgement

（1）Shenzhen University High-level University Phase III Construction Project (000001032029)
（2）Shenzhen University Special Funding for Ideological and Political Education (24MSZX09)

References

1. Abel, D., MacGlashan, J., & Littman, M. L. (2016, March). Reinforcement learning as a framework for ethical decision making. In *Workshops at the thirtieth AAAI conference on artificial intelligence*.
2. Acebes, F., González-Varona, J. M., López-Paredes, A., & Pajares, J. (2024). Beyond probability-impact matrices in project risk management: a quantitative methodology for risk prioritisation. *Humanities and Social Sciences Communications*, *11*(1), 1-13.
3. Alizadeh Alamdari, P., Klassen, T. Q., Toro Icarte, R., & McIlraith, S. A. (2022, May). Be considerate: Avoiding negative side effects in reinforcement learning. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems* (pp. 18-26).
4. Arfini, S., Spinelli, D., & Chiffi, D. (2022). Ethics of self-driving cars: a naturalistic approach. *Minds and Machines*, *32*(4), 717-734.
5. Baisero, A., & Amato, C. (2021). Unbiased asymmetric reinforcement learning under partial observability. *arXiv preprint arXiv:2105.11674*.
6. Beebe, J. R., & Buckwalter, W. (2010). The epistemic side-effect effect. *Mind & Language*, *25*(4), 474-498.
7. Beebe, J. R., & Jensen, M. (2012). Surprising connections between knowledge and action: the robustness of the epistemic side-effect effect. *Philosophical Psychology*, *25*(5), 689-715.
8. Bigman, Y. E., & Gray, K. (2020). Life and death decisions of autonomous vehicles. *Nature*, *579*(7797), E1-E2.
9. Choi, J. K., & Ji, Y. G. (2015). Investigating the importance of trust in adopting an autonomous vehicle. *International Journal of Human-Computer Interaction*, *31*(10), 692-702.
10. Contissa, G., Lagioia, F., & Sartor, G. (2017). The Ethical Knob: ethically-customisable automated vehicles and the law. *Artificial Intelligence and Law*, *25*, 365-378.
11. Cui, J., Yuan, L., He, L., Xiao, W., Ran, T., & Zhang, J. (2023). Multi-input autonomous driving based on deep reinforcement learning with double bias experience replay. *IEEE Sensors Journal*, *23*(11), 11253–11261.
12. Dalbauer, N., & Hergovich, A. (2013). Is what is worse more likely?-The probabilistic explanation of the epistemic side-effect effect. *Review of Philosophy and Psychology*, *4*, 639-657.
13. Dignum, V. (2019). *Responsible artificial intelligence: how to develop and use AI responsibly* (Vol. 2156). Cham: Springer.
14. 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.
15. Geisslinger, M., Poszler, F., Betz, J., Lütge, C., & Lienkamp, M. (2021). Autonomous driving ethics: from trolley problem to ethics of risk. *Philosophy & Technology*, *34*(4), 1033-1055.
16. Gogoll, J., & Müller, J. F. (2017). Autonomous cars: in favour of a mandatory ethics setting. *Science and engineering ethics*, *23*, 681-700.
17. Green, B. (2022). Escaping the impossibility of fairness: From formal to substantive algorithmic fairness. *Philosophy & Technology*, *35*(4), 90.
18. Hansson, S. O., Belin, M. Å., & Lundgren, B. (2021). Self-driving vehicles-an ethical overview. *Philosophy*

& Technology, *34*(4), 1383-1408.

19. Hoel, C. J., Wolff, K., & Laine, L. (2023). Ensemble quantile networks: Uncertainty-aware reinforcement learning with applications in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, *24*(6), 6030-6041.
20. Khan, A. M., Tariq, M. A., Rehman, S. K. U., Saeed, T., Alqahtani, F. K., & Sherif, M. (2024). BIM integration with XAI using LIME and MOO for automated green building energy performance analysis. *Energies*, *17*(13), 3295.
21. Kizilcec, R. F. (2016, May). How much information? Effects of transparency on trust in an algorithmic interface. in *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 2390-2395).
22. Klassen, T. Q., Alamdari, P. A., & McIlraith, S. A. (2023, May). Epistemic side effects: an ai safety problem. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems* (pp. 1797 (pp. 1797-1801).
23. Kneer, M. (2018). Perspective and epistemic state ascriptions. *review of Philosophy and Psychology*, *9*(2), 313-341.
24. Knobe, J. (2003a). Intentional action and side effects in ordinary language. *Analysis*, *63*(3), 190–194.
25. Knobe, J. (2003b). Intentional action in folk psychology: an experimental investigation. *Philosophical Psychology*, *16*(2), 309–324.
26. Knobe, J. (2004). Intention, intentional action and moral considerations. *Analysis*, *64*(2), 181–187.
27. Knobe, J., Buckwalter, W., Nichols, S., Robbins, P., Sarkissian, H., & Sommers, T. (2012). Experimental philosophy. *annual review of psychology*, *63*(1), 81-99.
28. Krakovna, V., Orseau, L., Ngo, R., Martic, M., & Legg, S. (2020). Avoiding side effects by considering future tasks. *Advances in Neural Information Processing Systems*, *33*, 19064-19074.
29. Krügel, S., & Uhl, M. (2024). The risk ethics of autonomous vehicles: an empirical approach. *scientific reports*, *14*(1), 960.
30. Liu, P., Du, M., & Li, T. (2021). Psychological consequences of legal responsibility misattribution associated with automated vehicles. *ethics and information technology*, *23*(4), 763-776.
31. Maćkiewicz , B., Kuś, K., Paprzycka-Hausman, K., & Zaręba, M. (2024). Epistemic side-effect effect: a meta-analysis. *Episteme*, *21*(2), 609-643.
32. Papadimitriou, E., Farah, H., van de Kaa, G., De Sio, F. S., Hagenzieker, M., & van Gelder, P. (2022). Towards common ethical and safe 'behaviour' standards for automated vehicles. *Accident Analysis & Prevention*, *174*, 106724.
33. Paprzycka-Hausman, K. (2020). Knowledge of consequences: an explanation of the epistemic side-effect effect. *Synthese*, *197*, 5457-5490.
34. Prem, E. (2023). From ethical AI frameworks to tools: a review of approaches. *AI and Ethics*, *3*(3), 699–716.
35. Rakoczy, H., Behne, T., Clüver, A., Dallmann, S., Weidner, S., & Waldmann, M. R. (2015). The side-effect effect in children is robust and not specific to the moral status of action effects. *ploS one*, *10*(7), e0132933.
36. Robinson, B., Stey, P., & Alfano, M. (2013). Virtue and vice attributions in the business context: an

experimental investigation. *Journal of Business Ethics*, *113*, 649-661.

37. Sahoo, S. K., & Goswami, S. S. (2023). A comprehensive review of multiple criteria decision-making (MCDM) Methods: advancements, applications, and future directions. *Decision Making Advances*, *1*(1), 25–48.
38. Schaffer, J., & Knobe, J. (2012). Contrastive knowledge surveyed. *noûs*, *46*(4), 675-708.
39. Shafik, W. (2025). Machine Learning Techniques for Multicriteria Decision-Making. In *Multi-Criteria Decision-Making and Optimum Design with Machine Learning* (pp. 165–194). CRC Press.
40. Stephen, K. (2019). The Effect of Religiosity on Decision Making in Self-Driving Cars: The Case of " The Ethical Knob".
41. Sui, T. (2023). Exploring moral algorithm preferences in autonomous vehicle dilemmas: an empirical study. *Frontiers in Psychology*, *14*, 1229245.
42. Tammewar, A., Chaudhari, N., Saini, B., Venkatesh, D., Dharahas, G., Vora, D., ... & Alfarhood, S. (2023). Improving the performance of autonomous driving through deep reinforcement learning. *Sustainability*, *15*(18), 13799.
43. Turri, J. (2014). The problem of ESEE knowledge. *ergo*, *1*(4), 101-127.
44. Vakili, E., Amirkhani, A., & Mashadi, B. (2024). DQN-based ethical decision-making for self-driving cars in unavoidable crashes: an applied ethical knob. *Expert Systems with Applications*, *255*, 124569.
45. Wallach, M. A., Kogan, N., & Bem, D. J. (1964). Diffusion of responsibility and level of risk-taking in groups. *The Journal of Abnormal and Social Psychology*, *68*(3), 263.
46. Wang, H., Khajepour, A., Cao, D., & Liu, T. (2020). Ethical decision making in autonomous vehicles: Challenges and research progress. *IEEE Intelligent Transportation Systems Magazine*, *14*(1), 6-17.
47. Wang, Z., Yan, H., Wei, C., Wang, J., Bo, S., & Xiao, M. (2024, August). Research on autonomous driving decision-making strategies based deep reinforcement learning. In *Proceedings of the 2024 4th International Conference on Internet of Things and Machine Learning* (pp. 211-215).
48. Weerts, H., Pfisterer, F., Feurer, M., Eggensperger, K., Bergman, E., Awad, N., ... & Hutter, F. (2024). Can fairness be automated? Guidelines and opportunities for fairness-aware AutoML. *Journal of Artificial Intelligence Research*, *79*, 639-677.
49. Whyte, G. (1991). Diffusion of responsibility: Effects on the escalation tendency. *Journal of Applied Psychology*, *76*(3), 408.