AI and Ethics
https://doi.org/10.1007/s43681-024-00592-6

ORIGINAL RESEARCH

Addressing ethical challenges in automated vehicles: bridging the gap with hybrid AI and augmented utilitarianism

Chloe Gros[1] · Leon Kester[2] · Marieke Martens[2,3] · Peter Werkhoven[1,2]

Received: 11 June 2024 / Accepted: 20 September 2024

Abstract
In the realm of automated vehicles (AVs), the focus is predominantly on the potential of sub-symbolic deep-learning-based artificial intelligence (AI) systems. Our study questions the suitability of this data-driven approach for AVs, particularly in embodying societal values in their behaviour. Through a systematic examination of sub-symbolic and symbolic AI, we identify key issues for AVs, including adaptability, safety, reliability, trust, fairness, transparency, and control. Deep learning systems' lack of adaptability and inherent complexities pose significant safety concerns and hinder meaningful human control. This limitation prevents humans from effectively updating AI decision-making processes to better reflect ethical values. Furthermore, deep learning systems are prone to biases and unfairness, leading to incidents that are difficult to explain and rectify. In contrast, symbolic, model-based approaches offer a structured framework for encoding ethical goals and principles within AV systems, thus enabling meaningful human control. However, they also face challenges, such as inefficiencies in handling large amounts of unstructured data for low-level tasks and maintaining explicit knowledge bases. Therefore, we advocate for hybrid AI, combining symbolic and sub-symbolic models with symbolic goal functions. We propose Augmented Utilitarianism (AU) as an ethical framework for developing these goal functions, aiming to minimise harm by integrating principles from consequentialism, deontology, and virtue ethics, while incorporating the perspective of the experiencer. Our methodology for eliciting moral attributes to construct an explicit ethical goal function engages collective societal values through iterative refinement, contributing to the development of safer, more reliable, and ethically aligned automated driving systems.

Keywords Artificial Intelligence · Hybrid AI · Automated vehicles · Symbolic AI · Ethical AI

1 Introduction

1.1 AVs: an overview

Automated Vehicles (AVs) are a prominent area of AI research nowadays. In today's landscape, AVs are at the forefront of innovation and disruption in the transportation industry. With advancements in artificial intelligence, sensor technology, and connectivity, AVs are rapidly evolving towards greater autonomy and reliability. Potential benefits of large-scale adoption of AVs include improving road safety, alleviating congestion, decreasing pollution, and reducing energy consumption [1]. Major automotive manufacturers, technology companies, and startups are investing heavily in research and development to bring automated vehicles to the market. However, challenges such as regulatory frameworks, safety concerns, and public acceptance need to be addressed alongside technological advancements. Recently, concerns have been raised about robotaxis causing accidents that human drivers would not have caused, such as driving into wet concrete in work zones or not making room for emergency vehicles [2].

There are 5 levels of vehicle automation as defined by SAE J3016 [3]. Level 1 introduces driver assistance systems such as adaptive cruise control, where the vehicle

✉ Chloe Gros
c.n.gros@uu.nl

1 Faculty of Science, Department of Information and Computing Sciences, Utrecht University, Princetonplein 5, Utrecht 3584 CC, The Netherlands

2 TNO Netherlands, Intelligent Autonomous Systems, Postbus 96864, The Hague 2509 JG, The Netherlands

3 Eindhoven University of Technology, Eindhoven, The Netherlands

Published online: 10 October 2024

can control one aspect (e.g. longitudinal control) of driving while the human manages others. Level 2 automation involves partial automation, where advanced driver assistance systems (ADAS) manage both steering and acceleration under specific conditions. However, a human driver remains responsible for the overall driving task, requiring constant vigilance, road monitoring, and typically maintaining hands on the wheel. Level 3 marks a significant technological and human-centred leap, where vehicles can temporarily perform the entire dynamic driving task while being in predefined conditions, defined as the Operational Design Domain. This means that the driver can temporarily be out of the loop and perform non-driving related tasks and only needs to take back control if warned by the vehicle.

Moving to Level 4, vehicles are included that can perform a minimum-risk manoeuvre if a driver does not respond to a take-over request. However, it also includes automated shuttles and robotaxis that are capable of self-driving within limited areas, such as urban environments or predefined locations, and the person inside is not required to have any active role in driving or taking over control. Level 5 vehicles represent the pinnacle of automation, as they suggest being able to handle any situation a human driver could handle in any environment.

Since levels 1 and 2 still require the driver to monitor the driving task and are formally responsible for all parts of the driving task, in this paper we focus particularly on levels 3 to 5, where the automated system can act upon the driving environment without the need for active human input or monitoring while driving. From this level 3 and above, the AV will need to decide how to act upon any situation within the operational design domain.

1.2 The role of AI in AVs

AI technology has emerged as a cornerstone in the development of AVs, revolutionising the way vehicles perceive, interpret, and navigate the world around them. AI is integral to the operation of AVs, contributing to various aspects of their functionality. AI algorithms process data from sensors such as cameras, LiDAR, radar, and GPS to perceive the vehicle's surroundings and to create a world model. Such a world model allows a cognitive system to better predict future sensory observations and optimise its actions based on those predictions. It encompasses not only the dynamics of the external environment but also bodily dynamics and social interactions [4].

Based on this world model, AI algorithms enable AVs to make real-time decisions. AVs analyse sensor data to predict the behaviour of surrounding objects and choose appropriate actions to determine the optimal paths and trajectories for AVs to reach their destinations safely and efficiently. This is referred to as path planning.

In addition, AI algorithms control the vehicle's actuators, such as steering, throttle, and brakes, to execute the planned trajectory. These algorithms strive to achieve smooth and precise control of the vehicle's movements, adjusting in real-time to changes in the environment, though this is not always guaranteed. AI is also used for simultaneous localisation and mapping (SLAM), allowing AVs to accurately determine their position and create maps of their surroundings in real-time. SLAM algorithms fuse data from sensors and use probabilistic techniques to estimate the vehicle's pose and map the environment.

Finally, AI algorithms can assist AVs in decision-making during unavoidable collisions by evaluating potential damages across various scenarios. For example, in their 2021 article, Perumal et al. highlight the critical role Crash Avoidance and Overtaking Advice (CAOA) systems within Advanced Driver Assistance Systems (ADAS) systems play in improving road safety by tackling challenges such as obstacle avoidance, overtaking, and lane changes [5]. Additionally, in their 2016 paper, Wiseman and Grinberg describe a real-time assessment method using computational geometry, which involves constructing intelligent simulation models of vehicles as simple polygons [6]. Finally, in 2022, Li et al. introduced an innovative integrated approach to enhance collision avoidance during emergencies [7]. Their strategy combines steering and braking mechanisms through a two-layer framework: an upper-level decision-making layer and a lower-level control layer.

In summary, AI technology is central to the evolution of automated vehicles, enabling them to perceive their surroundings, make real-time decisions, and navigate safely. By processing sensor data and creating detailed world models, AI facilitates precise path planning, control of vehicle actuators, and accurate localization and mapping. Recent advancements underscore AI's critical role in enhancing safety and optimizing collision avoidance strategies, reflecting its transformative impact on automated driving.

2 Deep-learning-based AI systems

Typically, common-day AVs use deep learning based-AIs, meaning that their world model is primarily based on sub-symbolic neural networks. Deep learning techniques involve training artificial neural networks with large amounts of data to recognise patterns and make decisions. This approach allows AVs to learn complex patterns and behaviours from sensor data without explicitly programmed rules. However, while deep learning excels in tasks like object recognition and classification, it may struggle with understanding

abstract concepts or adapting to novel situations. Deep learning can be described as a data-driven approach, also called a bottom-up approach or sub-symbolic approach. This means that the model does not develop explicit symbolic reasoning or understand the knowledge behind specific decisions, as it learns through examples without being given an explicit symbolic world model to optimise.

It is important to note that while sub-symbolic AI systems primarily learn through examples, they still require symbolic goal functions. A "goal function" is a comprehensive term encompassing all types of loss functions, objective functions, multi-objective functions, or utility functions used in various fields, especially in machine learning and artificial intelligence. It serves as a mathematical criterion that quantifies the performance or desirability of different outcomes or actions. For instance, in supervised learning, the loss function measures the difference between the predicted values and the actual values, guiding the optimisation process to minimise this difference and improve the model's accuracy. In multi-objective optimisation, the goal function involves balancing several objectives, often competing, to find an optimal trade-off solution. Utility functions assign values to different outcomes to reflect their relative preference or worth, driving the decision-making process towards maximising expected utility [8]. Thus, the goal function provides a formalised way to direct the learning or optimisation process toward achieving desired results, incorporating various constraints and priorities inherent to the problem at hand.

In deep-learning-based AI systems, goal functions are used beforehand to select relevant training data. Imagine an AV system designed to prioritise pedestrian safety. Goal functions based on symbolic rules can help choose training data that includes diverse pedestrian scenarios, such as crosswalks, school zones, and crowded city streets. This ensures the system learns from a wide range of critical situations. Additionally, symbolic evaluation criteria allow for systematic performance assessment. For example, after training, the AV's behaviour can be evaluated against predefined ethical guidelines, such as stopping for pedestrians at crosswalks or maintaining safe distances from other vehicles. If the AV fails to stop at a crosswalk during a test, this symbolic evaluation helps identify specific issues in the algorithm.

2.1 Case study: Wayve's gaia

Wayve is a software development company that uses AI to pioneer a next-generation approach to self-driving [9]. Wayve's autonomous driving technology, known as Gaia, operates on an approach called end-to-end learning. Unlike traditional methods that rely on handcrafted rules and mapping techniques, Gaia learns to drive directly from raw sensor data, such as camera images, using deep learning algorithms. This approach enables Gaia to perceive and understand the environment in a manner similar to how humans learn to drive, without extensive manual intervention or hand-crafted rules.

The GAIA-1 model utilises specialised encoders for various input modalities, including video, text, and action, to create a shared representation. These encoders project input data into a coherent timeline, ensuring alignment across different modalities. The core component of the model is the sub-symbolic world model, an autoregressive transformer trained to predict the next set of image tokens in a sequence by considering past image tokens, textual context, and action-based guidance. With 6.5 billion parameters, the world model generates visually coherent images aligned with textual and action-based input. Subsequently, a video decoder, employing a video diffusion model with 2.6 billion parameters, translates predicted image tokens into pixel space, enhancing the semantic meaning, visual accuracy, and temporal consistency of generated videos. GAIA-1 boasts over 9 billion parameters and was trained for 15 days on 64 NVIDIA A100s, utilising a dataset of 4,700 h of proprietary driving data collected in London, UK, between 2019 and 2023.

Wayve's Gaia represents a state-of-the-art example of modern automated vehicles that are solely based on deep learning and sub-symbolic AI. The formulation of the world modelling task in GAIA-1 is streamlined to focus on predicting the next token. This method underscores the reliance on sub-symbolic processes, highlighting the potential and current applications of deep learning in creating highly advanced automated driving systems.

Although models like Wayve's Gaia are at the forefront of automated driving technology, their reliance on deep learning proves inadequate for addressing the ethical challenges inherent in automated driving because of a lack of symbolic reasoning, explainability and therefore meaningful human control. We will elaborate on these shortcomings in the following section.

2.2 Ethical challenges

The advent of deep learning AI systems, characterised by approaches like end-to-end learning and deep reinforcement learning, has heralded remarkable advancements in various domains, including automated driving. However, alongside their potential benefits, these methods introduce profound ethical challenges. In most cases, challenges of deep learning-based AI systems highlighted in literature include technical challenges, but more attention is now also given to ethical issues such as unbiased data availability, limited

transfer between tasks, brittleness, but also explainability, trustworthiness and security [10–12].

Further, on 8 April 2019, the High-Level Expert Group on AI presented Ethics Guidelines for Trustworthy Artificial Intelligence [13]. This document highlights 4 ethical principles that serve as the foundations of Trustworthy AI: respect for human autonomy, prevention of harm, fairness, and explicability. To address these concerns comprehensively, we adopt a systematic approach that includes the following six critical topics: adaptability, reliability, fairness, explainability, trustworthiness, and security. These topics were chosen because they encompass the primary ethical and technical challenges unique to deep learning AI systems in automated driving. By addressing these six areas, we aim to provide a holistic evaluation of the ethical implications of deep learning AI systems in automated driving, ensuring alignment with the EU's guidelines for trustworthy AI.

2.2.1 Adaptability

Deep learning's demand for vast amounts of data starkly contrasts with human learning capabilities, which excel at abstracting relationships from limited examples. While humans can effortlessly grasp abstract concepts with minimal examples, deep learning systems require extensive datasets to achieve similar feats. Despite advancements, deep learning struggles with abstract reasoning and adaptability to new situations, relying heavily on extensive training data and appropriate examples. This leads to several challenges. If incorrect data are present in the training set, simply removing them can be difficult without potentially disrupting the system's learned behaviour, as deep learning models integrate information from all given data. Identifying and extracting erroneous data without losing valuable contextual information is a significant challenge, underscoring the limitations of deep learning in achieving robust and reliable performance across varied and unforeseen scenarios.

Moreover, deep learning's reliance on numerous hidden layers does not inherently imply conceptual depth; the representations acquired often lack nuanced understanding. Transfer tests reveal the superficial nature of deep learning's solutions, with systems failing when encountering minor perturbations outside their training set. Additionally, the explosion of state-action space in complex environments, such as the number of possible scenarios in automated driving situations, exacerbates these issues, making it difficult for deep learning models to generalise effectively across different situations. This inherent rigidity and dependency on specific training data highlight the need for more adaptable and flexible AI systems to ensure robust performance in dynamic and unpredictable real-world scenarios.

This stems from the inability of AI systems to explicitly update their world models due to a lack of understanding. Unlike humans, who can adapt their mental models of the world based on new information and experiences, deep learning systems operate within predefined frameworks and struggle to incorporate novel or unexpected situations into their existing models. This limitation comes from the inherent inability of deep learning algorithms to understand and create explanatory knowledge, which is needed in order to update their world model. As Roli et al. show in their 2022 paper, true general intelligence entails situational reasoning, perspective-taking, goal selection, and handling ambiguous information, all of which rely on identifying and exploiting new affordances—opportunities or impediments for goal attainment. However, as it is not possible to predefine a complete list of such uses, they cannot be treated purely algorithmically. Unlike organisms, deep learning agents do not possess the capacity to leverage new affordances [14]. As a result, AI systems may struggle to update their world models in real-time, leading to inaccuracies or errors in their perception and decision-making.

The lack of true understanding of deep learning AI systems poses significant challenges for AVs in real-world scenarios. Without an explicit world model, which defines the functional relationship between objects, the AI model will not be able to transfer knowledge from one situation to the next and create an efficient self-improving feedback loop. Hence, such an AI system will not be able to generalise to infinite situations. For instance, if an AV encounters unanticipated road closures or obstacles, its AI may struggle to adjust its world model to navigate safely. Additionally, changing traffic patterns, new road layouts, and unpredictable pedestrian or cyclist behaviour can further challenge AVs relying on implicit world models. Furthermore, when an incident occurs, the lack of an explicit understanding and transparent decision-making process means it cannot be easily explained in court. This lack of explainability additionally hampers our ability to learn from the incident and implement improvements. The inability to provide clear, understandable reasons for AI decisions undermines trust and accountability, crucial elements for the widespread acceptance and ethical deployment of AV technology.

2.2.2 Transparency

It has been contended that the opacity inherent in deep learning systems is not necessarily problematic [10]. For instance, in advisory AI systems, absolute transparency may not be imperative if these models show a high success rate. However, these systems involve the supervision of a human before any decision-making takes place. For example, consider the case of AI systems used in advanced

driver-assistance systems in vehicles. For instance, a lane-keeping assist feature might rely on sophisticated algorithms to help drivers stay within their lane. Absolute transparency into the exact workings of this AI might not be crucial, as long as the system performs reliably and the driver remains in control, with the AI providing assistance when needed. In contrast, AVs operate with a higher level of automation where real-time decision-making is critical. These vehicles continuously process a vast amount of sensory data and make instantaneous decisions about navigation, obstacle avoidance, and traffic interactions without human intervention. In this setting, transparency in the AI's decision-making process is essential to ensure that the system can be trusted to handle complex and dynamic driving conditions safely.

Moreover, the acceptability of this lack of transparency hinges upon the successful resolution of other technical challenges inherent in deep learning methodologies. In situations where data availability is limited or where the transferability of knowledge from one task to another is critical—particularly in high-risk scenarios—an opaque system could yield inconsistencies or erroneous decision-making, potentially resulting in catastrophic consequences. Furthermore, recent accidents involving robotaxis and even partially automated vehicles have raised significant public acceptance issues, highlighting the importance of transparent decision-making processes in automated systems to build trust and confidence among users and stakeholders [15].

Additionally, the inability to remove bad examples from training data in deep learning algorithms poses a significant challenge. Even if AVs are trained on diverse datasets, including scenarios with adverse weather or unusual road conditions, the presence of erroneous or misleading examples in the training data could lead to suboptimal decision-making in real-world situations. As a result, AVs may struggle to accurately interpret and respond to novel or ambiguous situations, highlighting the inherent limitations of relying solely on deep learning for automated driving.

The lack of transparency in deep learning AI systems presents significant challenges for humans in debugging, interpreting, controlling, and reasoning about them [10]. Without transparency, these systems operate as black boxes, concealing the underlying mechanisms driving their decisions and behaviours. This opacity hinders human efforts to debug errors or anomalies within the system, as understanding the root cause becomes elusive. As a result, interpreting the outputs or predictions of these systems becomes daunting, leaving humans unable to discern how or why certain decisions were made.

2.2.3 Control

Meaningful human control refers to the ability of humans to understand, influence, and ultimately take responsibility for the decisions and actions of automated systems, particularly in contexts where these systems have a significant impact on human lives or society as a whole. In the realm of AI, meaningful human control requires the ability to influence all relevant aspects of an automated system, access to information, and meeting strict control conditions, including knowledge and capacity [16]. This entails ensuring that humans have the capability to oversee and intervene in AI systems' decision-making processes, understand the rationale behind these decisions, and hold the companies responsible for the actions of their AI systems.

Operational control is generally interpreted in terms of having a driver control the vehicle. However, the concept of "human before the loop" goes beyond this traditional view. It emphasises meaningful human control through humans being involved in the design, development, and governance stages of AI systems. This ensures that ethical principles and societal values are embedded from the outset, rather than just having humans take control in real-time situations. Meaningful human control is essential for ensuring that AI systems serve human interests, align with societal values, and operate in a manner that is transparent, accountable, and trustworthy. It is a fundamental principle in the responsible development, deployment, and governance of AI technologies across various domains, including automated vehicles, healthcare, finance, and defence.

Relying solely on sub-symbolic implicit world models, as observed in certain automated driving systems like Wayve's Gaia, may impose limitations on meaningful control, particularly due to the inadequate specification of both world models and ethical goal functions. Without explicit representation of ethical considerations and higher-level goals, sub-symbolic models may make decisions that implicitly prioritise efficiency or convenience over safety, leading to potentially unsafe or unpredictable behaviour.

2.2.4 Security

In addition to the challenges posed by the inability of deep learning AI systems to have an explicit world model, there is also the concern of how this limitation could facilitate adversarial attacks. Adversarial attacks involve intentionally manipulating input data to deceive AI systems, causing them to make incorrect predictions or decisions. Since AI systems lack a comprehensive understanding of the world and rely on predefined frameworks, they are susceptible to adversarial attacks that exploit vulnerabilities in their perception and decision-making processes.

For example, in the context of AVs, adversaries could manipulate sensor inputs, such as traffic signs or road markings, in subtle ways that are imperceptible to humans but can lead to misinterpretations by AI algorithms. Without the ability to dynamically update their world models to adapt to these adversarial inputs, AI-powered AVs may be vulnerable to safety risks and security threats. As AI technology continues to advance, addressing these vulnerabilities and developing robust defence mechanisms against adversarial attacks will be crucial for ensuring the safety and reliability of AI-powered systems in real-world environments [10, 17].

2.2.5 Data bias and fairness

Deep learning models rely heavily on training data, and biases present in the data can propagate to the model's decisions. This raises concerns about fairness and equity, particularly if the automated system exhibits biased behaviour towards certain demographic groups or communities.

Bias is defined as a systematic error in decision-making processes that results in unfair outcomes [18]. AI systems can learn and replicate patterns of bias present in the data used to train them, resulting in unfair or discriminatory outcomes. Bias and fairness in AI are closely related to user and societal trust: When AI systems exhibit biases or unfairness due to biased training data or flawed algorithms, it can erode public trust in these systems. Users may lose confidence in AI-powered applications, questioning the reliability and impartiality of the decisions made by these systems.

Numerous instances of biased facial recognition technology have been observed in law enforcement agencies. For instance, research conducted by the National Institute of Standards and Technology (NIST) revealed that this technology exhibited significantly lower accuracy rates for individuals with darker skin tones, resulting in elevated false positive rates [19]. This bias can lead to significant outcomes, including unjust arrests or convictions, and an increased risk of collisions with AVs. We can easily see how this type of bias could translate into AVs in a deep-learning-based approach: a biased AI system trained mostly on a white sample of the population may be less able to identify individuals with darker skin tones, which may lead to an overrepresentation of injuries in this group. As a result, current disparities may be exacerbated, leading to increased discrimination against marginalised communities and restricting their access to critical services. The lack of transparency of the black box AI system increases the difficulty of identifying these biases.

Relying solely on deep learning AI in AVs can have significant consequences, particularly concerning biases present in the training data. Without appropriate measures, these biases may propagate into the AV's decision-making, leading to unfair outcomes, such as disproportionately affecting certain demographic groups, and potentially leading to wrongful arrests, injuries, or other forms of discrimination. Additionally, the lack of transparency in AI systems complicates the identification and mitigation of biases, highlighting the urgent need for proactive measures to ensure fairness and equity in AV technology.

2.2.6 Trust

One of the main arguments for the development of automated vehicles is that they would be substantially safer than conventional vehicles. A widely circulated statistic states that 94% of car crashes are due to human error [20]. Replacing humans with AVs would then lead to significantly fewer road injuries and fatalities. Some scholars have even posited in the literature that the safety advancements of AVs could lead to an obligation to transition from conventional transportation methods to AVs: if a new technology is introduced which is safer than previously existing alternatives, then this creates a duty to switch over to the safer alternative [21, 22].

However, the current public opinion seems reluctant to the adoption of AV technology. A recent survey commissioned by Forbes Advisor shows that 93% of Americans have concerns about some aspects of self-driving cars and that 62% of consumers have lost confidence in Tesla due to recent safety and technology recalls [23]. Additionally, 61% of respondents would not trust a self-driving car with their loved ones or children. The main cause of concern when it comes to self-driving cars is safety. Improving AV safety is therefore crucial for AV acceptance and to ensure that the positive benefits of AVs are met.

Trust in AI is crucial in overcoming this barrier, and transparency is key to building that trust. As Eschenbach describes, "A trusts B to do X only if A judges B to be trustworthy where trustworthy means that A has good reason to believe that B is competent in doing X and that B would act on A's behalf." We can draw a clear parallel with AVs, where A would be a human agent, trusting an AV (B), to drive in on their behalf (X) [24]. However, in a trust relationship, *how* the action X (driving) is performed by B (the AV), is as important as performing the action itself. If the AV compromises one's safety or that of others while driving on one's behalf, one may feel that their trust has been betrayed in the same manner as one would feel if the AV did not drive at all. Taking action on behalf of someone else implies aligning with, or at least not contradicting, their values and commitments, thereby offering justification for one's actions. Trust in another's ability to act on one's behalf is built upon understanding and insight into their motives and methods. Transparency in their actions or intentions is essential for evaluating their trustworthiness.

In a deep learning approach, the AI system is opaque due to the absence of any mechanisms to reproduce or explain decision-making processes: inputs go into the system, and outputs come out, but the process by which the inputs are transformed into outputs is not clear. This poses significant challenges when considering questions of trust. Such an AI system cannot be guaranteed to be reliable and free from bias. The challenges posed by the inability to comprehend or clarify errors, as well as the process leading to significant outcomes, greatly diminish confidence, and therefore trust, in these systems.

In conclusion, deep learning approaches, such as the data-driven bottom-up methods prevalent in current automated vehicle technologies like Wayve's Gaia, face inherent limitations in addressing several ethical challenges. These approaches lack explicit representation of ethical considerations and higher-level goals, relying solely on deep learning without incorporating domain knowledge or symbolic reasoning. Without an explicit ethical goal function and a comprehensive understanding of the underlying decision-making processes, deep learning AI systems may implicitly prioritise efficiency or convenience over safety, leading to potentially unsafe or unpredictable behaviour without humans even realising this. Additionally, the opacity of these systems makes it difficult to reproduce or explain their decision-making processes, diminishing confidence and trust in their reliability. Therefore, while deep learning approaches have demonstrated advancements in AI capabilities, they fall short in providing the meaningful human control and oversight necessary for ensuring ethical and responsible AI deployment in automated vehicles.

3 Model-based AI systems

Alternatively, a top-down or model-based approach offers a promising avenue for addressing some of the ethical challenges inherent in current bottom-up AI methodologies. Unlike data-driven approaches relying solely on deep learning, model-based approaches leverage symbolic AI techniques, reasoning with models, and incorporating domain knowledge to explicitly encode world models as well as ethical goals and principles within AI systems [25].

In model-based systems, we can use orthogonality-based disentanglement by separating the AI system's problem-solving capabilities from its ethical considerations [26]. This is achieved by defining two distinct axes: one for the technical performance and problem-solving abilities of the AI, and another for its adherence to ethical values as encoded by human stakeholders. The first axe includes developing accurate world models representing the vehicle's environment, enabling precise object detection, localisation, and mapping through sensor fusion. These world models can be developed without relying on deep learning by integrating various techniques, such as knowledge graphs that integrate prior knowledge (geometry, properties and rules) and sensor data to represent the AV-environment and symbolic reasoning and planning algorithms to interpret knowledge graphs and predict the environment. Sensor fusion combines data from multiple sensors like cameras and LiDAR, while traditional algorithms and symbolic AI interpret this data to map the environment. Probabilistic models and simulations further enhance these representations, providing transparency and predictability. Additionally, the AI's problem-solving capabilities extend to real-time decision-making processes, such as adapting to dynamic traffic situations, predicting the actions of other road users, and handling unexpected events.

The second axe involves explicitly encoding ethical goals within a dedicated ethical goal function. This function guides the AI's behaviour, ensuring that its actions align with predefined ethical principles. The separation allows for clear responsibility delineation among different stakeholders, such as developers focusing on the AI's technical proficiency and regulators overseeing its ethical compliance. This framework enhances transparency and accountability, facilitating more robust oversight and control over the AI's operations, and ensuring that ethical considerations are consistently prioritised in decision-making processes.

By explicitly representing the underlying mechanisms of the environment, such as the properties of objects and the relationship between them, and the dynamics of AV operation, model-based approaches enable more transparent and interpretable decision-making processes. This transparency enhances trust and confidence in AV systems by allowing human operators to understand and verify the rationale behind the system's actions. Furthermore, model-based approaches facilitate meaningful human control by providing a structured framework for encoding ethical goals and principles within AV systems. During training and even after deployment, humans can supervise and intervene in the system's actions more effectively, ensuring alignment with a priori defined societal values and preferences.

Additionally, model-based approaches enable real-time updates and adaptation of the AV's world model in response to changing environmental conditions or unforeseen events, enhancing the system's robustness and reliability. Model-based approaches offer greater flexibility and generalisation capabilities compared to purely deep learning methods. By encoding domain knowledge and physical laws into the model, AV systems can generalise across diverse scenarios and extrapolate their behaviour to novel situations not encountered during training. This adaptability is crucial for navigating complex and dynamic environments, where unforeseen challenges and uncertainties may arise.

One example of model-based approach is described in a recent paper from van der Ploeg et al. (2023), introducing an innovative trajectory planning method. By incorporating a novel application of a knowledge graph, which utilises a traffic-oriented ontology to reason about the risk of objects and infrastructural elements, the trajectory generator formulates adaptive trajectories validated through simulation. This method formalises the role of contextual information in motion planning, combining model-based predictive planning with a knowledge graph, and demonstrates robustness and real-time applicability through extensive simulation testing across four use cases with 309 variations [27].

However, despite its strengths in adaptability, explainability, reasoning, and knowledge representation, model-based or symbolic AI systems have notable limitations, particularly in handling low-level tasks like image classification. In contrast, deep learning-based AI excels in such tasks by leveraging large datasets to automatically learn and generalise from patterns without requiring predefined models. This data-driven approach allows deep learning models to achieve superior performance in such tasks, handling nuances and variations that model-based systems struggle with [25].

Moreover, the hand-coding of rules and knowledge creates a significant Knowledge Acquisition Bottleneck, requiring extensive human involvement and leading to high costs and time inefficiencies [28]. Acquiring explicit knowledge bases, typically from experts, is error-prone and expensive, limiting the scope of such systems. Additionally, the maintenance of rule bases poses challenges, as it necessitates complex verification and validation processes. Logic-based reasoning methods are subject to combinatorial explosions that limit both the number of axioms and the depth of reasoning that is possible, further constraining the efficiency and scalability of symbolic AI approaches. While model-based AI offers robustness and interpretability, its limitations highlight the need for complementary approaches to effectively address the complexities of modern AI applications [11].

4 Hybrid AI

The combination of both implicit and explicit models is called hybrid AI [11]. By integrating the strengths of symbolic methods with the adaptability and learning capabilities of sub-symbolic techniques, hybrid AI offers a promising solution to overcome the challenges faced by purely model-based or deep-learning-based AI systems. In this hybrid framework, the symbolic component oversees the deep learning part, providing oversight and guidance based on predetermined models. Unlike deep learning algorithms, which operate primarily based on learned patterns from data, the symbolic reasoning process allows for self-assessment and self-management. This capacity for reasoning enables the system to generalise knowledge and make decisions based on logical principles, enhancing its adaptability and robustness in various situations.

In hybrid AI-based AVs, the symbolic layer serves as a crucial bridge between technical functionalities and ethical considerations, guided by the principles of orthogonality-based disentanglement (see Fig. 1) [26]. It integrates technical elements such as sophisticated path-planning algorithms, leveraging symbolic reasoning to navigate complex environments efficiently while considering high-level goals and constraints. Simultaneously, the symbolic layer incorporates an ethical component, manifested in the form of an ethical goal function derived from the orthogonality-based disentanglement approach. This function quantitatively encodes human values and preferences, guiding AVs to prioritise actions aligned with ethical principles, such as pedestrian safety or environmental sustainability. Through the orthogonality-based disentanglement framework, the symbolic layer ensures a transparent division of responsibilities, enabling AVs to operate not only effectively but also responsibly and ethically in dynamic real-world scenarios.

Finally, the sub-symbolic layer primarily handles tasks requiring pattern recognition capabilities. This layer leverages deep learning techniques to process raw sensor data, such as images from cameras, radar signals, and LiDAR data, to perform essential functions. It identifies and classifies objects in the environment, integrates data from multiple sensors for accurate perception, and determines the vehicle's precise location using GPS and visual odometry. Additionally, it predicts the future movements of dynamic objects, such as other vehicles and pedestrians, to enhance safety and decision-making. By efficiently handling these data-intensive tasks, the sub-symbolic layer is crucial for the AV's overall performance and adaptability.

5 Moral decision-making for AVs

Efforts to address the moral implications of new technology require shifting focus towards considering moral issues during the design process, rather than as an afterthought. While laws, ethical codes, and theories offer frameworks for ethical assessment, they often lag behind technological advancements and primarily serve as tools for reviewing moral impact rather than proactively guiding design. Despite their utility, these top-down approaches are insufficient to keep pace with the rapid innovation in technology, necessitating a more proactive and integrated approach to moral design [29].

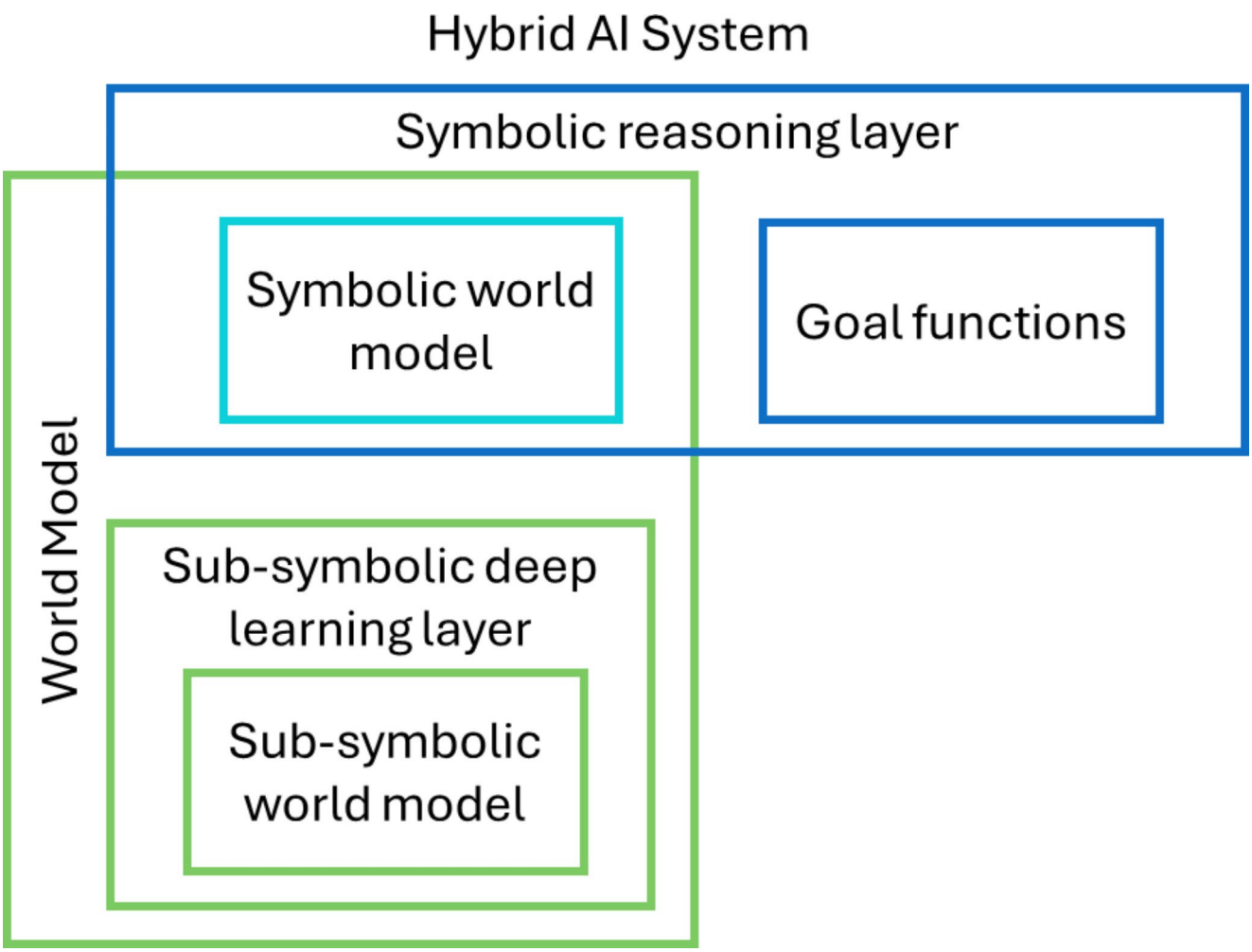

Fig. 1 Illustration of a hybrid AI system

Ethics must be integrated into the design process from the outset, a concept known as "ethics at the front door" [27], and continuously embedded throughout the development stages, referred to as "ethics by design." This combined approach ensures that ethical considerations shape the entire design process and the moral impact of the resulting technology from the very beginning and throughout its lifecycle. Recognising that technology's ethics are multifaceted—with influences from both the designer's values and the inherent moral implications of the artefact—highlights the importance of aligning values from the initial design stages and maintaining this alignment as the technology evolves. This comprehensive ethical integration helps ensure the development of morally sound technology that remains stable and ethically aligned throughout its lifecycle.

Crucially, in the context of high-risk AI systems where risk is defined as the probability of harm [30], the symbolic layer must explicitly incorporate representations of potential harm, in the shape of an ethical goal function. By integrating explicit representations of harm into the objective functions, the AI system can prioritise minimising harm while maximising the defined objectives. This approach relies on the AI system's utilisation of an implicit or explicit world model, which enables it to understand and navigate its environment while adhering to the established objective function.

Explicitly representing harm within AI systems necessitates a clear definition of what constitutes harm. While physical harm is a fundamental aspect, a comprehensive harm model must also encompass moral harm. Beyond tangible or physical damage, moral harm considers the ethical implications of AI decisions and actions, particularly in scenarios where human welfare, rights, or societal values are at stake. This dimension of harm extends beyond mere physical consequences to encompass the broader impact on individuals, communities, and society as a whole. By incorporating moral harm into the harm model, AI systems can more effectively evaluate the consequences of their actions and make decisions that align with ethical principles and societal norms. This holistic approach to harm

representation ensures that AI systems consider not only the immediate physical risks but also the broader ethical implications of their behaviour.

However, modelling moral harm within AI systems presents a significant challenge due to the inherent subjectivity and diversity of ethical theories. Different ethical frameworks, such as consequentialism, deontology, and virtue ethics, may yield divergent perspectives on what constitutes moral harm and how it should be prioritised. For example, consequentialism, specifically utilitarianism, prioritises maximising overall happiness or utility, often leading to decisions that may sacrifice the well-being of a few individuals for the greater good of the majority. In contrast, deontological ethics emphasises adherence to moral rules or principles, regardless of the consequences, which may lead to different judgments in morally complex situations. Similarly, virtue ethics focuses on the character traits or virtues of individuals, which may lead to judgments based on the intentions or motivations behind actions rather than their outcomes. By definition, virtue ethics cannot be applied to sub-symbolic AI.

The inherent variability among ethical theories poses a significant obstacle to developing a unified model of moral harm within AI systems, as the choice of ethical framework can profoundly influence the system's decision-making processes and outcomes.

5.1 Augmented utilitarianism

In response to these ethical challenges, a framework known as Augmented Utilitarianism (AU) [31] has emerged as a promising way forward. Unlike traditional utilitarian approaches that focus solely on maximising utility or outcomes, AU adopts a more nuanced perspective by prioritising harm minimisation while adhering to predefined ethical principles. This framework emphasises principles over specified outcomes and incorporates attributes and weights defined by society, rendering it a non-normative framework shaped by societal values. Grounded in moral psychology, cognitive neuroscience, and philosophy, AU aims to capture the diversity of human moral reasoning and ethical perspectives.

AU draws upon the theory of dyadic harm, as elucidated by Gray and Schein [32], to provide a comprehensive understanding of harm in ethical decision-making. Dyadic harm considers not only the consequences of actions, akin to consequentialism, but also factors such as the intentional agent, the action itself, and the perceptions of the observer (see Table 1). By integrating these dimensions of harm, AU offers a holistic approach to ethical decision-making that transcends simplistic utilitarian or deontological frameworks.

Table 1 Multiple ethical frameworks included in augmented utilitarianism [33]

Ethics framework/focus	Agent	Action	Outcome	Experiencer
Virtue ethics	x			
Deontological ethics		x		
Consequentialist ethics			x	
AU	x	x	x	x

Central to AU is the construction of an ethical goal function that guides the decision-making process of AV systems. This goal function is designed to be transparent, explainable, and grounded in societal values predefined by humans, often referred to as the "human before the loop" perspective. By explicitly encoding ethical principles into the decision-making system, AU enables AVs to navigate complex moral dilemmas and prioritise actions that align with societal norms and preferences even under new situations that were not encountered before or are not part of the training set. This ensures that AVs operate in a manner consistent with previously defined and explicit human values and ethical standards, fostering trust, accountability, and societal acceptance of automated technologies. Crucially, the ethical goal function derived from AU reflects societal values, ensuring that AVs prioritise actions that align with human preferences and ethical standards. By incorporating these values into the decision-making process, AVs become more transparent and accountable, as their actions can be traced back to predefined societal norms. Moreover, AU's emphasis on explainability allows for a clearer understanding of how decisions are made, reducing the potential for bias and promoting fairness in AV operations and the potential to learn from mistakes.

In essence, AU represents a significant advancement in ethical AI design for AVs, offering a principled and transparent framework for addressing the ethical challenges inherent in automated systems. By combining moral philosophy with insights from psychology and neuroscience, AU provides a robust foundation for developing AI systems that not only maximise utility but also uphold fundamental ethical principles and values.

5.2 Moral value elicitation

Building on AU, we propose a method to elicitate moral attributes to construct an explicit ethical goal function. To ensure alignment with societal values, this method allows for meaningful input from society, represented by a diverse sample of the population. This method comprises two key steps: first, defining an initial set of attributes, and second, refining this initial set through scenario-based attribute ranking and supplementation with any missing attributes (see Fig. 2).

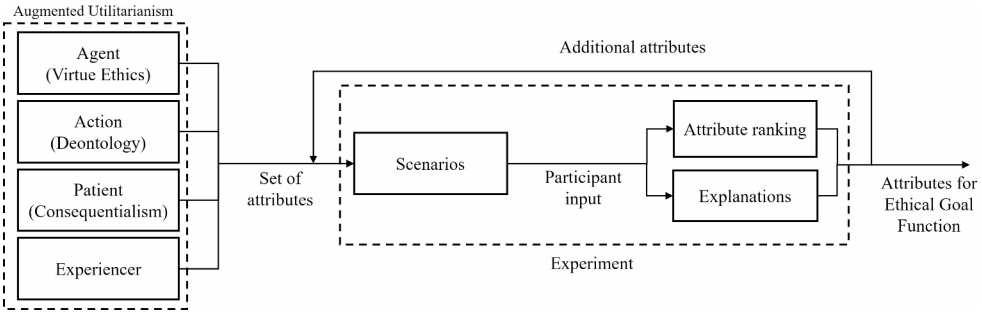

Fig. 2 Illustration of the experimental method

The initial phase of this experimental process entails defining a comprehensive set of attributes. While some studies have begun describing such a set, these frameworks often lack scientific grounding [34–37]. For example, the Moral Machine experiment asked thousands of participants to choose their preferred outcomes in various moral dilemmas [38]. The attributes used for decision-making included the number of individuals killed, their gender, age, and social status. However, the selection of these attributes was not scientifically grounded, leading to ambiguity in understanding the participants' decision-making processes. For instance, it is unclear whether children were favoured due to perceived vulnerability or based on the 'fair innings' philosophy, which advocates for everyone having the right to live a certain number of years. Additionally, the experiment's random comparison of attributes made it difficult to discern their individual impacts on decisions. Moreover, the study did not allow participants to suggest their own attributes or attempt to create an exhaustive set, further limiting its robustness and comprehensiveness.

In contrast, AU integrates insights from neuroscience, cognitive psychology, and ethical philosophy to ensure that moral attributes are both scientifically grounded and explainable. In addition, AU, being a non-normative framework, prioritises principles over normative ethical theories, aligning with the foundation of principlism. Principlism is an approach to ethics that emphasises the application of four core principles: autonomy, beneficence, non-maleficence, and justice [39]. These principles provide a flexible yet robust ethical framework that can be applied across various contexts without adhering to a single normative ethical theory [40].

This strategic alignment with principlism allows for the integration of these core ethical principles into the attribute definition process, ensuring that the attributes reflect essential ethical considerations. For instance, autonomy ensures that the attributes respect individual decision-making and personal freedom, beneficence promotes the well-being and positive outcomes for individuals and society, non-maleficence ensures that the attributes prevent harm and minimise potential negative impacts, and justice guarantees fairness and equity in the treatment of individuals and distribution of benefits and burdens.

By incorporating these principles, the process establishes a robust ethical foundation for the subsequent stages of attribute refinement and scenario analysis, ensuring that the ethical dimensions are thoroughly considered and integrated into the development of the technology. This principled approach ensures that the resulting attributes are not only technically sound but also ethically robust, aligning with broader societal values and ethical standards. This approach led to the identification of attributes such as physical damage, psychological damage, moral responsibility, legality, and damage to the vehicle [41].

The next challenge involves determining how to accurately capture societal values and who is responsible for defining them. Understanding and integrating societal values into the attribute definition process is crucial for developing ethically sound technologies. This can be achieved through participatory methods that involve diverse stakeholders, including the public, experts, and policymakers, to ensure a broad and inclusive representation of societal values.

To address this challenge, we propose an experimental process that allows individuals to voice their opinions and suggest additional attributes. This participatory approach ensures that the attributes reflect a wide range of perspectives and values. The defined attributes undergo rigorous testing across two distinct scenarios: high-risk and low-risk situations. High-risk scenarios might include situations where the potential for significant harm or ethical dilemmas is greater, such as emergency decision-making by AVs. Low-risk scenarios, on the other hand, could involve everyday situations with minimal ethical stakes.

Participants in the experiment are invited to rank the attributes based on their perceived importance in these scenarios, elucidate their decision-making process, and contribute additional attributes as needed. This approach not only captures the initial set of societal values but also allows for the identification of new attributes that may emerge from the participants' feedback. The collected feedback is then used to refine and expand the attribute set, ensuring it remains

comprehensive and reflective of evolving societal values [41].

To ensure that the evolving societal values are accurately reflected, the experiment incorporates a feedback loop. This iterative cycle allows for the ethical goal function to be dynamically adjusted, ensuring ongoing relevance and adaptability to changing norms and ethical considerations. This loop involves continuously revisiting and refining the attributes based on participant feedback and changes in societal norms and values. By doing so, the process remains dynamic and responsive to societal shifts, ensuring that the technology remains aligned with current ethical standards.

After participants contribute to defining and refining the attributes, regulators review these inputs to ensure they comply with existing laws and ethical guidelines. This oversight helps bridge the gap between experimental findings and legal requirements, providing a formal mechanism for integrating public input into regulatory frameworks. By validating the attributes, regulators ensure that the ethical decision-making processes of AVs are not only comprehensive and representative but also legally sound. Additionally, regulators can adapt these attributes as laws evolve, ensuring that AV technologies remain compliant and ethically responsible in the face of changing legal landscapes. This collaborative approach between public input and regulatory oversight ensures that AV decision-making is both ethically grounded and legally robust.

The rationale behind this approach lies in the belief that with a sufficiently large and representative sample of participants, an exhaustive list of attributes can be generated to fully define the decision-making processes of AVs. This inclusivity ensures that diverse perspectives are considered, addressing any inadequacies in the initial set of attributes through participant contributions and ensuring comprehensive coverage. While complete agreement among participants cannot be realistically achieved, by prioritizing transparency and inclusivity, we aim to build a consensus that, while not perfect, reflects a broad spectrum of societal values and ethical considerations. This dynamic approach ensures that the decision-making framework for AVs can adapt over time, accommodating shifts in societal norms and maintaining alignment with evolving ethical standards.

Ultimately, the objective of this endeavour is to cultivate a robust and transparent framework that incorporates societal values into the decision-making processes of AVs. By embedding ethical considerations at the core of AV design and operation, our aim is to bolster public trust and acceptance of automated vehicles while ensuring alignment with societal values and priorities. Through ongoing collaboration and iteration, we aspire to contribute to the responsible and ethical development of automated vehicle technology for the betterment of society as a whole.

6 Conclusion

In conclusion, this paper has addressed the ethical challenges confronting AV technologies and proposed viable solutions to surmount them. By highlighting the constraints of deep learning-based AI and advocating for transparent, interpretable, and ethically grounded methodologies like hybrid AI, we underscore the significance of ethical considerations in real-time AV decision-making. Technical hurdles such as data limitations and explainability issues, coupled with broader apprehensions regarding transparency and trustworthiness, underscore the intricacy of seamlessly integrating AI into AVs. Model-based approaches offer a promising avenue by explicitly embedding ethical principles and prioritising safety, fairness, and meaningful human control. However, they encounter challenges, particularly in terms of inefficiency when confronted with substantial amounts of unstructured data. Hybrid AI, blending symbolic and subsymbolic methodologies, presents a compelling strategy to effectively address these challenges. Augmented Utilitarianism is proposed as an ethical framework for AVs, with a focus on harm minimisation while upholding ethical tenets. Our method for eliciting moral attributes strives to construct an explicit ethical goal function, steering AV decision-making in harmony with societal values. By infusing ethics into the design and operation of AVs, we can bolster public trust and contribute to the conscientious advancement of automated vehicle technology.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Othman, K.: Exploring the implications of autonomous vehicles: a comprehensive review. Innov. Infrastruct. Solut. **7** (2022). https://doi.org/10.1007/s41062-022-00763-6
2. Schneider, B.: Robotaxis are Here. It's Time to Decide What to do About them. MIT Technology Review. (2023). https://www.

technologyreview.com/2023/06/23/1074270/obotaxis-decision-time/ Accessed 2 May 2024
3. On-Road Automated Driving (ORAD), Committee, J.: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles—SAE International, (2021). https://doi.org/10.4271/J3016_202104
4. Taniguchi, T., Murata, S., Suzuki, M., Ognibene, D., Lanillos, P., Ugur, E., Jamone, L., Nakamura, T., Ciria, A., Lara, B., Pezzulo, G.: World models and predictive coding for cognitive and developmental robotics: frontiers and challenges. Adv. Robot. **37**, 780–806 (2023). https://doi.org/10.1080/01691864.2023.2225232
5. Perumal, P.S., Sujasree, M., Chavhan, S., Gupta, D., Mukthineni, V., Shimgekar, S.R., Khanna, A., Fortino, G.: An insight into crash avoidance and overtaking advice systems for autonomous vehicles: a review, challenges and solutions. Eng. Appl. Artif. Intell. **104** (2021). https://doi.org/10.1016/j.engappai.2021.104406
6. Wiseman, Y., Grinberg, I.: Circumspectly crash of autonomous vehicles. In: IEEE International Conference on Electro Information Technology (2016). https://doi.org/10.1109/EIT.2016.7535271
7. Li, H., Zheng, T., Xia, F., Gao, L., Ye, Q., Guo, Z.: Emergency collision avoidance strategy for autonomous vehicles based on steering and differential braking. Sci. Rep. **12** (2022). https://doi.org/10.1038/s41598-022-27296-3
8. Eckersley, P.: Impossibility and uncertainty theorems in AI value alignment (or why your AGI should not have a utility function). SafeAI. (2019). https://doi.org/10.48550/arXiv.1901.00064
9. Hu, A., Russell, L., Yeo, H., Murez, Z., Fedoseev, G., Kendall, A., Shotton, J., Corrado, G.: GAIA-1: a generative world model for autonomous driving (2023). https://doi.org/10.48550/arXiv.2309.17080
10. Dengel, A., Etzioni, O., DeCario, N., Hoos, H., Li, F.F., Tsujii, J., Traverso, P.: Next Big challenges in core AI technology. In: Lecture notes in Computer Science (including subseries lecture notes in Artificial Intelligence and Lecture notes in Bioinformatics). Springer Sci. Bus. Media Deutschland GmbH. 90–115 (2021). https://doi.org/10.1007/978-3-030-69128-8_7
11. Van Harmelen, F., Ten Teije, A.: A boxology of design patterns for hybrid learning and reasoning systems. In: CEUR Workshop Proc (2019). https://doi.org/10.13052/jwe1540-9589.18133
12. Marcus, G., Critical Appraisal, D.L.A.: CoRR (2018). https://doi.org/https://doi.org/10.48550/arXiv.1801.00631
13. European Commission: Directorate-General for Communications Networks, Content and Technology, Ethics guidelines for trustworthy AI (2019). https://doi.org/https://data.europa.eu/doi/10.2759/346720
14. Roli, A., Jaeger, J., Kauffman, S.A.: How organisms come to know the World: fundamental limits on Artificial General Intelligence. Front. Ecol. Evol. **9** (2022). https://doi.org/10.3389/fevo.2021.806283
15. Hawkins, A.J.: Waymo's robotaxis are under investigation for crashes and traffic law violations. (2024). https://www.theverge.com/2024/5/14/24156238/waymo-nhtsa-investigation-crash-wrong-side-road Accessed 23 May 2024
16. de Sio, F.S., van den Hoven, J.: Meaningful human control over autonomous systems: A philosophical account. Front. Rob. AI. **5** (2018). https://doi.org/10.3389/frobt.2018.00015
17. Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., Song, D.: Robust. Physical-World Attacks Deep Learn. Models. (2017). https://doi.org/10.48550/arXiv.1707.08945
18. Ferrara, E.: Fairness and Bias in Artificial Intelligence: A brief survey of sources, impacts, and mitigation strategies. Sci. **6** (2024). https://doi.org/10.3390/sci6010003
19. Schwartz, M.S.: Ethical decision-making theory: An Integrated Approach. J. Bus. Ethics. **139** (2016). https://doi.org/10.1007/s10551-015-2886-8
20. Yang, C.Y.D., Fisher, D.L.: Safety impacts and benefits of connected and automated vehicles: How real are they? J. Intell. Transp. Systems: Technol. Plann. Oper. **25**, 135–138 (2021). https://doi.org/10.1080/15472450.2021.1872143
21. Nyholm, S.: The ethics of crashes with self-driving cars: a roadmap, I. Philos. Compass. **13** (2018). https://doi.org/10.1111/phc3.12507
22. Sparrow, R., Howard, M.: When human beings are like drunk robots: Driverless vehicles, ethics, and the future of transport. Transp. Res. Part. C Emerg. Technol. **80**, 206–215 (2017). https://doi.org/10.1016/j.trc.2017.04.014
23. Bieber, C.: 93% Have Concerns About Self-Driving Cars– Forbes Advisor. (2024). https://www.forbes.com/advisor/legal/auto-accident/perception-of-self-driving-cars/ Accessed 2 May 2024
24. von Eschenbach, W.J.: Transparency and the Black Box Problem: Why we do not trust AI. Philos. Technol. **34**, 1607–1622 (2021). https://doi.org/10.1007/s13347-021-00477-0
25. Alam, M., Groth, P., Hitzler, P., Paulheim, H., Sack, H., Tresp, V.: Symbolic Vs Sub-symbolic AI Methods: Friends or Enemies? In: International Conference on Information and Knowledge Management, Proceedings, Association for Computing Machinery, pp. 3523–3524 (2020). https://doi.org/10.1145/3340531.3414072
26. Aliman, N.M., Kester, L., Werkhoven, P., Yampolskiy, R.: Orthogonality-based disentanglement of responsibilities for ethical intelligent systems. In: Hammer, P., Agrawal, P., Goertzel, B., Iklé, M. (eds.) Artificial General Intelligence. AGI 2019. Lecture Notes in Computer Science(), pp. 22–31. Springer (2019). https://doi.org/10.1007/978-3-030-27005-6_3
27. van der Ploeg, C., Braat, M., Masini, B., Brouwer, J., Paardekooper, J.-P.: Connecting the Dots: Context-Driven Motion Planning Using Symbolic Reasoning. In: 2023 IEEE Intelligent Vehicles Symposium (IV) (2023). https://doi.org/10.1109/IV55152.2023.10186794
28. Cullen, J., Bryman, A.: The Knowledge Acquisition Bottleneck: Time for reassessment? Expert Syst. **5** (1988). https://doi.org/https://doi.orghttps://doi.org/10.1111/j.1468-0394.1988.tb00065.x
29. Wernaart, B.F.W.: 1. An introduction to moral design and technology. In: Moral Design and Technology, Brill, pp. 13–23. Wageningen Academic (2022). https://doi.org/10.3920/978-90-8686-922-0_1
30. Risk (Stanford Encyclopedia of Philosophy): (2022). https://plato.stanford.edu/entries/risk/ Accessed 23 May 2024
31. Aliman, N.-M., Kester, L.: Crafting a flexible heuristic moral meta-model for meaningful AI control in pluralistic societies. In: B. Wernaart (Ed.), Moral Design and Technology, pp. 63–80. Wageningen Academic (2022). https://doi.org/10.3920/978-90-8686-922-0_4
32. Schein, C., Gray, K.: The theory of Dyadic Morality: Reinventing Moral Judgment by redefining harm. Personality Social Psychol. Rev. **22** (2018). https://doi.org/10.1177/1088868317698288
33. Aliman, N.-M., Kester, L.: Augmented utilitarianism for AGI Safety. In: Hammer, P., Agrawal, P., Goertzel, B., Iklé, M. (eds.) Artificial General Intelligence. AGI 2019. Lecture Notes in Computer Science. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-27005-6_2
34. Bergmann, L.T., Schlicht, L., Meixner, C., König, P., Pipa, G., Boshammer, S., Stephan, A.: Autonomous vehicles require socio-political acceptance—an empirical and philosophical perspective on the problem of moral decision making. Front. Behav. Neurosci. **12** (2018). https://doi.org/10.3389/fnbeh.2018.00031
35. Li, J., Zhao, X., Cho, M.J., Ju, W., Malle, B.F., SAE International: from trolley to autonomous vehicle: perceptions of

responsibility and moral norms in traffic accidents with self-driving cars. In:: SAE Technical Papers (2016). https://doi.org/10.4271/2016-01-0164
36. Faulhaber, A.K., Dittmer, A., Blind, F., Wächter, M.A., Timm, S., Sütfeld, L.R., Stephan, A., Pipa, G., König, P.: Human decisions in Moral dilemmas are largely described by Utilitarianism: Virtual Car driving study provides guidelines for Autonomous Driving vehicles. Sci. Eng. Ethics. **25** (2019). https://doi.org/10.1007/s11948-018-0020-x
37. Kallioinen, N., Pershina, M., Zeiser, J., Nosrat Nezami, F., Pipa, G., Stephan, A., König, P.: Moral judgements on the actions of self-driving cars and human drivers in dilemma situations from different perspectives. Front. Psychol. **10** (2019). https://doi.org/10.3389/fpsyg.2019.02415
38. Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.F., Rahwan, I.: The Moral Machine experiment. Nature. **563**, 59–64 (2018). https://doi.org/10.1038/s41586-018-0637-6
39. Beauchamp, T.L., Childress, J.F.: Principles of Biomedical Ethics, 8th edn. Oxford University Press (2019)
40. Scher, S., Kozlowska, K.: The rise of bioethics: a historical overview. In: Rethinking Health Care Ethics, pp. 31–44. Springer Singapore (2018). https://doi.org/10.1007/978-981-13-0830-7_3
41. Gros, C., Werkhoven, P., Kester, L., Martens, M.: Defining a method for ethical decision making for automated vehicles. In: ICAIL2023 (2023). https://doi.org/10.13140/RG.2.2.34735.71844