



Human-Robot Dialogue that Elicits The Alignment of Moral Principles For Driverless Vehicles

Georgios Angelopoulos

Interdepartmental Center for Advances in Robotic Surgery
Università degli Studi di Napoli Federico II (UNINA)
80125 Naples, Italy
georgios.angelopoulos@unina.it

Vladimir Estivill-Castro

Dept of Information and Communications Technologies
Universitat Pompeu Fabra (UPF)
08018 Barcelona, Spain
vladimir.estivill@upd.edu

ABSTRACT

The emergence of autonomous vehicles has raised ethical considerations regarding their controlling software. The focus is on defining ethical settings that determine the response to moral dilemmas, akin to the Trolley Problem. The study explores the interaction with a robot that engages users in a dialogue about such ethical dilemmas, allowing users to align vehicles' behaviour with their ethical preferences. Additionally, providers of these vehicles can have a codified version of user preferences to address potential real-world issues, for instance, by adjusting insurance premiums. The research details designing and implementing a human-robot interaction system for eliciting ethical settings in autonomous vehicles.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; **User centered design**; • **Computer systems organization** → **Robotic autonomy**; **Robotic autonomy**.

KEYWORDS

Ethical Dilemma, Trolley Problem, Human-Robot Interaction

ACM Reference Format:

Georgios Angelopoulos and Vladimir Estivill-Castro. 2024. Human-Robot Dialogue that Elicits The Alignment of Moral Principles For Driverless Vehicles. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24 Companion)*, March 11–14, 2024, Boulder, CO, USA. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3610978.3640641>

1 INTRODUCTION

“Roboethics” [31, 33] focuses on the positive and negative repercussions of robotic technologies on society. The term roboethics encompasses ethical considerations guiding the moral conception, development, and utilization of robots, particularly intelligent and autonomous ones. Key issues include the dual potential of robots for beneficial or harmful applications, the anthropomorphisation of robots, the evolving dynamics of human-robot symbiosis, the

mitigation of the socio-technological disparity, and the impact of robotics on equitable wealth and power distribution.

The initial ethical framework for robots draws inspiration from Asimov’s three laws [4], which, although fictional, are inherently anthropocentric, emphasizing robots’ role in serving humanity. These laws imply that robots possess sufficient intelligence, including perception and cognition, allowing them to autonomously make morally informed decisions based on specified rules, regardless of the complexity of encountered situations.

Advances in logical reasoning have raised expectations for ethical behaviour in artificial agents [2, 3]. Progress has been notable in utilizing various logics to mechanize moral problem-solving [21, 30], including addressing classic dilemmas like the Trolley Problem [29]. This dilemma involves a runaway trolley heading towards several individuals, and an observer must decide whether to divert it to save five lives at the expense of causing harm to one person on an alternate track. The ongoing philosophical debate highlights the conflict between consequentialist and non-consequentialist perspectives. With the emergence of autonomous vehicles, such ethical scenarios may potentially become real-world challenges [6, 7]. The study addresses real-world concerns about user acceptance of autonomous vehicles if settings do not align with users’ preferences. We employ human-robot dialogues on moral dilemmas, such as the Trolley Problem. Our results indicate that interacting with a robot enables users to align the behaviour of autonomous vehicles with their ethical preferences.

2 RELATED WORK

In the rapidly advancing field of autonomous vehicles and agents, ethical decision-making has emerged as a critical area of research. Goodall *et al.* [17] has provided a foundation for ethical considerations in the decision-making processes of automated vehicles. Awad *et al.* [5] have further contributed to this discussion by establishing the “Moral Machine”, an online platform that presents various moral traffic scenarios. This interactive tool allows users to dictate the vehicle’s responses in these situations, offering a unique perspective on moral judgement. The social dilemma of decision-making in autonomous vehicles was highlighted by Bonnefon *et al.* [7]. Their research revealed a preference among participants for utilitarian algorithms in automated vehicles. However, this preference shifted when participants were faced with scenarios that risked their own lives [22]. In such cases, they demanded that automated vehicles prioritise passenger safety above all else. This led to a general opposition towards enforcing utilitarian regulations, with participants indicating they would be unlikely to purchase vehicles operating under such guidelines.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

HRI '24, March 11–14, 2024, Boulder, Colorado

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0323-2/24/03...\$15.00
<https://doi.org/10.1145/3610978.3640641>

Frison *et al.* [14] utilised the Trolley Problem in a driving simulation experiment to assess if individuals would be prepared to give up their own lives for the benefit of others. Their findings contrast with others [7, 22] and indicate that individuals would lean towards a utilitarian decision, even if it implies sacrificing their own life. The rationale behind such decisions encompassed the anticipation of a system behaving logically and personal implications (for instance, the reluctance to own a vehicle that could potentially harm others). In a different approach, Fournier [13] proposed the development of “command profiles” for each driver, which could be established through a series of questions demanding ethical decisions under varying and challenging driving conditions. This profile could subsequently be applied to the autonomous vehicle.

After an examination of the ethical considerations related to decision-making algorithms in automated vehicles, Mirnig *et al.* [20] argue that these algorithms should not solely focus on driving maneuvers based on environmental data but also take ethical aspects into account. Specifically, they address situations where all available decisions have fatal consequences, resembling the Trolley Problem dilemma. They advocate for research and design efforts to avoid scenarios akin to the Trolley Problem altogether, rather than attempting to find optimal solutions to an inherently unsolvable dilemma. The discussion also touches on alternative approaches to feasibly address ethical concerns in automated agents. Additionally, a German government commissioned report [10], proposes attributing any tragedy involving autonomous vehicles to humans rather than machines, even in situations resembling the Trolley Problem. While the Trolley Problem helps analyse the concerns regarding how to morally program autonomous vehicles, arguments have also been made for why ethical considerations surrounding design algorithm for autonomous vehicles diverge from the Trolley Problem’s philosophical underpinnings [23].

Despite the significant strides made in the field of ethical decision-making in autonomous agents, there is still a need for a more personalised model. Fournier [13] insists artificial autonomous agents should understand and respond in the same way their human users do to ethical dilemmas. However, the specific methodologies and implementations for creating such a personalised model remains an open area of research. Our study aims to explore this and contribute to creating ethical decision-making models in autonomous agents.

3 PROPOSED APPROACH

Various proposals have been put forward advocating the design of AI systems that are capable of reasoning about and explaining the virtue of their actions [18] or otherwise, exhibit moral agency [1, 15, 35]. A second challenge is the “value alignment problem”, which requests that AI systems shall display moral values that align with those of humanity. For these two challenges, some researchers propose to supply AI with moral-decision making ability using deontic reasoning [8, 26] or conceptual spaces. An example for the latter approach is the idea to construct a geometric space of moral principles from a set S of already solved ethical cases; then we make decisions on a new case c by finding the case s in S nearest to c and adopting for c the decision as per s [25]. This can be either categorized as case-based reasoning [27] or instance-based learning [36]. Our approach follows these lines, but it introduces

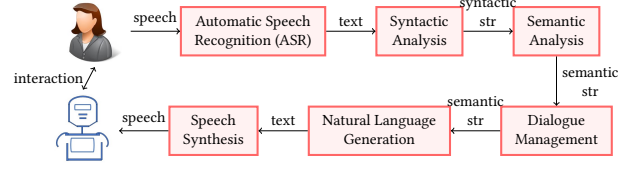


Figure 1: Our architecture for our frame-based conversational system.

a mechanism for creating a personalized model. That is, we solve align the moral values to one user and deploy a robot that understands and responds to human ethical dilemmas in the same way their user would. This approach is based on two distinct methods: the *ki*-Nearest Neighbours (kNN) algorithm (the most well-known methods for instance-based learning) and an in-house developed algorithm to provide the metric among four ethical principles. For this, we designed and deployed a frame-based conversational system [16] on a humanoid robot, Aldebaran’s Pepper (already used in real-world applications such as hotel receptionist, car & coffee-machine salesman, and tourism guide) since the embodiment can help facilitate more natural and engaging interactions [34].

Pepper (equipped with our software) interacts with a person, and the system elicits the information regarding personal views on variants of the Trolley Problem. Thus, the interaction enables learning the boundaries for solutions to ethical dilemmas. The dialogue model of the system guides the system in characterising the user input and performing an appropriate response. Figure 1 depicts the dialogue management architecture.

We first selected the *k*-Nearest Neighbors (kNN) algorithm for its simplicity and minimal user input requirement. This type of lazy learning algorithm does not require offline training. The algorithm calculates a weighted Euclidean distance between a new example e and all training data to identify the k nearest neighbours to e and produces the classification output as the majority vote of those k nearest neighbours. The weighted Euclidean distance between points A and B in a feature space is calculated using the normalized Euclidean metric. Let A and B be represented by feature vectors $A = (x_1, x_2, \dots, x_m)$ and $B = (y_1, y_2, \dots, y_m)$, where m is the dimensionality of the feature space. The normalized Euclidean metric is used to calculate the distance between A and B is defined by

$$\text{dist}(A, B) = \sqrt{\frac{\sum_{i=1}^m (x_i - y_i)^2}{m}} \quad (1)$$

Aimed at preventing both underfitting and overfitting of the model, through multiple iterations, we identified the optimal value for k representing the number of neighbours, to be 3.

Our second method operates based on four criteria: Saving More Lives, Protecting Passengers, Avoiding Intervention, and Upholding the Law. Each factor is represented by χ_i , and each time a human responds to a robot’s question, a weight w_i is applied to each factor. The weights are encoded in a sequence of pairs

$$\langle (w_1, \chi_1), (w_2, \chi_2), (w_3, \chi_3), (w_4, \chi_4) \rangle. \quad (2)$$

We selected these criteria from our insights derived from the Moral Machine [5]. However, we intentionally excluded some factors (age

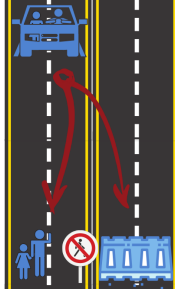


Figure 2: The user is presented with a case and asked to decide whether the human driver shall save the passengers or the pedestrians.

and gender) to circumvent potential ethical biases that could arise from their consideration [24].

The algorithm generates questions to discern an individual's preferences when two weights are similar. In a specific scenario depicted in Figure 2, a question is posed concerning the driving control faced with the dilemma of causing harm to pedestrians or passengers. Through dialogue and interaction, the method reaches a setting where each weight significantly differs from another, revealing a clear preference hierarchy from the user. Consequently, the method effectively elicits the moral value preferences of the individual, acknowledging the inherent variability in such dilemmas among different people. The target is the value pairs so that $\forall i \neq j, (w_i, \chi_i) \neq (w_j, \chi_j)$. For instance, if $(w_1 = w_2)$, the robot will create a condition where the criteria (w_1) and (w_2) are in conflict with one another. The setting is presented on the tablet of the robot and explained to the user before asking which of the two evils the human finds acceptable in contrast to the other.

We experimented with this two approaches to determine which method captures more accurately user's beliefs on variants of the Trolley Problem. The kNN algorithm provides a solid foundation for classifying responses based on their proximity in the feature space. It can be particularly effective when decision boundaries are irregular and need to be learned from the data. On the other hand, the in-house developed algorithm allows for explainability of human ethical considerations. It operates based on the four factors. By applying weights to each factor based on the human's response to a robot's question, the algorithm can distinguish the individual's preferences and create a clear preference hierarchy. These preferences are used to make decisions in new cases by applying the hierarchy in order of the weights. An example of user's preferences is depicted in Figure 3.

4 EVALUATION

4.1 The learning algorithms

In our preliminary assessment, we surveyed eleven volunteers with 4-Likert scale question to gauge their level of agreement with the robot's decisions when utilizing our two learning methods: our in-house algorithm to elicit the ranking of criteria and the machine learning algorithm (kNN). Each participant was exposed to both algorithms. The presentation order between the two was randomized, ensuring an unbiased assessment of their preferences. The findings

indicated a clear preference for our in-house algorithm, where the mean level of agreement was $M = 3.73$ with a standard deviation of $SD = 0.47$. This indicates a high level of agreement among the participants with the decisions made by the robot when using our in-house algorithm. On the other hand, for the kNN algorithm, the mean level of agreement was $M = 2.91$ with a standard deviation of $SD = 0.55$. This suggests a lower level of agreement among the participants with the decisions made by the robot when using the kNN algorithm. Although this evaluation corresponds to a limited number of participants, our results, however, indicate some strong direction for the ranking of criteria as a more acceptable approach to elicit the moral values of a user and to achieve alignment of moral decision with a user.

4.2 The Human-Robot interaction

The iterative assessment of robot interaction revealed that relying solely on verbal feedback led to a diminished sense of anthropomorphism, making the interaction feel unnatural and failing to enhance lifelikeness, engagement, and likability. Human face-to-face conversations typically involve both verbal and non-verbal expressions. Verbal feedback serves various purposes, such as seeking information, responding to questions, presenting new information, and expressing understanding or uncertainty. Backchannels like 'ah,' 'uhu,' and 'mhm' are used to encourage the speaker. Importantly, non-verbal cues such as gestures and eye gaze complement verbal expressions intentionally and play a significant role in communication, rather than being mere artifacts in a conversation.

Thus, we evolved our prototype to effectively engage with humans. We noticed the Pepper should possess the capability for both verbal and non-verbal communication. Thus, we used the ALAnimated Speech API. This module facilitated the integration of gesturing and expressing emotions while participants were speaking into the humanoid. Additionally, for a more dynamic Human-Robot Interaction scheme, the robot required a gaze behaviour designed to create the impression of eye contact with the human interlocutor.

In the next iteration, we utilized the preinstalled ALFaceTracker library, to implement the gaze behaviour in the 'eye contact' condition. Since this library enables face identification in the images/video captured by Pepper's camera, our software could detect the user's face, and steer Pepper's head to maintain the detected face in the centre of the image, creating the impression that Pepper is directing its gaze toward the participant when delivering utterances. Our findings indicate that the robot receives more positive evaluations when displaying eye contact and non-verbal behaviours such as hand and arm gestures in conjunction with speech.

4.3 SWOT analysis

We applied the SWOT analysis tool [12], to categorize both internal and external factors related to the design outcomes of our prototype.

S:Strengths : The decision-making knowledge is explicit and can be used for explanations, contributing to the perception of smart robots. The architecture delegates intensive computing tasks to the cloud, leading to extended battery life.

W:Weaknesses Functionality is dependent on networking connectivity and remote services, with no cybersecurity.

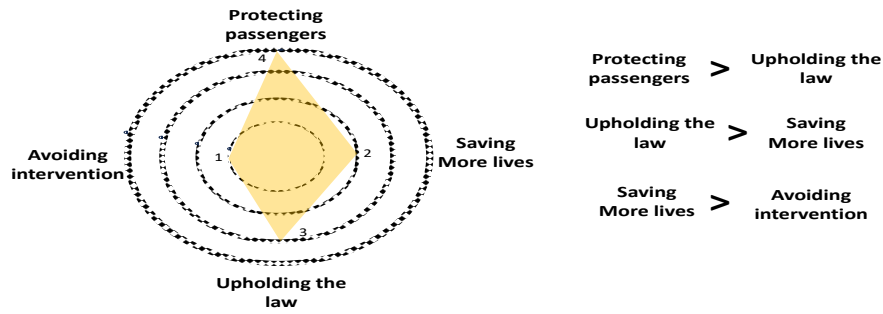


Figure 3: In this example, the human prioritises the protection of passengers.

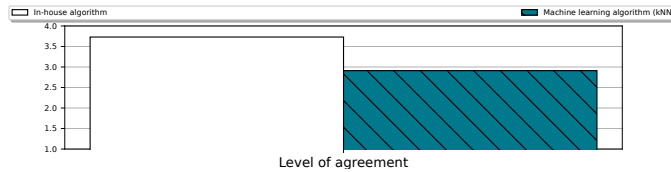


Figure 4: Level of agreement

A dialogue with a humanoid robot may not be the optimal mechanism by which human users' values are aligned with those of their autonomous vehicles.

O:Opportunities This interaction could result in more awareness by the public of some of the ethical challenges posed by emerging technologies.

T:Threats The interaction with a robot to elicit the moral criteria from users of an autonomous vehicle may impact how genuine are the participants' responses. The small size sample of 11 participants implies low confidence in the results, a larger set of participants with comprehensive characteristics (diversity in gender, culture, age, etc.) is necessary to confirm this preliminary results.

5 FINAL REMARKS

5.1 Next steps

In future research, we aim to gain a deeper understanding of how participants perceive and trust robots that are equipped with our proposed ethical decision-making mechanism. To achieve this, we will equip again the Pepper robot with our proposed mechanism, thereby creating a unique profile for each participant based on their ethical decisions. Subsequently, participants will be guided to complete the Robotic Social Attributes Scale (RoSAS) Questionnaire, which is designed to assess their perceptions of the robot's social attributes [9]. This questionnaire will provide valuable insights into how participants perceive the robot's warmth, competence, and discomfort. We will also administer the Multi-Dimensional Measure of Trust (MDMT) v2 questionnaire to evaluate participants' levels of trust in the robot [32]. The MDMT questionnaire is a comprehensive tool that measures trust across multiple dimensions, providing a nuanced understanding of participants' trust levels.

5.2 Conclusions

In the near future, numerous mobile robots will operate in public spaces, making swift ethical decisions that impact humans [11]. For real-world use, these robots should acquire ethical behaviour skills. We propose that the dialogue of the robot with its users (on matters of ethical dilemmas) results in a robot that aligns its moral values with those of that user. Given that humans participate in society with moral values that diverge (even on matters of autonomous vehicles behaviour [28]), it seems acceptable that robots in human environments would behave with moral values aligned with their users. Our setup grants the robot the ability to engage with humans and capture their ethical perspectives (engagement between users and their driverless cars has been postulated as the basis of ethics for driverless cars [19]). We have reported here on our prototype software for a humanoid robot. Our experiments show that the robot achieves a satisfactory level of alignment with its user (at least for the challenging variants of the Trolley problem).

REFERENCES

- [1] Alison Adam. 2008. Ethics for things. *Ethics and Information Technology* 10, 2 (2008), 149–154. <https://doi.org/10.1007/s10676-008-9169-3>
- [2] Sabah S. Al-Fedaghi. 2008. Typification-based ethics for artificial agents. In *2008 2nd IEEE International Conference on Digital Ecosystems and Technologies* (Phitsanuloke, Thailand). IEEE, 482–491. <https://doi.org/10.1109/DEST.2008.4635149>
- [3] Sofia Almpiani, Petros Stefaneas, and Panayiotis Frangos. 2022. Argumentation-Based Logic for Ethical Decision Making. *Studia Humana* 11, 3-4 (2022), 46–52.
- [4] Isaac Asimov. 1942. Runaround. *Astounding Science Fiction* 29, 1 (1942), 94–103.
- [5] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The moral machine experiment. *Nature* 563, 7729 (2018), 59–64.
- [6] Choton Basu, Manu Madan, and Balaji Sankaranarayanan. 2019. Are driverless cars truly a reality? A technical, social and ethical analysis. *Issues in Information Systems* 20, 4 (2019), 56–64.
- [7] Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. 2016. The social dilemma of autonomous vehicles. *Science* 352, 6293 (2016/11/29 2016), 1573–1576. <https://doi.org/10.1126/science.aaf2654>
- [8] S. Bringsjord, N. G. Sundar, B. F. Malle, and M. Scheutz. 2018. Contextual Deontic Cognitive Event Calculi for Ethically Correct Robots. In *International Symposium on Artificial Intelligence and Mathematics, ISAIM* (Fort Lauderdale, Florida, USA). University of Virginia.
- [9] Colleen M. Carpinella, Alisa B. Wyman, Michael A. Perez, and Steven J. Stroessner. 2017. The robotic social attributes scale (RoSAS) development and validation. In *Proceedings of the 2017 ACM/IEEE International Conference on human-robot interaction* (Vienna, Austria). IEEE, 254–262.
- [10] U. Di Fabio et al. 2017. *Ethics Commission Automated and Connected Driving*. Technical Report. Federal Ministry of Transport and Digital Infrastructure, Germany. www.mbd.de
- [11] Gábor Erdélyi, Olivia Johanna Erdélyi, and Vladimir Estivill-Castro. 2021. Randomized Classifiers vs Human Decision-Makers: Trustworthy AI May Have to Act Randomly and Society Seems to Accept This. *CoRR* abs/2111.07545 (2021).

- arXiv:2111.07545 <https://arxiv.org/abs/2111.07545>
- [12] Lawrence G. Fine. 2010. *The SWOT analysis: using your strength to overcome weaknesses, using opportunities to overcome threats*. Kick It.
 - [13] Tom Fournier. 2016. Will my next car be a libertarian or a utilitarian?: Who will decide? *IEEE Technology and Society Magazine* 35, 2 (2016), 40–45.
 - [14] Anna-Katharina Frison, Philipp Wintersberger, and Andreas Riener. 2016. First person trolley problem: Evaluation of drivers' ethical decisions in a driving simulator. In *Adjunct proceedings of the 8th international conference on automotive user interfaces and interactive vehicular applications* (Ann Arbor, MI, USA) (*AutomotiveUI '16 Adjunct*). Association for Computing Machinery, New York, NY, USA, 117–122.
 - [15] D. Fuenmayor and C. Benzmlüller. 2020. Normative Reasoning with Expressive Logic Combinations. In *ECAI 2020 - 24th European Conference on Artificial Intelligence* (Santiago de Compostela, Spain) (*Frontiers in Artificial Intelligence and Applications, Vol. 325*), G. De Giacomo, A. Catalá, B. Dilkina, M. Milano, S. Barro, A. Bugarin, and J. Lang (Eds.). IOS Press, Amsterdam, 2903–2904.
 - [16] David Goddeau, Helen Meng, Joseph Polifroni, Stephanie Seneff, and Senis Busayapongchai. 1996. A form-based dialogue manager for spoken language applications. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96* (Philadelphia, PA, USA), Vol. 2. IEEE, 701–704.
 - [17] Noah J Goodall. 2014. Ethical decision making during automated vehicle crashes. *Transportation Research Record* 2424, 1 (2014), 58–65.
 - [18] J. S. Hall. 2011. Chapter 3: Ethics for Machines. In *Machine Ethics*, M. Anderson and S. L. Anderson (Eds.). Cambridge University Press, UK, 28–44.
 - [19] Neil McBride. 2016. The Ethics of Driverless Cars. *ACM SIGCAS Computers and Society* 45, 3 (2016), 179–184. <https://doi.org/10.1145/2874239.2874265>
 - [20] Alexander G Mirnig and Alexander Meschtscherjakov. 2019. Trolled by the trolley problem: on what matters for ethical decision making in automated vehicles. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–10.
 - [21] Luis Moniz Pereira and Ari Saptawijaya. 2018. *Programming Machine Ethics*. Springer International Publishing AG, Switzerland.
 - [22] S. Moore. 2016. Driverless cars should sacrifice their passengers for the greater good – just not when I'm the passenger. <https://theconversation.com/driverless-cars-should-sacrifice-their-passengers-for-the-greater-good-just-not-when-im-the-passenger-61363>
 - [23] Sven Nyholm and Jilles Smids. 2016. The Ethics of Accident-Algorithms for Self-Driving Cars: An Applied Trolley Problem? *Ethical Theory and Moral Practice* 19, 5 (2016), 1275–1289. <https://doi.org/10.1007/s10677-016-9745-2>
 - [24] Giulia Perugia, Stefano Guidi, Margherita Bicchì, and Oronzo Parlangeli. 2022. The shape of our bias: Perceived age and gender in the humanoid robots of the ABOT database. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 110–119.
 - [25] Martin Peterson. 2019. The Value Alignment Problem: A Geometric Approach. *Ethics and Inf. Technol.* 21, 1 (March 2019), 19–28. <https://doi.org/10.1007/s10676-018-9486-0>
 - [26] Y. U. Ryu and R. M. Lee. 1995. Defeasible deontic reasoning and its applications to normative systems. *Decision Support Systems* 14, 1 (1995), 59 – 73. [https://doi.org/10.1016/0167-9236\(94\)00002-A](https://doi.org/10.1016/0167-9236(94)00002-A)
 - [27] Ram D. Sriram. 1997. Analogical and Case-Based Reasoning. In *Intelligent Systems for Engineering: A Knowledge-based Approach*. Springer London, London, 285–334.
 - [28] Kazuya Takaguchi, Andreas Kappes, James M Yearsley, Tsutomu Sawai, Dominic JC Wilkinson, and Julian Savulescu. 2022. Personal ethical settings for driverless cars and the utility paradox: An ethical analysis of public attitudes in UK and Japan. *Plos one* 17, 11 (2022), e0275812.
 - [29] Judith Jarvis Thomson. 1985. The Trolley Problem. *The Yale Law Journal* 94, 6 (1985), 1395–1415.
 - [30] Suzanne Tolmeijer, Markus Kneer, Cristina Sarasua, Markus Christen, and Abraham Bernstein. 2021. Implementations in Machine Ethics: A Survey. *ACM Comput. Surv.* 53, 6, Article 132 (dec 2021), 38 pages. <https://doi.org/10.1145/3419633>
 - [31] Spyros G. Tzafestas. 2018. Ethics in robotics and automation: a general view. *International Robotics & Automation Journal* 4, 3 (2018), 229–234.
 - [32] Daniel Ullman and Bertram F Malle. 2019. Measuring gains and losses in human-robot trust: Evidence for differentiable components of trust. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 618–619.
 - [33] Gianmarco Veruggio. 2005. The birth of roboethics. In *Proceedings of ICRA'2005: IEEE International Conference on Robotics and Automation: Workshop on Roboethics* (Barcelona, Spain). 1–4.
 - [34] Joshua Wainer, David J Feil-Seifer, Dylan A Shell, and Maja J Mataric. 2006. The role of physical embodiment in human-robot interaction. In *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 117–122.
 - [35] W. Wallach and C. Allen. 2009. *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, Oxford.
 - [36] I. H. Witten, E. Frank, and M. A. Hall. 2011. *Data Mining: Practical Machine Learning Tools and Techniques* (3rd ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.