Consensus-determinacy Space and Moral Components for Ethical Dilemmas

Yongxu Liu[1], Yan Liu[1], Gong Chen[1], Yuexian Hou[2], Sheng-hua Zhong[3]

[1] Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China
[2] College of Computer Science and Technology, Tianjin University, Tianjin, China
[3] College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
yongxu.liu@connect.polyu.hk,{csyliu,csgchen}@comp.polyu.edu.hk,yxhou@tju.edu.cn,csshzhong@szu.edu.cn

ABSTRACT

This paper models ethical dilemmas, which take place in a decision-making context where any of the available options requires the agent to violate or compromise on their ethical standards. A famous dilemma, the "Trolley Problem", has been studied quantitatively and systematically since it represents machine ethics and has many applications in autonomous vehicles. We design a psychological paradigm to collect the data on the decisions and confidence of the participants facing the dilemmas, and a consensus-determinacy space is defined for human ethics. Then, we formulate the moral principle analysis as the matrix factorization problem, and a new model is proposed to discover moral components. Based on the embedded moral principles, we explore the possibility of providing decisions that are more consistent with the behavior of human beings, even if the available dataset is small and incomplete. Several experiments have been conducted on the proposed model to discuss the necessity and feasibility of the research in machine ethics.

KEYWORDS

Machine ethics; artificial morality; AI ethics in practice

ACM Reference Format:
Yongxu Liu[1], Yan Liu[1], Gong Chen[1], Yuexian Hou[2], Sheng-hua Zhong[3]. 2022. Consensus-determinacy Space and Moral Components for Ethical Dilemmas. In *Proceedings of 1st ACM SIGKDD Workshop on Ethical Artificial Intelligence (EAI-KDD' 22).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

With the development of artificial intelligence (AI), every aspect of people's daily life has been impacted, from social media to daily commutes to online shopping [10]. However, ethical concerns about these widely seen and used autonomous intelligent systems still lack efficient and explicit solutions [17], resulting in public stress and anxiety about artificial intelligence [21, 27]. The AI-powered chatbot Tay launched by Microsoft was forced offline and blocked just 24 hours after its operation. Tay could not recognize whether its statements are offensive and have contents related to racial discrimination or not [19].

Some foreseeing researchers have realized that if we want AI to be trustworthy, AI systems need the ability to tell wrong from right or the explicit moral framework enabling the ethics governance [1, 6, 9]. For this reason, there is a growing research interest in machine ethics, which is defined as one principle or a set of ethical principles guiding the work of intelligent machines [5, 25, 26]. Current research on machine ethics can be roughly divided into two kinds. The first type of work is about designing a general paradigm for studying machine ethics based on ethical principles. The most well-known ethical principles are Asimov's Laws. The most important law of Asimov's Laws is that robots are not allowed to injure humans [18]. Deng et al. [8] designed a dilemma in which one robot embedded with Asimov's laws chooses which of two human robots is going to die. Past autonomous vehicles dilemma studies have mostly discussed utilitarianism, debating whether to minimize unhappy behavior or to maximize happiness behavior [20]. The question is whether the autonomous car would choose to save the majority of lives when a crash is unavoidable or not. Bonnefon et al. [4] conducted online surveys to study whether autonomous vehicles are designed to be utilitarian from the perspective of passengers. The survey questions are about the typical dilemmas named "Trolley Problem". This problem has been argued for nearly half a century in human morality [24] and extended to artificial morality [7, 25]. The Moral Machine [3] provided a public online experimental platform simulating moral dilemmas and mass data about moral decisions. Anderson et al. [2] evaluated the effectiveness of the ethics framework by comparing ethicists' decisions with machines' decisions. The second kind of machine ethics research focuses on the moral decision for specific ethical events. Loreggia et al. [16] stated the importance of moral principles. When moral principles conflict with personal preferences, the principles stand out and become dominant. Guarni [11] conducted works on a neural network of machine ethics, which classified moral events into two categories: acceptable or unacceptable.

Motivated by the latest research works on machine ethics, this paper proposes a general framework to study ethical dilemmas, specifically on the "Trolley Problem". Since dilemmas faced by machines originated from humans and autonomous machines are designed by humans, the moral decisions made by humans and the principal human moral components must be studied. Therefore, this work studies ethical dilemmas from these two perspectives. We collect the decisions from survey participants on the "Trolley Problem" questions in [3]. Specifically, we require the participants to make decisions on the dilemmas and declare their confidence in

EAI-KDD' 22, August 15, 2022, Washington, DC

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

their decisions. We construct consensus-determinacy space based on the questionnaire data to describe the ethical events in the same set of standards. We formulate the moral principle analysis as the matrix factorization problem. In particular, we extract a latent space to describe principal components.

2 PROBLEM FORMULATION

2.1 Psychological Paradigm

During a series of ethical events, participants are required to provide the decisions they made and the confidence they had in them. Participants are assumed not to be experts in ethics. Therefore, they can only answer a limited number of questions related to ethical dilemmas. Let $\mathcal{X} = [\![x_{m,k,n}]\!] \in \mathbb{R}^{M\times K\times N}$ be the survey data tensor, where the first-order represents M users, the second-order represents the K options, and the third-order represents N events. For the user m and the event n, $x_{m,k,n}$ is set as the confidence value showed in the survey when the option k is selected, otherwise, $x_{m,k,n}$ is set as 0. We define the decision matrix $\mathbf{D} \in \mathbb{R}^{M\times N}$ as:

$$\mathbf{D}_{m,n} = \begin{cases} 1, & \text{Cross the lane} \\ -1, & \text{Stay on the lane} \\ 0, & \text{no decision} \end{cases}. \tag{1}$$

Note that in Eq. 1, we take the case of "no decision" into consideration. Some users refuse to make moral decisions for some events in our survey. Unlike most previous works that abandon the "no decision" cases, we try to use its information by treating it as the third option, which is setting it as 0. Based on the survey data tensor $\mathcal{X} = [\![x_{m,k,n}]\!]$, we can also define the confidence matrix $\mathbf{C} \in \mathbb{R}^{M\times N}$, where the element $\mathbf{C}_{m,n}$ is given as follows:

$$\mathbf{C}_{m,n} = \begin{cases} 0, & \text{no decision} \\ \sum\limits_{k} x_{m,k,n}, & \text{otherwise} \end{cases}. \tag{2}$$

$\mathbf{C}_{m,n}$ is set as 0 when the participant makes no decision, keeping a consistency between decision matrix $\mathbf{D}$ and confidence matrix $\mathbf{C}$.

2.2 Consensus-Determinacy Construction

We denote the survey data of event n by an event matrix $\mathbf{E}_n = \mathbf{X}_{::n}$, where $\mathbf{X}_{::n}$ is the frontal slice n of the tensor $\mathcal{X}$. The rows of $\mathbf{E}_n$ represent users and the columns represent options. For a given event matrix $\mathbf{E}$, consensus $\theta(\cdot)$ and determinacy $\psi(\cdot)$ are defined:

$$\begin{aligned} \theta(\mathbf{E}) &= \Big(\frac{\max\limits_{k}\Big\{\sum\limits_{m} e_{m,k}\Big\}}{\sum\limits_{m}\sum\limits_{k} e_{m,k}} - \frac{1}{K}\Big)\Big/\Big(1-\frac{1}{K}\Big) \\ \psi(\mathbf{E}) &= \frac{1}{M}\sum_{m}\sum_{k} e_{m,k} \end{aligned} \tag{3}$$

where M is the total number of users, K is the total number of event options. The value of θ and ψ can be represented by point (θ, ψ) in a two-dimensional space. The horizontal dimension in this spatial metaphor is the consensus dimension, and the vertical dimension is the determinacy. According to Eq. 3, the consensus θ is the extent that there is a dominative option. In other words, it represents the level of agreement between users. Determinacy ψ is determined by the degree of confidence that users have in their decision.

3 ETHICS PRINCIPLE FACTORIZATION

3.1 Ethics Principle Factorization

As mentioned in Section 2, we formulate two matrices $\mathbf{D} \in \mathbb{R}^{M\times N}$ and $\mathbf{C} \in \mathbb{R}^{M\times N}$ according to Eq. 1 and Eq. 2. We propose a novel matrix factorization method to project survey participants and moral events into a common latent space with these two matrices as input. The overflow is visualized in Figure 1.

Each row vector of the decision matrix $\mathbf{D} \in \mathbb{R}^{M\times N}$ represents a participant's moral decisions across all moral events. Each column vector represents the moral decisions of the moral event across all survey participants. Motivated by collaborative filtering (CF) [22], we decompose the decision matrix into latent matrices is as: $\mathbf{D} \approx \mathbf{P}\mathbf{V}^T$, where $\mathbf{P} \in \mathbb{R}^{M\times c}$, $\mathbf{V} \in \mathbb{R}^{N\times c}$. M is the number of users, N is the number of events, and c is the number of latent factors, the number of moral principle components. Then $\mathbf{p}_m$ can be treated as a latent morality representation of the m^{th} participant, and $\mathbf{v}_n$ can be treated as a latent morality representation of n^{th} moral event.

Referring to the confidence matrix $\mathbf{C} \in \mathbb{R}^{M\times N}$, each row vector represents a participant's confidence levels of their moral decisions across all moral events. Each column vector represents the made moral decisions' confidence levels of the event across all survey participants. We can also apply matrix factorization to decompose the confidence matrix into latent matrices as: $\mathbf{C} \approx \mathbf{Q}\mathbf{V}_1^T$, where $\mathbf{Q} \in \mathbb{R}^{M\times c}$, $\mathbf{V}_1 \in \mathbb{R}^{N\times c}$, c is the number of latent factors which is also the number of moral principle components. Then $\mathbf{q}_m$ can be treated as a latent ethical confidence representation of the m^{th} participant, and $\mathbf{v}_{1n}$ can be treated as a latent ethical confidence representation of the n^{th} moral event.

Our proposed model considers not only moral dilemmas' options, but also confidence levels. According to our knowledge, it is the first trial to evaluate moral dilemmas using both the decision and confidence dimensions. We directly use the participant-event matrix as the input matrix. Since each row vector represents a participant's decisions or confidence across all events, this vector can be extended or further represented as an indicator of a person's morality or ethical preference. The moral decision or people's confidence level towards a moral event is represented via each column vector. It can be extended to indicate an event's moral characteristics. Inspired by collective matrix factorization (CMF) [23], we decompose $\mathbf{D}$ and $\mathbf{C}$ jointly to get a shared latent space of the moral events' principal components, which is $\mathbf{V} = \mathbf{V}_1$.

3.2 Loss Function

We aim to minimize the loss between decision matrix $\mathbf{D}$ and the product of $\mathbf{P}\mathbf{V}^T$ in the latent space. Meanwhile, we aim to minimize the loss between confidence matrix $\mathbf{C}$ and the product of $\mathbf{Q}\mathbf{V}^T$ in the latent space. Therefore, our loss function is defined as:

$$L = \min_{\mathbf{P},\mathbf{Q},\mathbf{V}} \frac{\lambda}{2}\left\|\mathbf{D} - \mathbf{P}\mathbf{V}^T\right\|^2 + \frac{1-\lambda}{2}\left\|\mathbf{C} - \mathbf{Q}\mathbf{V}^T\right\|^2 + R(\mathbf{P},\mathbf{Q},\mathbf{V}), \tag{4}$$

where $0 \le \lambda \le 1$, which is a parameter averaging the matrix factorization errors between the two matrices. $\|\cdot\|^2$ denotes the Euclidean distances. $R(\mathbf{P},\mathbf{Q},\mathbf{V})$ denotes the regularization function to avoid $\mathbf{P}$, $\mathbf{Q}$, and $\mathbf{V}$ being over-complex. All three terms in Eq. 4 are differential for the following gradient descent during

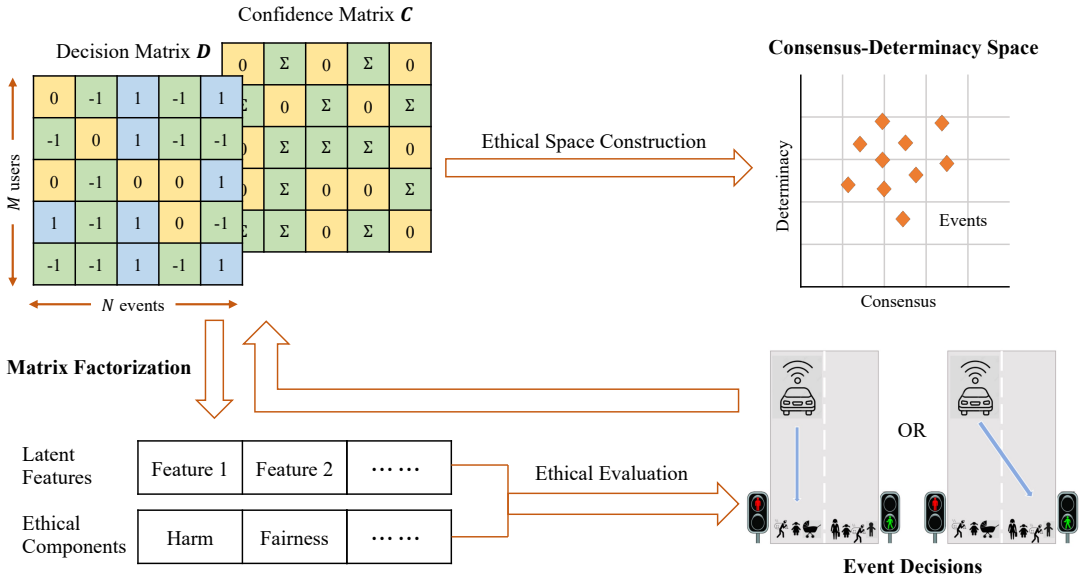

Figure 1: Overview of Our Proposed Ethics Framework

the optimization process. The regularization function $R(\mathbf{P}, \mathbf{Q}, \mathbf{V}) = \frac{\lambda_1}{2}\|\mathbf{P}\|^2 + \frac{\lambda_2}{2}\|\mathbf{Q}\|^2 + \frac{\lambda_3}{2}\|\mathbf{V}^T\|^2$, where $\lambda_1, \lambda_2, \lambda_3 \geq 0$, controlling the terms in the regularization function. In this paper, we equally assign 0.33 to λ_1, λ_2, and λ_3. The optimization problem in Eq. 4 is an unconstrained nonconvex optimization problem with three matrices $\mathbf{P}$, $\mathbf{Q}$, and $\mathbf{V}$. We differentiate the loss with respect to each matrix as follows:

$$\nabla_{\mathbf{P}_m} L = -\lambda \sum_{n=1}^{N} (\mathbf{D}_{m,n} - \mathbf{P}_m \mathbf{V}_n^T)\mathbf{V}_n^T + \lambda_1 \mathbf{P}_m$$

$$\nabla_{\mathbf{V}_n} L = -\lambda \sum_{m=1}^{M} (\mathbf{D}_{m,n} - \mathbf{P}_m \mathbf{V}_n^T)\mathbf{P}_m - (1-\lambda) \sum_{m=1}^{M} (\mathbf{C}_{m,n} - \mathbf{Q}_m \mathbf{V}_n^T)\mathbf{Q}_m + \lambda_3 \mathbf{V}_n^T$$

$$\nabla_{\mathbf{Q}_m} L = -(1-\lambda) \sum_{n=1}^{N} (\mathbf{C}_{m,n} - \mathbf{Q}_m \mathbf{V}_n^T)\mathbf{V}_n + \lambda_2 \mathbf{Q}_m$$

4 EXPERIMENTS

4.1 Psychological Experiments

4.1.1 Moral Dilemmas. Guidelines for autonomous vehicles should take into account not only engineers' and ethicists' morality, but also the public's [4]. The design of ethical problems is quite challenging, as these problems are highly diverse. Some people prefer to save more lives than the fewer [4], while others prefer to save the young than the elderly [15].

Motivated by [3], the dilemmas in our paper also consider nine factors. Nine factors are included in the analysis: humanity preference (humans vs. pets), action preference (stay in lane vs. cross the lane), passenger preference (passengers vs. pedestrians), utilitarianism (more lives vs. fewer lives), gender preference (men vs. women), age preference (elderly vs. young), law and order preference (cross legally vs. illegally), and property preference (machine vs. human). Considering all factors, we designed ten dilemma problems to test the nine moral principle preferences. People are shown only two possible actions when unavoidable accidents.

4.1.2 Participants. We recruited 40 ethnically diverse participants: 15 females and 25 males, and collected participants' ethically related information to eliminate personal ethical discrimination. Moreover, the survey participants were not notified beforehand. All responses and opinions are immediate and real-time. We collect forty effective questionnaires, and each questionnaire has ten moral dilemmas.

4.1.3 Questionnaire. In our paper, we want to test the significance order of the nine moral principle components affecting the public morality of autonomous vehicles. There are two datasets: the decision and confidence data for further analysis. Each dataset is based on 40 individuals' attitudes towards ten autonomous vehicle dilemmas. Two questions were formulated for each dilemma concerning people's decision and confidence in their choices: 'Stay on the lane or Cross the lane?', 'What level of confidence do you have in your decision?' Answers were rated at five levels from 0 (Not at all confident) to 100% (Very confident).

4.2 Matrix Factorization Results

4.2.1 Moral Principle Factorization.

Impact of Parameter. In the first experiment, we study the impact of parameters in our factorization loss function (Eq. 4). The parameter sensitivity, the number of components in the ethics latent space, on the loss value is studied in Figure 2. This parameter is tested from 1 to 20. According to the presented result, we can find that our ethics principles factorization model performs well when the number of components falls in the range [5, 10]. That is to say, five principal components or more may have good explanations of the public morality of dilemmas in our paper.

Explainable Moral Principle Components. The critical issue of our ethics principle factorization is extracting a latent space to describe moral principle components. According to Figure 2, we denote the

number of principal components as five. Motivated by [12, 13], we define the first five ethics principle components as: Harm, Fairness, Loyalty, Respect, and Purity. Harm takes the first position over the other four ethics domain because people are sensitive to suffering. In Figure 2, the loss value decreases significantly with only one principal component: 'Harm', by around 41%. According to Holstein [13], there are eight ethical concerns regarding autonomous vehicles, namely safety, security, privacy, and so on. The first two concerns are "how much damage autonomous vehicles would cause?". Thus, only 'Harm,' can decrease loss value significantly. Adding the 'Fairness', the loss value keeps decreasing by around 38%. If we consider all five ethics domains, the loss value decreases sharply. Continuing to decompose additional fifteen moral principle components does not significantly affect the loss value reduction.

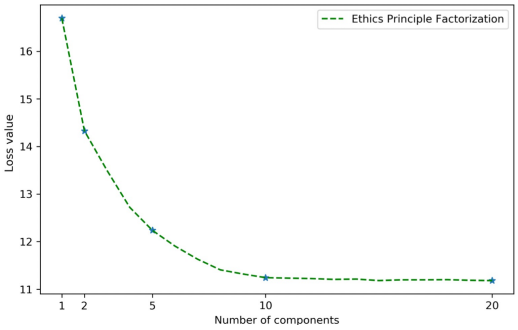

Figure 2: Loss Value Varying the Number of Components

4.2.2 *Moral Choices Inference.*

The Effects of Morality on Moral Justification. The cosine similarity between the latent vectors extracted by our ethics principle factorization model calculates a person's morality. The moral justification similarity calculation method is similar, using the cosine similarity between the moral justification vectors: decision vectors and confidence vectors. As shown in Figure 3, persons with more similar morality tend to make more similar justifications. The correlation between morality and moral justification is measured by the Pearson Coefficient $r = 0.8835$, indicating high linearity.

Inference of Public Implicit Morality of Dilemmas. Referring to our two collected datasets: **D** and **C**, both have missing values (*Nan*). Among the 40 participants, ten of them refuse to make decisions on some events. In the decision matrix, there are 29 missing values. While for the confidence data, six participants did not state their confidence in some decisions they made. There are 21 missing values in the confidence matrix. Although people do not make decisions or show confidence, the missing values still have implicit moral meanings.

This experiment is to predict the missing morality values based on the user latent factors' similarities. As shown in Figure 3, moral justification is highly correlated with personal morality. Using neighbors of the participant to predict implicit morality is reasonable. In this paper, since we want to test the public morality of dilemmas, all other users' decisions on the specific dilemmas are involved in predicting the missing moral decision value, introduced by the user similarity as weights of missing value prediction.

After predicting and filling missing values, we have two new decision and confidence matrices without *Nan* values. Then we reuse our ethics principle factorization model to extract new latent factors representing user morality, moral characteristics of dilemmas, and principal components involved in moral justification. The parameter sensitivity of loss value on the usage of processed morality matrix is similar to Figure 2. Collectively factorizing matrices without missing values has a relatively lower loss value during the training algorithm.

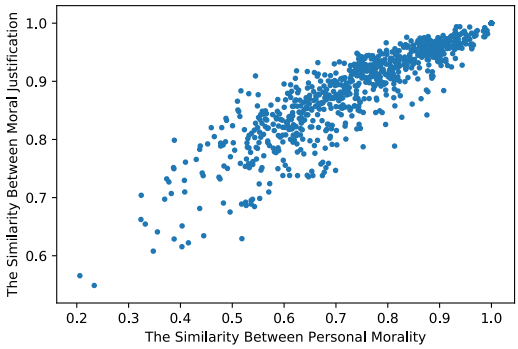

Figure 3: Impact of Personal Morality on Moral Justification

Predict Public Morality of Unseen Moral Events. Our proposed *Ethics Principle Factorization* model extracts the latent vectors to represent the moral characteristics of dilemmas. As shown in Figure 2, five principal moral components explain most considerations of the public in justifying dilemmas. For the unseen new moral event, based on its violation degree of the five ethics domain: Harm, Fairness, Loyalty, Respect, and Purity, we can represent the event as a five-dimensional latent vector. Further, we can predict public morality based on the similarity between latent event vectors of the dilemmas which have already been justified by people and the unseen ones without any former moral justification.

4.3 Consensus-Determinacy Space

The consensus and determinacy distribution of the ten events is shown in Figure 4. The positions of the consensus and determinacy converge from blue circles to red squares. Significant increases in the consensus value are found for all of the ten events. It indicates that our proposed algorithm can discover the common implicit trend underlying the diverse feedback of all users. Besides the consensus, the determinacy value also shows an increasing trend. It also suggests a more appropriate moral event measurement behind the noisy survey data, as the higher confidence is found via our moral learning method.

There are also some interesting observations based on the individual event. For example, event 8 is to choose between three passengers including a baby and two homeless persons. Event 8 shows a shallow consensus (0.2734) value, but its determinacy is relatively high (0.5161). Shallow consensus and high determinacy imply a dangerous signal for the event. A low consensus value indicates a significant difference between the opinions of different users. Meanwhile, relatively high determinacy indicates that both sides are very confident in their opinions. The design of AI systems should avoid this event since it can lead to extreme conflict.

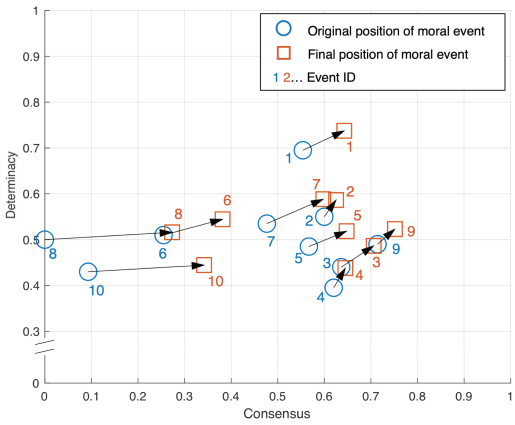

Figure 4: The Consensus-Determinacy space and the distribution of ten moral events.

However, event 1 is a much better case, whose consensus (0.6428) and determinacy (0.7375) are relatively high. It is safe to let machines make decisions for this type of case. Event 5 also shows a consensus value (0.6479) similar to event 1, but with its lower determinacy value. Considering that the option "cross the lane" in event 1 does not cause life loss; however, the same option in event 5 causes at least three life losses. This difference between the determinacy of events 1 and 5 is intuitive. It shows that people would not be that confident when their selection causes death. Although most participants agree that "crossing the lane" can reduce the life loss in events 1 and 5, death still makes them less confident. The coordinate of event 4 also evidences the decrease of determinacy when the better option causes loss of life.

In summary, the consensus dimension indicates the possibility of providing the decision when facing a dilemma. If a dilemma has a high consensus, the machine can decide according to the preference of human beings. The most dangerous situation is the upper left area of the Consensus-Determinancy space, where the consensus is low, but at the same time, the determinacy is high. Generally, we have to avoid that kind of design in an intelligent system.

5 CONCLUSION

This paper does not intend to provide the so-called standard answer to any ethical dilemma, which is still an unresolved problem in human society. The direction of the effort is to explore machine ethics in practice when the world has been full of intelligence and autonomy. In cases with no consensus on the most ethical way to act, the machine should not be allowed to act autonomously. However, not acting does not imply the moral conundrum is avoided because the decision not to act also has a moral dimension. So, in this paper, we try to propose a general framework to study the ethical dilemmas by deep research of one typical dilemma that needs urgently with the fast growth of autonomous vehicles. Specifically, we would like to emphasize the following points:

- Consensus-determinacy space unifies different ethical events in the same set of standards, which benefits the comparison and awareness of ethical events. Autonomous machines cannot only be affected by their designers' moralities. The human morality consensus must be heard.
- Combing the latent features with social science findings, we can find that harm and fairness are the two ethical principal components. Harm is related to the ordered loss across different groups (e.g., people, animals, self, property, etc.) [25]. Fairness is across human individuals in various demographic groups (e.g., by gender, race, age, etc.) [14].

6 ACKNOWLEDGMENTS

This work was funded by Innovation and Technology Fund (ITS/110/19) from the Innovation and Technology Commission of Hong Kong.

REFERENCES

[1] Michael Anderson and Susan Leigh Anderson. 2011. *Machine ethics.* Cambridge University Press.
[2] Michael Anderson and Susan Leigh Anderson. 2014. GenEth: a general ethical dilemma analyzer. In *Twenty-Eighth AAAI Conference on Artificial Intelligence.*
[3] 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.
[4] Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. 2016. The social dilemma of autonomous vehicles. *Science* 352, 6293 (2016), 1573–1576.
[5] Nick Bostrom and Eliezer Yudkowsky. 2014. The ethics of artificial intelligence. *The Cambridge handbook of artificial intelligence* 1 (2014), 316–334.
[6] Joan Casas-Roma, Jordi Conesa, and Santi Caballé. 2021. Education, ethical dilemmas and AI: from ethical design to artificial morality. In *HCII.* 167–182.
[7] Vincent Conitzer, Walter Sinnott-Armstrong, Jana Schaich Borg, Yuan Deng, and Max Kramer. 2017. Moral decision making frameworks for artificial intelligence. In *AAAI.*
[8] Boer Deng. 2015. The robot's dilemma. *Nature* 523, 7558 (2015), 24.
[9] Clàudia Figueras, Harko Verhagen, and Teresa Cerratto Pargman. 2021. Trustworthy AI for the People?. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society.* 269–270.
[10] Keith Frankish and William M Ramsey. 2014. *The Cambridge handbook of artificial intelligence.* Cambridge University Press.
[11] Marcello Guarini. 2006. Particularism and the classification and reclassification of moral cases. *IEEE Intelligent Systems* 21, 4 (2006), 22–28.
[12] Jonathan Haidt and Jesse Graham. 2007. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research* 20, 1 (2007), 98–116.
[13] Tobias Holstein, Gordana Dodig-Crnkovic, and Patrizio Pelliccione. 2018. Ethical and Social Aspects of Self-Driving Cars. *arXiv preprint arXiv:1802.04103* (2018).
[14] Rashidul Islam, Shimei Pan, and James R Foulds. 2021. Can We Obtain Fairness For Free?. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society.* 586–596.
[15] Olof Johansson-Stenman and Peter Martinsson. 2008. Are some lives more valuable? An ethical preferences approach. *AJHE* 27, 3 (2008), 739–752.
[16] Andrea Loreggia, Nicholas Mattei, Francesca Rossi, and K Brent Venable. 2018. Preferences and ethical principles in decision making. In *AAAI.*
[17] James H Moor. 2006. The nature, importance, and difficulty of machine ethics. *IEEE intelligent systems* 21, 4 (2006), 18–21.
[18] Robin Murphy and David D Woods. 2009. Beyond Asimov: the three laws of responsible robotics. *IEEE Intelligent Systems* 24, 4 (2009), 14–20.
[19] Gina Neff and Peter Nagy. 2016. Automation, algorithms, and politics| talking to Bots: Symbiotic agency and the case of Tay. *International Journal of Communication* 10 (2016), 17.
[20] Frederick Rosen. 2005. *Classical utilitarianism from Hume to Mill.* Routledge.
[21] Stuart Russell. 2016. Should We Fear Supersmart Robots? *Scientific American* 314, 6 (2016), 58–59.
[22] Badrul Munir Sarwar, George Karypis, Joseph A Konstan, John Riedl, et al. 2001. Item-based collaborative filtering recommendation algorithms. *WWW* (2001).
[23] Ajit P Singh and Geoffrey J Gordon. 2008. Relational learning via collective matrix factorization. In *KDD.* ACM, 650–658.
[24] Judith Jarvis Thomson. 1976. Killing, letting die, and the trolley problem. *The Monist* 59, 2 (1976), 204–217.
[25] Suzanne Tolmeijer, Markus Kneer, Cristina Sarasua, Markus Christen, and Abraham Bernstein. 2020. Implementations in machine ethics: a survey. *ACM Computing Surveys (CSUR)* 53, 6 (2020), 1–38.
[26] Steve Torrance. 2005. A robust view of machine ethics. In *Technical reportmachine ethics: papers from the AAAI fall symposium; FS0506, American Association of Artificial Intelligence, Menlo Park, CATuring A (1950) Computing machinery and intelligence. Mind,* Vol. 59. 433460.
[27] Han Yu, Zhiqi Shen, Chunyan Miao, Cyril Leung, Victor R Lesser, and Qiang Yang. 2018. Building Ethics into Artificial Intelligence. In *IJCAI.*