Robin Chan, Radin Dardashti, Meike Osinski, Matthias
Rottmann, Dominik Brüggemann, Cilia Rücker, Peter Schlicht,
Fabian Hüger, Nikol Rummel und Hanno Gottschalk

What Should AI See? Using the Public's Opinion to Determine the Perception of an AI

June 21, 2022

http://www.imacm.uni-wuppertal.de

What Should AI See?

Using the Public's Opinion to Determine the Perception of an AI

Robin Chan[1], **Radin Dardashti**[2], **Meike Osinski**[3], **Matthias Rottmann**[1], **Dominik Brüggemann**[4], **Cilia Rücker**[1], **Peter Schlicht**[5], **Fabian Hüger**[6], **Nikol Rummel**[3,7], and **Hanno Gottschalk**[1]

1: IZMD, Faculty of Mathematics and Natural Sciences, University of Wuppertal, Germany
2: IZWT, School of Humanities, University of Wuppertal, Germany
3: Institute of Educational Research, Ruhr-University Bochum, Germany
4: IZMD, Chair of Reliability Engineering and Risk Analytics, University of Wuppertal, Germany
5: CARIAD SE, Wolfsburg, Germany
6: Volkswagen AG, Wolfsburg, Germany
7: Center of Advanced Internet Studies, Bochum, Germany

Abstract

Deep neural networks (DNN) have made impressive progress in the interpretation of image data, so that it is conceivable and to some degree realistic to use them in safety critical applications like automated driving. From an ethical standpoint, the AI algorithm should take into account the vulnerability of objects or subjects on the street that ranges from "not at all", e.g. the road itself, to "high vulnerability" of pedestrians. One way to take this into account is to define the cost of confusion of one semantic category with another and use cost-based decision rules for the interpretation of probabilities, which are the output of DNNs. However, it is an open problem how to define the cost structure, who should be in charge to do that, and thereby define what AI-algorithms will actually "see". As one possible answer, we follow a participatory approach and set up an online survey to ask the public to define the cost structure. We present the survey design and the data acquired along with an evaluation that also distinguishes between perspective (car passenger vs. external traffic participant) and gender. Using simulation based F-tests, we find highly significant differences between the groups. These differences have consequences on the reliable detection of pedestrians in a safety critical distance to the self-driving car. We discuss the ethical problems that are related to this approach and also discuss the problems emerging from human–machine interaction through the survey from a psychological point of view. Finally, we include comments from industry leaders in the field of AI safety on the applicability of survey based elements in the design of AI functionalities in automated driving.

Index Terms

AI-based perception • computer vision • automated driving • safety aware interpretation of probabilities • cost-based decision rules • participatory approach to ethics by design • practical ethics • human machine interaction

I. INTRODUCTION

When human beings and robots interact in public space, robots should take the ethical values of humans into account. As robots are programmed, this requires anticipation of situations where robots could be in conflict with those values and also anticipation of the decision space for such a situation. Ultimately, an algorithmic approach to find and execute a compliant decision is needed. This particularly applies to the case of automated driving, where wrong decisions of the robotic car could be harmful or even deadly to humans. Approaching what has been said above, software engineers encounter various problems:

1) In the stage of programming, software engineers might not be aware of the consequences that certain technical settings imply. Ethical problems might just be disregarded. In light of existing *safety by design* procedures in industry, this could be seen as an organizational task to enable software designers to develop *ethics by design* [1].

2) There is no unique ethical system the software developer could refer to. Neither the existing normative theories of ethics need to come to the same decision [2]–[4] nor is it clear, *who* should apply any of these ethical

systems: the software engineer, the software company, a philosopher with a PhD degree in ethics, bodies of technical standardization, politicians, judges, or the public? Also the relevance of representation [5] and regional variations in ethical decision making [6] might play a central role.

3) Whoever makes the decision requires an understanding of the technical matters that determine that decision. What communication strategies could enable e.g. citizens to come to a qualified decision?

4) The perception of a robot and its representation of the world fundamentally differs from human perception. Due to this circumstance, the application of ethical systems developed by and for humans is not straightforward. For instance, the technical perception of a robotic car is usually based on deep neural networks (DNN) that interpret various sensors based on probabilities [7].

Concerning this last point, from a technical point of view, DNNs in computer vision output probabilities given high resolution images. Coming to conclusions on the basis of probabilities itself involves ethical decisions, since the consequences of e.g. the confusion of a pedestrian with the street could be deadly, whereas the opposite confusion would trigger an unnecessary emergency braking. The perception of the street scene derived from the predicted probabilities therefore changes if the cost of confusion is determined in different ways [7], [8]. In particular, different configuration of confusion costs that prioritize the safety of the passengers of the self-driving car over the safety of other external road users may lead to significant changes in perception.

Just leaving confusion costs aside and just choosing the class with the highest probability using the maximum a posteriori probability principle, also known as the Bayes decision rule, does not avoid the ethical problem, since this amounts to costs of confusion treating any type of confusion the same. Such a cost structure deviates from common ethical intuition. Nevertheless, this "robotistic" decision rule is standard in present artificial intelligence (AI) technologies, which gives a good illustration to the first point of disregarded ethical problems mentioned above. In particular, this problem is not explicitly mentioned in any of the recommendations of ethics commissions in autonomous driving on German [9], European [10], or international level [11], [12].

Elaborating further on [7], where the ethical problem of decision rules is described and the decision space is presented, in this article we discuss a participatory approach to determine costs of confusions. Between June 2019 and November 2020, participants in our online survey[1] entered 5045 proposals as to how severe confusions of random instances adhering to one semantic category (e.g. human) with other semantic categories (e.g. road) should be judged. The presented instances are extracted from the popular computer vision dataset called Cityscapes [13], which is commonly used to train an AI for the perception of street scenes. With our survey, we present an approach to use the public's opinion to determine what an AI system will see. In this article, we experiment with this specific solution strategy to problem point 2) and we also position this approach in the field of the ethics of AI, where participatory approaches are recognized as a crucial ingredient for developing trustworthy technologies.

As a further contribution to the second point, we use personal information provided by the survey participants to conduct significance tests for the difference of confusion costs between groups, similar to [6]. More precisely, we employ statistical tests to compare the variation within groups to the variation between groups by means of simulation based F-tests that are tailored to the data structure of confusion costs. The different confusion cost structures lead to differences in the consequences with regard to human instances in safety critical zones in front of a self-driving car. These are evaluated using software developed in [14]. This allows us to explore the extent to which the judgment of different groups of people and also the "robotistic" view can potentially lead to consequences with ethical significance. In general, however, direct statistical assessment of safety of AI systems in automated driving is known to be difficult in practice [15].

Concerning point 3), we put the online survey in the perspective of current research on human-machine-interaction (HMI) by evaluating the feedback of the survey participants. We also involve industry leaders who give a comment on how our conducted participatory approach can contribute to an actual industrial solution.

Our article does not intend to deliver any solution to the problem of defining confusion costs, i.e. to determine the perception of an AI applied in automated driving. In particular, we do not recommend or disrecommend to use the confusion cost matrices obtained from our survey in automated driving. Rather we construct and confront interdisciplinary perspectives on a problem of practical ethics, which emerges from the application of AI in safety critical applications, and investigate potential tensions as well as concordance between those. It should also be clear

---

[1]**Visit** http://www-ai.math.uni-wuppertal.de/~chan/poll/ or http://uni-w.de/1hx for the online survey
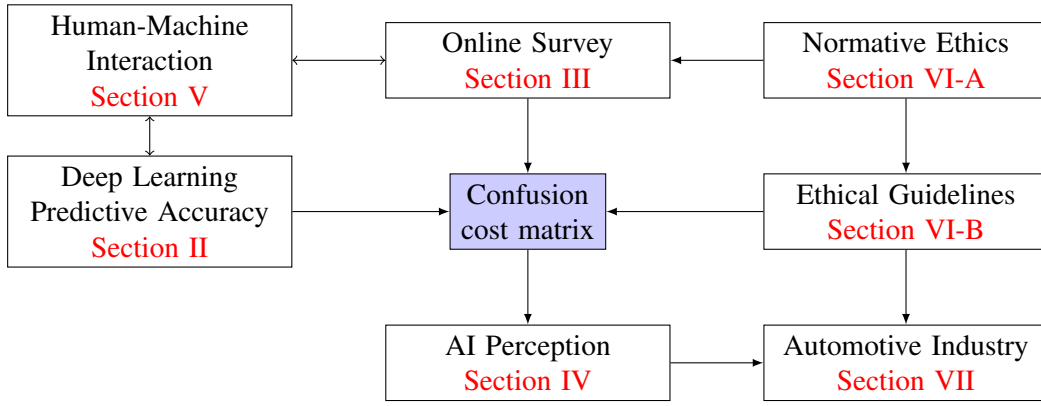
Fig. 1: The perception of an artificial intelligent (AI) system depends on a *confusion cost matrix*, which is in turn determined by various factors. The sections, in which we will discuss the different factors, are given in this figure.

that in this work we only investigate one AI perception modality, namely camera based semantic segmentation. Autonomous vehicles, however, are complex systems with various levels of redundancy within and across sensor modalities, including camera, radar, lidar, and ultrasonic detectors. It is beyond the scope of this work to evaluate the consequences of different choices of the cost structure of confusions on the level of such complex systems.

We provide an overview of the structure of this article in Fig. 1. This work is organized as follows: In Section II we give a mathematical definition of cost-based decision rules and thereby define the space of alternatives. In the subsequent Section III we present our survey-based approach and the survey design. Section IV presents statistical evidence for the difference between certain groups of people as well as the *robotistic* view, and also evaluates to what extent these differences are safety-relevant. Section V considers the relevant HMI aspects of the survey and explores the difficulties that emerge from just asking the people from the people's perspective. Section VI places this article's approach in the field of ethical discussions on the application of AI, with special emphasis on automated driving. Similarities and Dissimilarities to the infamous trolley problem are discussed and the applicability of existing ethical guidelines are probed. Section VII discusses the usefulness and the feasibility of participation based approaches from an industrial point of view and relates such approaches to ethical, legal and technical regulatory frameworks.

II. COST-BASED DECISION RULES IN DEEP LEARNING FOR COMPUTER VISION

The introduction of deep learning, a sub field of machine learning, has enabled advances in many applications of artificial intelligence (AI) that have been considered intractable before, such as computer vision. Computer vision (CV) can be described as the task that deals with enabling machines to gain an high-level understanding of scenes from digital image data. This includes the detection and localization of objects by means of images captured by cameras, which essentially determines what an AI can "see". The nowadays established CV pipelines are all based on deep learning and organized around deep neural networks (DNNs). Employing such a type of model contains several gateways by which ethical difficulties may enter. This includes for instance the distribution of object classes in the training data, class weights in the training objectives, or the decision rules incorporated in DNNs to obtain final class predictions. In this section, we elaborate on how ethical problems enter through the latter gateway relating to the concept of decision rules.

A. Mathematical Formulation of Cost-based Decision Rules and Connection to Standard Decision Principle

From a technical point of view, DNNs are typically used as statistical models to estimate a probability that a given input belongs to a certain object category. More formally, let us denote the input by x. In computer vision this could e.g. be an entire image or even only a single pixel of an image, depending on the type of problem. Further, let us denote the object category corresponding to the input by y. Such a class label is usually encoded with a natural number, e.g. $y \in \{1, \ldots, N\}$. Here, $N \in \mathbb{N}$ denotes the number of different classes, which is a design parameter chosen before creating datasets, and thus also before training DNNs to recognize objects in images. In this regard,
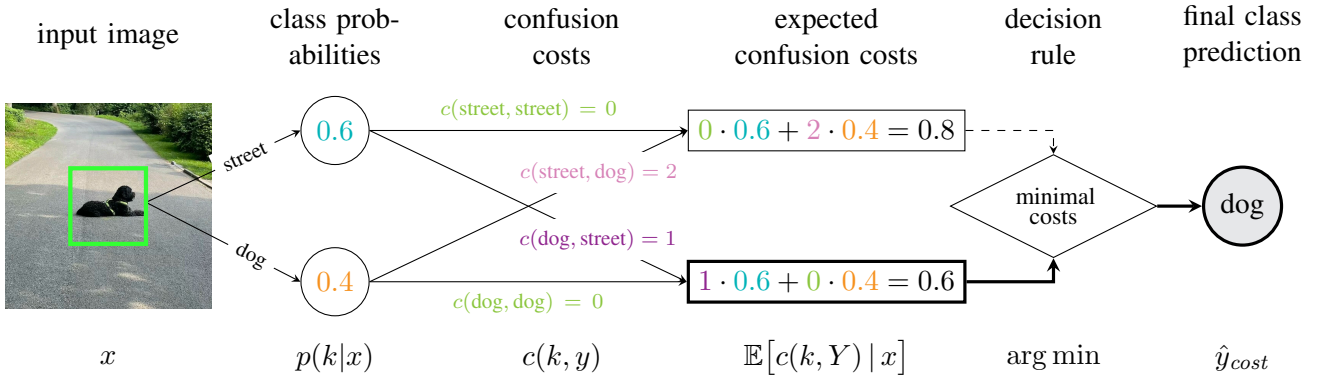
Fig. 2: Illustration of a simplified example of a cost-based decision rule in binary classification. Here, the task is to classify the content within the green box in the input image either as *street* (class 1) or *dog* (class 2). In this particular example it is assumed that confusing the class dog with the class street is twice as severe as the other way round, which can be realized by the confusion costs $c(\text{street}, \text{dog}) = 2$ and $c(\text{dog}, \text{street}) = 1$. Then, a cost-based decision rule selects the class that has the lowest expected confusion costs. This results in the class dog as final class prediction although the class probabilities are higher for street (60% vs. 40%). We refer to Eq. (2) for the general formula for the final class prediction by means of a cost-based decision rule.

we train DNNs to estimate $p(y|x)$, which can be understood as the probability for a given input x having class affiliation y. Given the probabilistic output of DNNs over all potential classes, the final class prediction is then usually obtained by selecting the class that is assigned the highest probability. This approach is also commonly referred to as *maximum a-posteriori probability* (MAP) principle or *Bayes decision rule* [8], [16], which yields the final class prediction

$$\hat{y}_{Bayes}(x) := \underset{k \in \{1,...,N\}}{\arg\max} \; p(k|x) \; . \tag{1}$$

From decision theory, the Bayes decision rule is known to be merely one example of the more general concept of *cost-based decision rules*. The latter decision principle selects the class that is associated with the lowest expected costs with respect to a classification mistake. To this end, a quantification of the costs of confusion between classes is required. In more detail, given the cost of confusion $c(k,y) \in \mathbb{R}_{\geq 0}$ between the two classes $k, y \in \{1, \ldots, N\}$, the final class prediction for the input x via a cost-based decision rule is then obtained by

$$\hat{y}_{cost}(x|c) := \underset{k \in \{1,...,N\}}{\arg\min} \; \mathbb{E}\big[c(k,Y)\,|\,x\big] = \underset{k \in \{1,...,N\}}{\arg\min} \sum_{y=1}^{N} c(k,y)\, p(y|x)\,, \tag{2}$$

where $\mathbb{E}[c(k,Y)\,|\,x]$ denotes the expected costs of confusion with respect to class k given input x, or, in other words, the expected costs for confusing the considered class k with any other possible class $y \in \{1, \ldots, N\}$. Note that the cost assignments can also be expressed compactly in form of a *confusion cost matrix* $(c(k,y))_{k,y \in \{1,...,N\}}$ of size $N \times N$, which will be the subject of discussion in the following sections. For a simplified example illustrating the process of decision making by means of cost-based decision rules, we refer to Fig. 2. At this point, we want to emphasize that the final class prediction via cost-based decision rules is not affected by the absolute values but by the relative differences of confusion costs. More formally, the cost-based decision remains unaffected by a constant non-negative factor $\lambda \in \mathbb{R}_{\geq 0}$ applied to the confusion costs, i.e.

$$\hat{y}_{cost}(x|c) = \underset{k \in \{1,...,N\}}{\arg\min} \; \mathbb{E}\big[c(k,Y)\,|\,x\big] = \underset{k \in \{1,...,N\}}{\arg\min} \; \lambda\, \mathbb{E}\big[c(k,Y)\,|\,x\big] = \underset{k \in \{1,...,N\}}{\arg\min} \; \mathbb{E}\big[\lambda\, c(k,Y)\,|\,x\big] \; . \tag{3}$$

Still, a question that naturally arises in this context is how to choose the quantities $c(k,y) \; \forall \; k, y \in \{1, \ldots, N\}$. For example in scenarios of automated driving, for the two classes "street" and "human", how should the cost of the confusion be valuated? Moreover, should the cost of confusion between these two classes be symmetric, e.g. is overlooking humans in favor of the street as severe as the other way round? Common human intuition would

suggest that confusion costs should be different depending on the type of confusion. However, it remains an open question, what values should explicitly be used.

As a matter of fact, however, the confusion costs are already implicitly defined in DNNs when employing the standard Bayes decision rule. Returning now to the mentioned statement that the Bayes decision rule is merely one example of a cost-based decision rule, it turns out that this *standard decision principle incorporates constant confusion costs, weighting each type of confusion equally serious*. More precisely, for each type of confusion between two classes k and y, the Bayes decision is based on the cost valuation

$$c_{robot}(k, y) = \begin{cases} 0 & \text{if } k = y \\ 1 & \text{if } k \neq y \end{cases} \quad \forall \; k, y \in \{1, \ldots, N\} \; , \tag{4}$$

resulting in the connection between Bayes decision rule and the latter cost function

$$\hat{y}_{Bayes}(x) \overset{(1)}{=} \underset{k \in \{1,\ldots,N\}}{\arg\max} \; p(k|x) \overset{(4)}{=} \underset{k \in \{1,\ldots,N\}}{\arg\min} \; \mathbb{E}\big[c_{robot}(k, Y) \,|\, x\big] \overset{(2)}{=} \hat{y}_{cost}(x|c_{robot}) \; , \tag{5}$$

cf. also [7], [16]. In this way, the Bayes decision principle aims at *minimizing the chance of any incorrect prediction*, since there is no distinction in the type of error according to Eq. (4), which is equivalent to *maximizing the model's predictive accuracy*. We refer to this kind of decision making as the *robotistic* attitude. Furthermore, as revealed in [7], the Bayes decision rule leads to the ethical dilemma of machine learning models in general that, on the one hand, it is not evident what confusion cost values should explicitly be used, but on the other hand, confusion cost values in conflict with common human intuition are already set implicitly by default. In this work, we study different confusion cost valuations in the context of street scenes, with the costs determined by the public in an online survey. While doing so, we study qualitative differences in the obtained cost matrices for different groups of survey participants.

B. Values and Tradeoffs in Cost-based Decision Rules

Value judgments are ubiquitous in science and technology and have been a main issue in discussions on the ethics and philosophy of science and technology (e.g. [17]–[19]). They can enter at various steps of the inquiry and development when different epistemic and non-epistemic goals are in conflict, thereby leading to tradeoff situations. A specific choice within this tradeoff space amounts to a possible entry point for values. Quite frequently, these choices remain implicit or are not even recognized as "choices" and their corresponding tradeoff space remains similarly unknown. These contingencies of the process of inquiry should in morally significant situations be the subject of reflective evaluation [19].

These tradeoff situations have already been recognized in the work of software engineers that use machine learning techniques, see e.g. [20]. Issues regarding the opacity of the AI system and the potential biases present in the dataset from which the algorithm learns, have been central topics of discussion in the ethics of AI literature [21]. Once these potentially ethical issues in the choice of these tradeoff situations have been recognized, a plethora of policy recommendations and ethical principles for more transparency and explainability in the design process have been suggested, see e.g. the ethics guidelines of AI High-level expert group of the European Commission [22] or for an overview [23]. As was shown in [23] there seems to be a global convergence regarding the principles that should underlie the ethical use of AI, while a divergence was recognized regarding specific implementations of these. Thus, a detailed ethical analysis complementing available ethical guidelines is essential as it is explicitly requested in [10, p. 23].

In Section II-A the possible choices for the parameters in the costs of confusion in Eq. (2) provides the tradeoff space within which various epistemic and non-epistemic values need to be assessed and possibly weighed against each other. This can easily be illustrated by a comparison between the recall and the precision in the case of human classification, cf. also Fig. 3. By varying the costs of confusion, one may maximize the recall (or equivalently the sensitivity) to the point where there are no false negatives. This, however, goes hand in hand with a decrease in the precision, as the number of false positives similarly increases with the variation of cost values. This is a generic feature of any non-perfect but realistic classifier. So one may be able to identify all humans correctly, but only at the cost of identifying additionally many non-human things as humans. This would lead to a significant decrease in the practicability of autonomous vehicles, as it would mistakenly hit the brakes too frequently. Here practicability enters as a value, which if not satisfied would significantly diminish the usability of any autonomous vehicle.

(a) computation of the classification performance metrics recall and precision

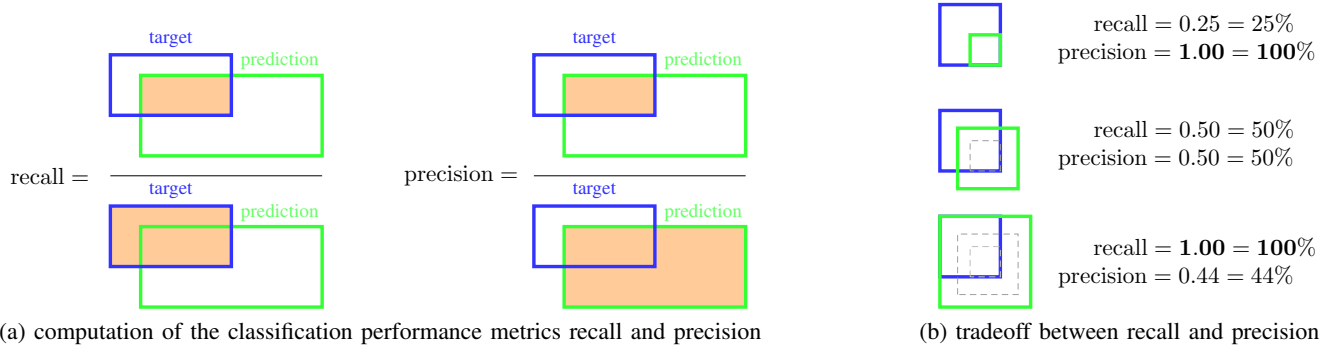(b) tradeoff between recall and precision

Fig. 3: Illustration of the classification performance metrics recall and precision. (a) The recall is the fraction of the amount of overlap between target and prediction divided by the amount of the target, while the precision is the fraction of the amount of overlap between target and prediction divided by the amount of the prediction. (b) The sensitivity of predictions can be adjusted by varying thresholds (or implicitly by varying confusion costs), which subsequently increases the amount of predictions. In this way the recall can be maximized, however to the detriment of decreasing the precision, illustrating the tradeoff between these two classification performance metrics.

In this particular context, an ethical assessment of the outlined tradeoff is non-trivial as there are utilitarian reasons for the introduction of autonomous vehicles which would warrant an increase in practicability (cf. [9], although these are critically discussed [24], [25]), thereby making practicability a requirement, which nevertheless may be in tension with the individual lives that are put at risk by decreasing the recall.

This tradeoff between the recall and precision, or, the identification of humans and the practicability of the autonomous vehicle, respectively, illustrates just one of many values that may impact, if made explicit, the determination of the confusion cost matrix. Other examples would be fuel efficiency and speed but also issues of privacy may impact the tradeoff space. In this paper we are particularly concerned with the tradeoff between accuracy, as implemented by the Bayes decision rule, cf. Eq. (5), and the public's opinion, as implemented in a survey-based approach in determining the cost confusion matrix. We will further elaborate on the tradeoff and consider the guidelines provided by the ethics commission [9].

III. Survey Design for the Valuation of Confusion Costs in Street Scenes

In the context of automated driving, cameras are one of the main components for perceiving the environment of a self-driving car. To this end, AI algorithms are deployed in order to interpret the captured images. They classify objects into predefined classes based on estimated probabilities that an object belongs to a certain class. As described in the previous Section II, the AI is usually programmed to select the class that has the highest probability [8], [16], [26], see also Eq. (1). Hence, all types of confusion of classes are implicitly treated to be equally serious, cf. Eq. (5). However, in terms of safety this decision principle with constant costs of confusion is not necessarily optimal and ethically differs from common human intuition.

For instance, assume the case in which the AI suggests the class "road" and the class "human" for an object of interest with an estimated probability of 51% and 49%, respectively. In this particular case, strictly trusting the AI and choosing *road over human* because of the slightly higher probability could lead to fatal consequences. From a safety point of view, it therefore seems recommendable that confusions are assessed according to their type as well as the occurrence of possible dangers.

In this section, we present the conducted survey, in which we ask the survey participants to provide confusion costs. We aim at examining the ethical attitude of the public in regard of safety-critical confusions in street scenes. Ultimately, the submitted values should help the AI to adapt its perception to human ethical intuition. Starting now with the design of the survey, we distinguish between two possible perspectives to view street scenarios from, either

1) *passenger of the self-driving car*,   2) or *external traffic participant*.

We focus on these two perspectives as putting passenger first vs. putting other road users first has already been subject of intense public debate [27]. One of these two perspectives is randomly assigned to each participant at the

1: drivable ⟨ road

2: nondrivable ⟨ sidewalk, terrain

3: static ⟨ building, wall, fence, pole, vegetation

4: info ⟨ traffic light, traffic sign

5: human ⟨ person, rider

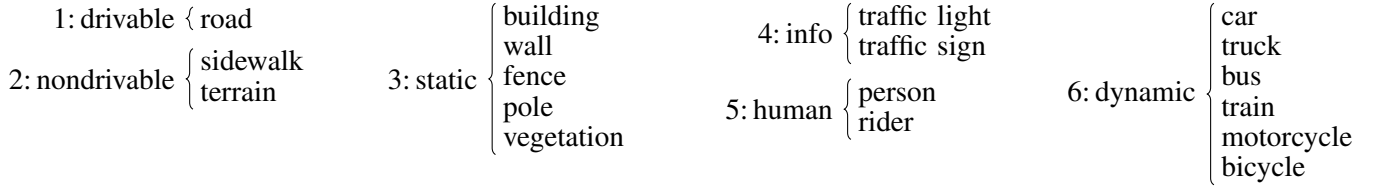6: dynamic ⟨ car, truck, bus, train, motorcycle, bicycle

Fig. 4: Class aggregates of Cityscapes object classes that are used in the survey. Note that the class *sky* is ignored in the survey since the non-detection of the sky does not cause hazardous street scenarios.

Fig. 5: An example of a traffic scene that the survey participants evaluate. Marked with red is the object class, in this case human, for which a possible confusion is assessed.
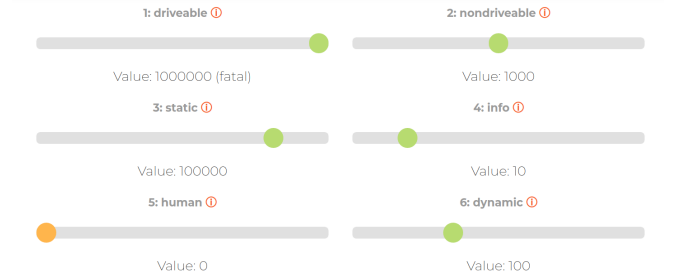

Fig. 6: The survey participants assess confusions in the displayed street scene via sliders with confusion cost values $1 = 10^0, 10 = 10^1$ up to $1M = 10^6$. For the correct class label the cost value is fixed at 0.

start of the survey and last until the survey is stopped. Based on the perspective, potential confusions in different street scenes are evaluated according to the participants' subjective sense of the severity of consequences. The street scenes that are shown during the survey are extracted from the publicly available Cityscapes dataset [13]. Cityscapes is a large collection of images containing diverse images recorded in urban street scenarios in 50 different European cities and it is a popular computer vision benchmark to evaluate how well deep neural networks perform at interpreting complex traffic scenes. There are 19 object classes available in this dataset to be classified. For the sake of simplicity, we reduce the number of classes in the survey by considering class aggregates. We distinguish between six (more superficial) object categories, which are namely 1) driveable, 2) nondriveable, 3) static, 4) info, 5) human, and 6) dynamic, see also Fig. 4 for an overview.

In each of the street scene images displayed in the survey, one instance corresponding to exactly one of the six outlined classes is highlighted, see Fig. 5 for an example. Given this displayed image, the task of the survey participants is to assess a potential confusion with an instance from another class, see Fig. 6 for an illustration of the interface from the survey. With a confusion between car and bus (a confusion between two objects within the same object category) being a reference mistake, the survey participants should assess how severe it is if the highlighted instance is incorrectly assigned to one of the remaining five classes. In other words, the *severity of overlooking* the highlighted object must be valuated. One can choose out of 10 (*nearly harmless*), 100 (*fairly harmless*) up to 1M (*fatal*) times more severe than the reference mistake, or 1 (*marginal*) if the confusion is as serious as the reference mistake. Thus, in total there are seven available levels of severity to choose from. With respect to the same class, naturally no confusion costs should incur, therefore the cost value is fixed to 0. Note that the exponential scale of confusion costs is due to the particular architecture of deep neural networks, which have shown gradual changes in their outputs only with confusion costs of this exponential magnitude, cf. [7].

For a more in-depth analysis of the survey data, the participants are additionally asked to voluntarily provide the following personal information on:

- gender,
- age,
- graduation,
- field of work/study,
- whether they own a driver's license,
- and what they use as main transportation.

$$\bar{C}_P = \begin{pmatrix} 0 & 10^{4.12} & 10^{3.97} & 10^{3.75} & 10^{4.74} & 10^{3.97} \\ 10^{3.90} & 0 & 10^{2.70} & 10^{3.18} & 10^{3.42} & 10^{3.30} \\ 10^{3.34} & 10^{2.50} & 0 & 10^{2.70} & 10^{3.51} & 10^{3.07} \\ 10^{2.96} & 10^{2.96} & 10^{2.76} & 0 & 10^{3.51} & 10^{3.13} \\ 10^{3.72} & 10^{3.00} & 10^{3.05} & 10^{3.41} & 0 & 10^{3.18} \\ 10^{3.41} & 10^{3.17} & 10^{3.05} & 10^{3.41} & 10^{3.18} & 0 \end{pmatrix} \begin{matrix} \text{"drive"} \\ \text{"nondrive"} \\ \text{"static"} \\ \text{"info"} \\ \text{"human"} \\ \text{"dynamic"} \end{matrix}$$

$$\bar{C}_E = \begin{pmatrix} 0 & 10^{4.42} & 10^{4.36} & 10^{4.00} & 10^{5.51} & 10^{4.71} \\ 10^{3.71} & 0 & 10^{2.72} & 10^{3.06} & 10^{3.99} & 10^{3.40} \\ 10^{3.70} & 10^{2.13} & 0 & 10^{2.41} & 10^{3.56} & 10^{3.46} \\ 10^{2.97} & 10^{2.46} & 10^{2.79} & 0 & 10^{3.70} & 10^{3.08} \\ 10^{4.03} & 10^{3.04} & 10^{3.16} & 10^{3.40} & 0 & 10^{3.50} \\ 10^{3.77} & 10^{2.84} & 10^{3.14} & 10^{3.34} & 10^{3.14} & 0 \end{pmatrix}$$

$$\bar{C}_F = \begin{pmatrix} 0 & 10^{4.45} & 10^{4.25} & 10^{4.23} & 10^{5.65} & 10^{4.47} \\ 10^{4.03} & 0 & 10^{2.72} & 10^{3.32} & 10^{4.06} & 10^{3.79} \\ 10^{3.83} & 10^{2.26} & 0 & 10^{2.48} & 10^{3.65} & 10^{3.36} \\ 10^{3.29} & 10^{2.92} & 10^{2.81} & 0 & 10^{3.84} & 10^{3.41} \\ 10^{4.14} & 10^{3.09} & 10^{3.28} & 10^{3.54} & 0 & 10^{3.44} \\ 10^{4.03} & 10^{3.07} & 10^{3.23} & 10^{3.56} & 10^{3.22} & 0 \end{pmatrix}$$

$$\bar{C}_M = \begin{pmatrix} 0 & 10^{4.33} & 10^{4.03} & 10^{3.58} & 10^{4.60} & 10^{4.17} \\ 10^{3.74} & 0 & 10^{2.76} & 10^{2.98} & 10^{3.39} & 10^{3.04} \\ 10^{3.33} & 10^{2.43} & 0 & 10^{2.64} & 10^{3.48} & 10^{3.15} \\ 10^{2.74} & 10^{2.54} & 10^{2.76} & 0 & 10^{3.39} & 10^{2.88} \\ 10^{3.73} & 10^{3.02} & 10^{3.04} & 10^{3.32} & 0 & 10^{3.24} \\ 10^{3.30} & 10^{2.98} & 10^{3.03} & 10^{3.24} & 10^{3.17} & 0 \end{pmatrix}$$

Column labels (left to right): "drivable", "nondrivable", "static", "info", "human", "dynamic". For $\bar{C}_E$: *true target class label in columns*; *prediction in rows*.
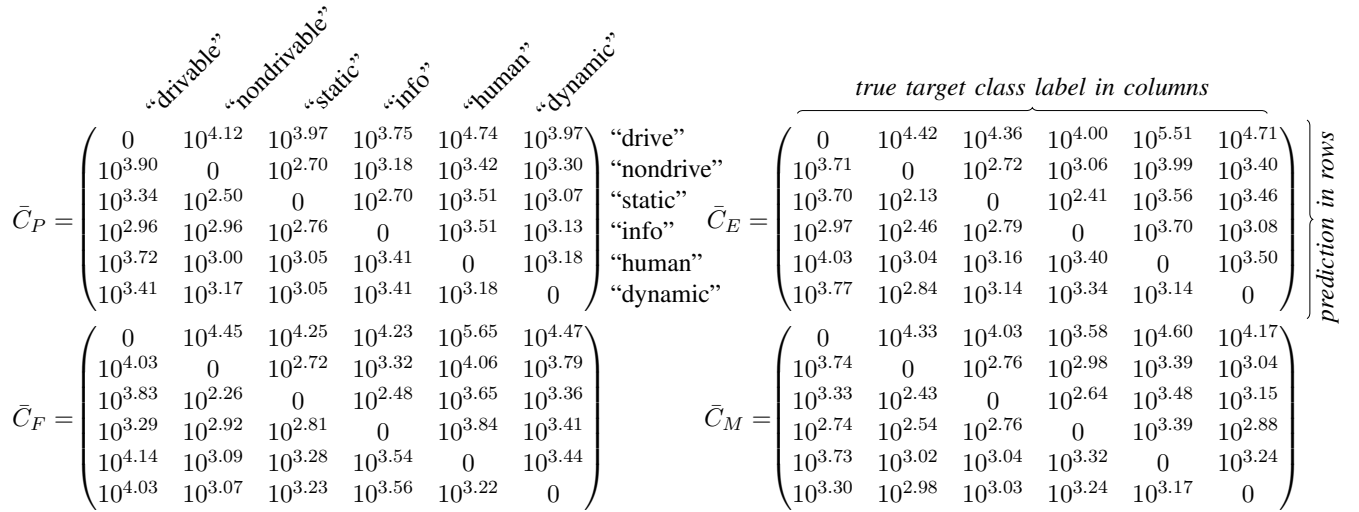
Fig. 7: Average confusion cost matrices determined by different groups formed by the assigned perspective and gender of the survey participants, namely these groups are *passenger* ($\bar{C}_P$), *external* ($\bar{C}_E$), *female* ($\bar{C}_F$), and *male* ($\bar{C}_M$), cf. also Section IV. The matrices are read as follows: e.g. for the fifth entry in the first row, how fatal is the confusion if the model predicts the class *driveable* but the actual true class label is *human*?

Besides the perspective, such meta data can also be used in order to form groups and analyze the collected survey data for statistical relationships, which will be subject of discussion in what follows.

IV. EVALUATION METHODOLOGY FOR CONFUSION COSTS AND NUMERICAL RESULTS

At the time of writing 520 people have participated in the survey, answering 5,045 questions in total. Given this collected data, we inspect for differences in the confusion costs valuation in this section. To this end, we form the groups to investigate by restricting the survey data by means of the assigned perspective and based on the gender of the survey participants. These two characteristics form data subsets of roughly the same size:

- *passenger*: 2744 answers,
- *external*: 2301 answers,
- *female*: 2444 answers,
- *male*: 2523 answers,

besides being sufficiently large to compare them statistically. Note that the groups *passenger* and *external* denote the perspectives corresponding to a passenger of the self-driving car and an external traffic participant, respectively.

A. Comparison of Confusion Cost Values provided by the Survey Data

The average confusion cost values of the outlined groups are displayed in the form of cost matrices in Fig. 7. Comparing the matrices provided by the assigned perspective, i.e. *passenger of the self-driving car vs. external traffic participant*, we observe that the cost valuations are in a similar range. Some noteworthy exceptions are related to predictions of the class "drivable". More precisely, the confusion of either "human" or "dynamic" in favor of "drivable" is valuated considerably more severe by the group *external* (costs of $10^{5.51}$ and $10^{4.71}$ vs. $10^{4.74}$ and $10^{3.97}$, respectively). Focusing now only on the class "human", confusions involving this vulnerable class are generally assigned higher costs by the group *external* as well. The corresponding column in the respective matrices (which represents the submitted cost values when a human instance is highlighted in the image displayed in the survey) let us conclude that by means of the submitted values, the group *external* tends towards a more human sensitive cost assignment than the group *passenger*.

Similar observations can be made when comparing the cost matrices provided by gender, i.e. *female* vs. *male*. Here, the cost values of the group *female* show to be higher in general for most types of confusion. In light of the absolute values, this is most significant for the confusion "driveable" as prediction and "human" as true target class label with cost values $10^{5.65}$ and $10^{4.60}$, respectively. Focusing again only on the class "human", confusions involving this vulnerable class are also assigned clearly higher costs by the group *female*. In fact, with respect to

the class "human", none of the cost valuations provided by the group *male* exceed those of the group *female*. This let us conclude that by means of the submitted values, the group *female* tends towards a more human sensitive cost assignment than the group *male*.

This leads to another finding that is related to the tradeoff between recall and precision. As already introduced in Section II-B, recall can be maximized by increasing the prediction sensitivity, but possibly sacrificing precision. In light of confusion cost matrices, the prediction sensitivity with respect to one class can be varied by the cost values within the corresponding column. This consequently counteracts with the cost values in the row of the considered class in the confusion matrix, which in turn controls the predictive precision. We have already concluded that both the groups *external* and *female* show to have a more human sensitive cost assignment than the groups *passenger* and *male*, respectively. This might be still valid given the submitted cost values and the design of the survey, in which the participants explicitly provided cost values for the severity of overlooking humans, cf. Section III.

However, from a technical point of view, this does not necessarily imply better recall with respect to human classification via cost-based decision rules. In this context, it is crucial to understand that increasing prediction sensitivity always refers to both increasing *and* decreasing confusion cost values. More precisely, the relative difference is critical according to Eq. (3). Otherwise, confusions are valuated symmetrically as in the robotistic cost valuation, cf. Eq. (4). Having now another look at the average cost matrices from the survey data in Fig. 7, we realize that with respect to the class "human" the groups *external* and *female* have higher costs in the column as well as in the row compared to *passenger* and *male*, respectively. Whether such confusion costs valuation still translates to more human sensitive AI perception will be subject of the qualitative as well as consequential analysis in Section IV-C and Section IV-D, respectively, after the statistical analysis of the survey data in the following.

B. Statistical Evaluation of Different Confusion Cost Valuations

In order to study correlations between characteristics of the survey participants and their assessment of confusions, we statistically evaluate the survey data using an analysis of variance (ANOVA). We use this analysis technique to find differences in the means of data between two groups (one-way ANOVA). As we deal with matrices (of size 6×6), which are not suitable for a standard F-test, we apply a modification thereof, see Appendix A for mathematical details. To put it another way, we perform a statistical test to analyze the degree of variability between two average cost matrices. To this end, the degree of variability is given by the F-statistic

$$F := \frac{MS_B}{MS_W} = \frac{\textit{between-groups variance}}{\textit{within-groups variance}} \in \mathbb{R}_{\geq 0} \; . \tag{6}$$

The greater the F-statistic, the greater the variance of the confusion valuations between different groups, implying that the examined groups differ more significantly. After shuffling the data to randomly assign groups to any cost valuations, we determine how often the random F-statistic (denoted as F_{random}) is greater than the actual calculated F-statistic. In other words, we test the actual survey data against randomness. This approach is also known as bootstrapping [28] and provides the approximated p-value

$$p := \frac{\#\{F_{random} > F\}}{\#\{F_{random}\}} = \frac{\textit{number of times } F_{random} > F}{\textit{total number of simulated } F_{random}} \in [0, 1] \; , \tag{7}$$

which indicates how likely two average cost matrices are drawn from the same distribution, i.e. how likely there is *no* difference between the examined groups with respect to their confusion cost valuations.

In our analysis, we test for differences when comparing the groups *passenger of the self-driving car* vs. *external traffic participant* and *female vs. male*. The results of our conducted statistical tests are summarized in Fig. 8. Regarding the evaluation of the perspective the p-value is $p = 0.52\%$, while regarding the gender $p = 0.00\%$. Such small p-values indicate that there is significant evidence for statistical differences between the average cost matrices associated with the examined groups. This is particularly significant when comparing the groups *female* vs. *male* since in our analysis there has not been a single randomly simulated F-statistic that exceeds the actual calculated F-statistic. Based on the findings of these statistical tests, we conclude that the characteristics related to the assigned perspective as well as the gender of the survey participants both likely have an impact on their respective confusion costs valuation.

However, the evaluation methodology in this subsection does not provide information regarding by what amount and for which specific type of confusion the costs differ. To this end, we further investigate the differences between
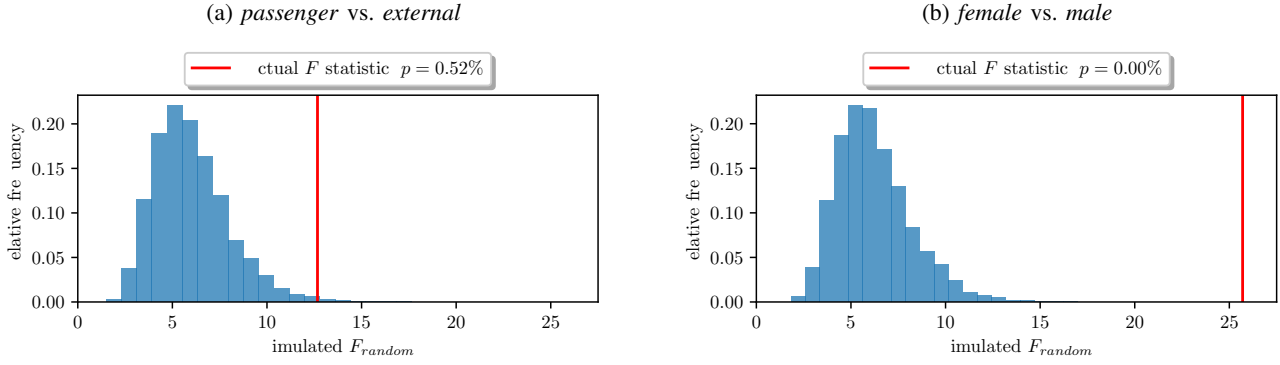
(a) *passenger* vs. *external*   (b) *female* vs. *male*

Fig. 8: The distributions of simulated F-statistics, denoted by F_{random}, when the group affiliation of cost valuations in the survey data is randomly assigned. To this end, the survey dataset is shuffled 1M times in total. The greater the F-statistic, the greater the variance of the confusion cost valuations between the groups in the respective comparison. The red line corresponds to the actual calculated F-statistic according to the original, untouched survey data.

(a)
$$\Delta_{\substack{passenger \\ external}} = \begin{pmatrix} 0 & 0.1313 & 0.1377 & 0.1971 & 0.2054 & 0.3032 \\ 0.1054 & 0 & 0.3394 & 0.1407 & 0.0097 & 0.1528 \\ 0.0655 & 0.1746 & 0 & 0.0267 & 0.0275 & 0.0761 \\ 0.1303 & 0.1245 & 0.0939 & 0 & 0.0768 & 0.0046 \\ 0.0098 & 0.3672 & 0.2676 & 0.0319 & 0 & 0.0895 \\ 0.0207 & 0.3129 & 0.1524 & 0.1831 & 0.0622 & 0 \end{pmatrix} \begin{matrix} \text{"drivable"} \\ \text{"nondrivable"} \\ \text{"static"} \\ \text{"info"} \\ \text{"human"} \\ \text{"dynamic"} \end{matrix} \qquad \Sigma_{\substack{passenger \\ external}} = 4.0221$$

(b)
$$\Delta_{\substack{female \\ male}} = \begin{pmatrix} 0 & 0.1131 & 0.3222 & 0.0829 & 0.0762 & 0.1193 \\ 0.1733 & 0 & 0.1663 & 0.1228 & 0.0353 & 0.3988 \\ 0.2960 & 0.0442 & 0 & 0.1706 & 0.0953 & 0.2013 \\ 0.0282 & 0.1386 & 0.1776 & 0 & 0.0088 & 0.0028 \\ 0.0127 & 0.0192 & 0.0787 & 0.0165 & 0 & 0.2112 \\ 0.0574 & 0.1554 & 0.1380 & 0.0326 & 0.1427 & 0 \end{pmatrix} \begin{matrix} \text{"drivable"} \\ \text{"nondrivable"} \\ \text{"static"} \\ \text{"info"} \\ \text{"human"} \\ \text{"dynamic"} \end{matrix} \qquad \Sigma_{\substack{female \\ male}} = 3.6394$$

(c)
$$\Delta_{\substack{all\ survey \\ robot}} = \begin{pmatrix} 0 & 0.4195 & 0.0366 & 0.1705 & 0.6560 & 0.2055 \\ 1.2728 & 0 & 1.1589 & 0.5589 & 0.3640 & 0.3462 \\ 0.7875 & 0.7304 & 0 & 0.5362 & 0.4256 & 0.3986 \\ 0.5388 & 0.3456 & 0.8982 & 0 & 0.0495 & 0.0855 \\ 1.6849 & 0.1799 & 0.1685 & 0.2839 & 0 & 0.2705 \\ 0.9179 & 0.1568 & 0.2481 & 0.2635 & 0.2434 & 0 \end{pmatrix} \begin{matrix} \text{"drivable"} \\ \text{"nondrivable"} \\ \text{"static"} \\ \text{"info"} \\ \text{"human"} \\ \text{"dynamic"} \end{matrix} \qquad \Sigma_{\substack{all\ survey \\ robot}} = \mathbf{14.4036}$$

Fig. 9: Value-wise differences of confusion costs assessment between different groups. Here, Δ denotes the difference matrix of confusion cost values for the comparisons of the groups (a) *passenger of the self-driving car* vs. *external traffic participant*, (b) *female* vs. *male*, and (c) *all survey participants* vs. *robotistic cost valuation / Bayes decision rule*. Furthermore, Σ denotes the sum of entries in the difference matrices of the respective comparisons.

average cost valuations for specific types of confusion. In particular, we additionally compare the average confusion cost matrix of *all survey participants* against the *robotistic confusion costs valuation*, i.e. constant costs for each possible type of confusion, cf. Eq. (4). The difference matrices are given in Fig. 9. The previously conducted statistical tests might suggest that the difference in the confusion cost valuation between different groups of survey participants tend to be significant, but we observe that the difference between all survey participants as unit and the robotistic cost valuation is even more drastic. The total sum of entries in the corresponding difference matrix yields 14.40 compared to 4.02 and 3.69 for the perspective and gender, respectively. This result let us conclude

survey group \ performance metric	mean IoU	mean recall	mean precision	human IoU	human recall	human precision
passenger of the self-driving car	82.3	94.7	85.9	79.9	**94.6**	85.0
external traffic participant	81.1	93.9	85.1	82.8	93.7	86.2
female survey participants	80.9	94.1	84.8	82.9	94.0	86.0
male survey participants	82.0	94.6	85.8	82.7	94.2	85.6
all survey participants	81.6	94.4	85.3	82.7	94.3	85.5
robot / Bayes decision rule	**90.0**	**95.2**	**94.0**	**84.1**	91.8	**88.2**

TABLE I: Semantic segmentation performance on the Cityscapes validation data when incorporating different costs of confusion provided by different groups. The most commonly used performance metric for this computer vision task is the *mean IoU*, which is the class-wise intersection over union (IoU) averaged over all available classes. For further insights, particularly with respect to performance tradeoffs (cf. Section II-B) we additionally include the metrics *recall* as well *precision* (cf. Fig. 3), and also restrict the evaluation to the human class only. Note that the values in this table denote percentages (%). For all performance metrics higher values indicate better semantic segmentation performance with 100% being the score of a perfect classifier.

that the subjective sense of the survey participants with respect to confusion costs in the context of street scenes *clearly disagrees with the robotistic confusion costs valuation*, and thus with the Bayes decision rule, which is used by default in any machine learning model.

C. Qualitative Evaluation of Different Confusion Cost Matrices

In this subsection, we investigate the actual changes in AI perception of street scenes obtained by using the confusion cost matrices from the survey, cf. Fig. 7. We test the quality of AI perception by integrating the given confusion cost matrices into the cost-based decision rule within a deep neural network (DNN) for semantic segmentation, cf. Eq. (2). In this context, semantic segmentation can be described as pixel-wise image classification task, i.e. assigning an object category to each single pixel of an input image, and thus providing the finest level of detection and localization of objects in scenes. As underlying semantic segmentation DNN we employ the state-of-the-art `DeepLabV3+` model with `WideResNet38` backbone trained by Nvidia [29].

We report numerical semantic segmentation performance results in Table I. As expected, using the Bayes decision rule yields the best overall segmentation performance. The commonly used performance metric mean IoU drops by 8.4 percent points when using the confusion costs from the survey. Among the examined survey groups, the group *external* achieves the best score with 82.3%. Restricting the evaluation on human classification only, the Bayes decision rule still performs best, but by smaller margin of 1.4 percent points over the cost valuations of all survey participants. Among the survey groups, the group *female* achieves the best score with 82.9%. Noteworthy, all groups improve recall over the Bayes decision rule by up to 2.8 percent points for the group *passenger*.

For visual inspection, we provide exemplary semantic segmentation masks for one scene of the Cityscapes dataset [13] in Fig. 10. In general, we realize that all semantic segmentation masks look similar in large parts. In particular with respect to human classification, differences in the outputs are barely visible. The small changes compared to the standard Bayes decision rule include slightly larger segment predictions for the class human, which is a consequence of the increased sensitivity towards human predictions already discussed in Section IV-A. Between the examined survey groups the visual differences in the semantic segmentation masks are marginal, particularly in regard to safety relevant classes such as "road", "info" and "human", which let us conclude that the cost matrices provided by the survey participants do not significantly impact the segmentation quality via deep neural networks.

D. Consequential Evaluation of Different Confusion Cost Matrices

For the consequential evaluation of different cost matrices, we use the safety-aware evaluation tool for the perception of scenes introduced in [14]. We again test the confusion cost matrices from the survey data on the Cityscapes dataset [13], which provides image data with humans and their associated distance to the self-driving car. Furthermore, we compare two groups of survey participants against each other as well as the single groups against the robotistic cost valuation / Bayes decision rule. The evaluation tool generates the bird's-eye view graphics displayed in Fig. 11 and Fig. 12, in which the complete area in front of the self-driving car as captured by the camera is shown. The described area is of circular form with a center angle of 60°, which is naturally given by the

(a) passenger of the self-driving car

(b) external traffic participant

(c) female survey participant

(d) male survey participant

(e) all survey participants
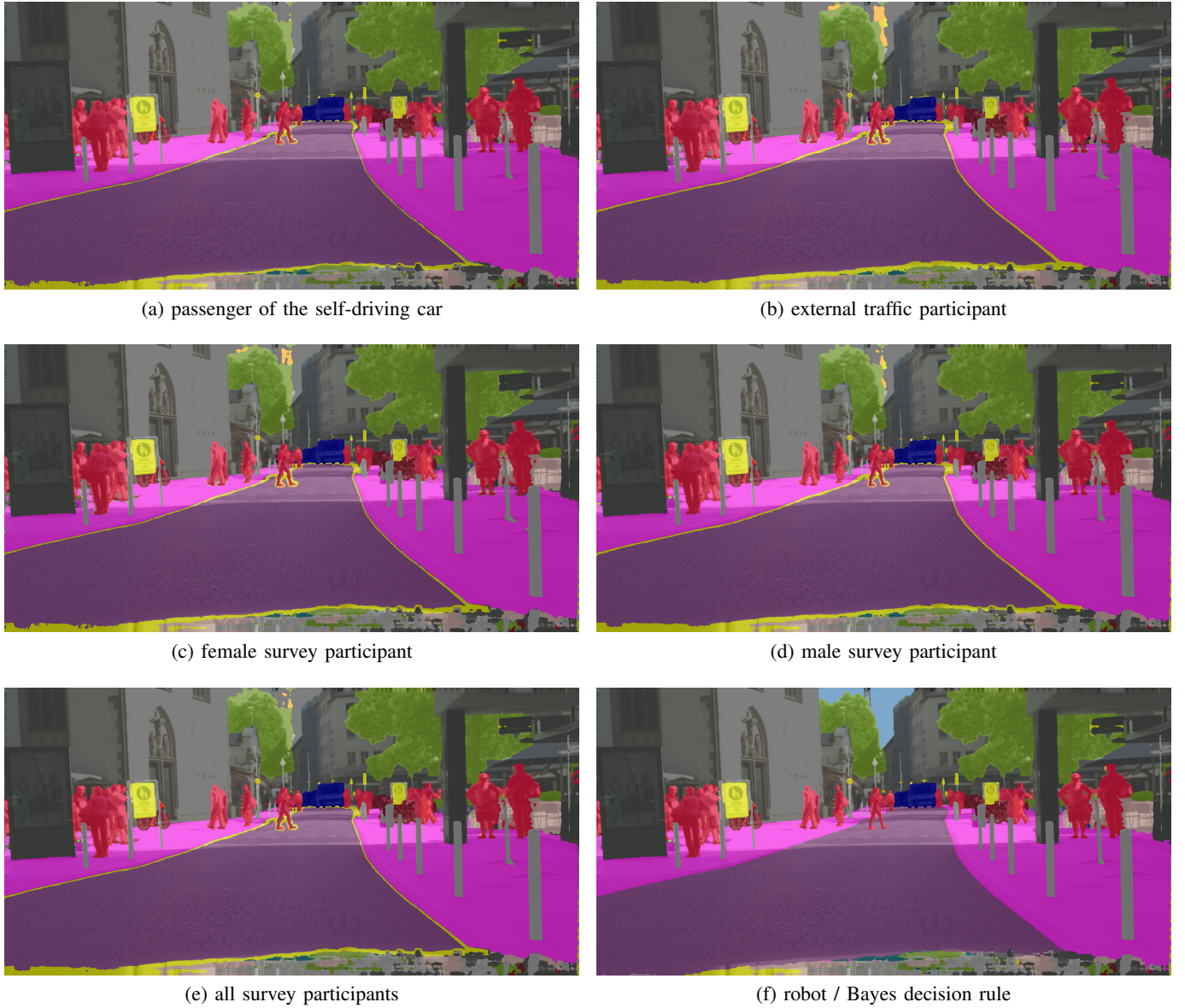
(f) robot / Bayes decision rule

Fig. 10: Changes in AI perception of an urban street scene by using different confusion cost matrices, cf. also Fig. 7, in the cost-based decision rule integrated within a state-of-the-art deep neural network for semantic segmentation.

field of vision of the camera. We subdivide the area in front of the self-driving car into two safety-critical zones, that are 1) the braking distance at a speed of up to 50 km/h and 2) the braking distance at a speed of up to 30 km/h. Accordingly, we assume braking distances of 46.5 meters and 20.6 meters, respectively, which are indicated by the shaded zones in the bird's-eye view graphics.

Moreover, the crosses and dots in the graphics indicate the position of actual human instances relative to the self-driving car. In this regard, we consider a human instance as detected if the associated recall exceeds a threshold of 50%, cf. Fig. 3, otherwise we consider that human instance to be overlooked.[2] Then, gray crosses represent human instances that are overlooked by both considered groups in the comparison, while colored dots represent human instances that are detected by only one group.

As discussed in Section II-B, the interaction between recall and precision is always an interaction between two opposing metrics. Therefore, we additionally consider the precision in order to examine to what extent this metric

[2]**Disclaimer**: Note that the outlined score is also commonly referred to as detection threshold and it is commonly set to 50% in the field of computer vision. However, using any other detection threshold between 0% and 100% is obviously also possible.
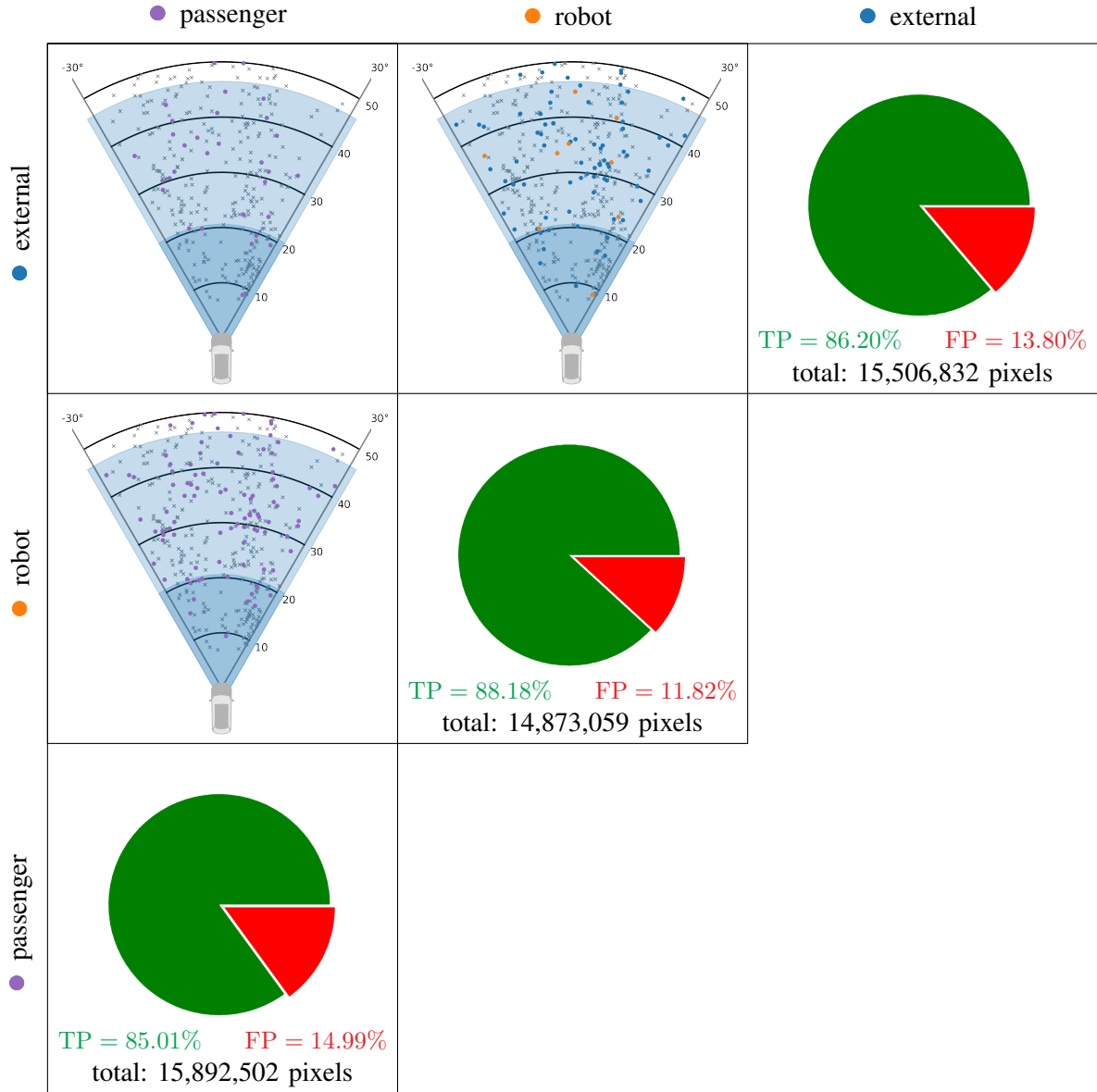
Fig. 11: Consequential comparisons between the cost assignments of the groups *passenger*, *external*, and *robot* with respect to the detection of humans. On the diagonal, the pie charts indicate the precision of human predictions as practicability measure for the respective cost assignments. The graphs on the off diagonal display the perceivable area ahead of the ego-vehicle from a bird's eye perspective. Here, gray crosses show human instances that are overlooked by both considered groups in the comparison, whereas colored dots show human instances that are detected only by the group corresponding to the respective color.

is sacrificed to improve recall. In Fig. 11 and Fig. 12, the precision is illustrated as pie charts, which we also refer to as *practicability measure* of the self-driving car. In this light, a low precision could yield an autonomous vehicle that mistakenly brakes all the time, since the underlying AI incorrectly identifies human instances all the time, and thus resulting in an impractical application of AI in automated driving. Via these just described visualizations, we aim at presenting different safety related consequences of AI perception with respect to human classification in a comprehensible and compact way. Besides, we report numerical results in Table II.

We have seen in Section IV-C that different confusion cost valuations result in only marginal visual differences in the semantic segmentation masks. However, the consequential evaluation in this subsections reveals that the different confusion cost valuations from the survey yield sufficient changes in AI perception such that the number of overlooked human instances in safety-critical distances to the self-driving car may differ. We observe that the
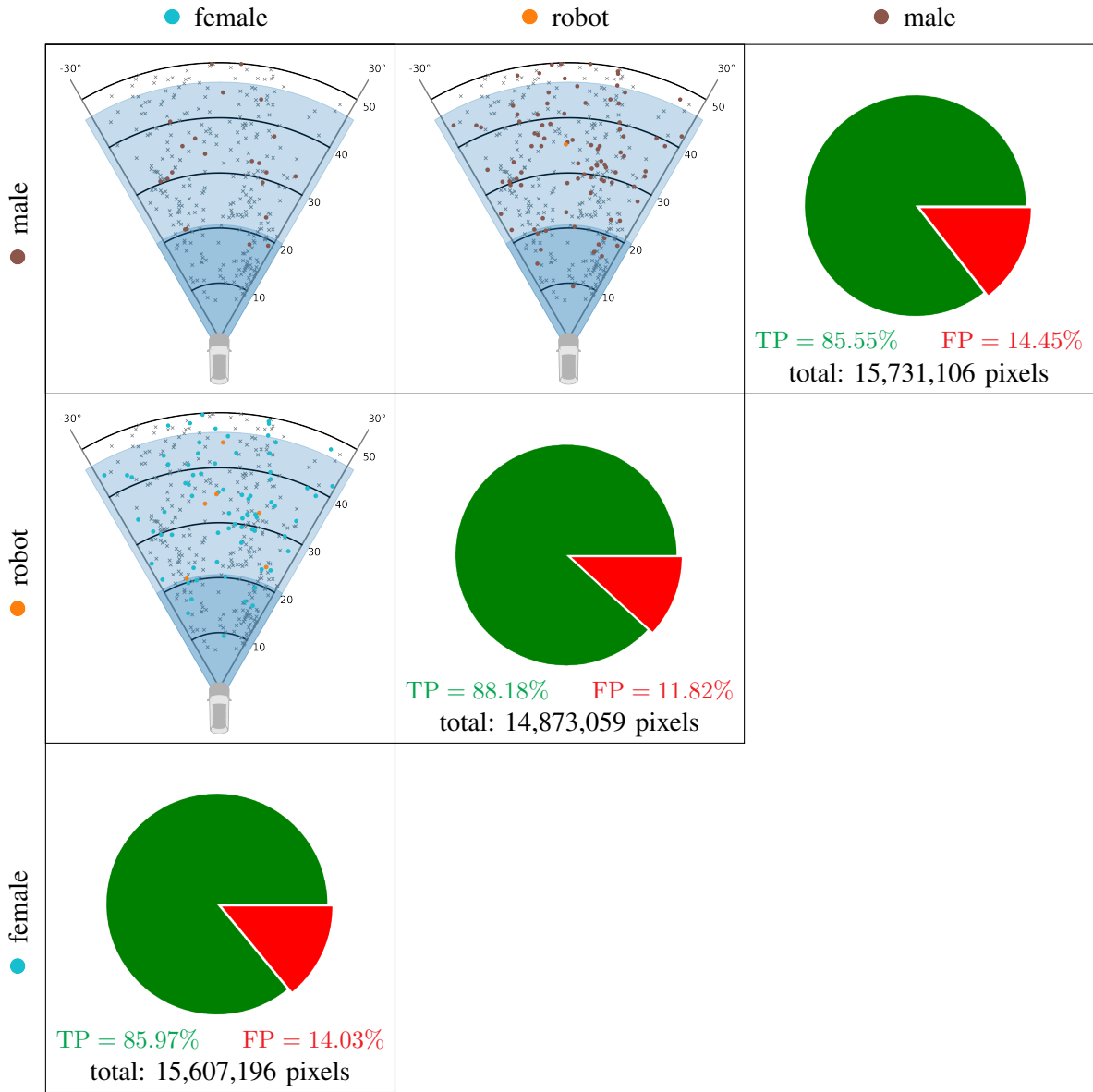
Fig. 12: Consequential comparisons between the cost assignments of the groups *female*, *male*, and *robot* with respect to the detection of humans. On the diagonal, the pie charts indicate the precision of human predictions as practicability measure for the respective cost assignments. The graphs on the off diagonal display the perceivable area ahead of the ego-vehicle from a bird's eye perspective. Here, gray crosses show human instances that are overlooked by both considered groups in the comparison, whereas colored dots show human instances that are detected only by the group corresponding to the respective color.

groups *passenger of the self-driving car* and *male survey participants* provide more human sensitive predictions compared to the groups *external traffic participant* and *female survey participants*. The number of overlooked human instances differ by up to 73 within the braking distance of 50 km/h and up to 5 within the braking distance of 30 km/h. More significantly, all the examined groups produce more human sensitive predictions than the *robotistic cost valuation / Bayes decision rule*. With respect to the group *passenger of the self-driving car*, the standard Bayes decision rule overlooks 211 and 13 human instances within the braking distance of 50 km/h and 30 km/h, respectively. As already seen in Table I, the reduction of overlooked human instances, i.e. gain in recall, is accompanied by a sacrifice of 3.2 percent points in precision. The latter loss could still be considered "acceptable", particularly in light of preventing life-threatening street scenarios.

safety-critical zones	total number of human instances	number of overlooked human instances				
		passenger	external	female	male	robot / Bayes
within braking distance of 50 km/h	2394	**534** (13.5%)	607 (15.4%)	599 (15.2%)	550 (13.9%)	745 (18.9%)
within braking distance of 30 km/h	817	**77** (9.4%)	82 (10.0%)	81 (9.9%)	79 (9.7%)	90 (11.0%)

TABLE II: The amount of overlooked human instances in the Cityscapes validation dataset when using the confusion cost valuations provided by the different groups of survey participants as well as the Bayes decision rule.

V. Reflection and Limitation of the Survey from a Psychological Point of View

In the development of responsible AI technology, active participation and involvement of users and stakeholders take on a new significance. This is especially true if the AI technology is expected to make decisions or act in accordance with human values and not purely based on probabilities, as it is the case in automated driving [30], cf. also Section II-A. Then, it is not solely a matter of achieving acceptance of the new technology among human users but also a matter of AI technology learning from users, their values and ways of thinking. With the survey conducted in the present study, we aim at gaining insights for the development of AI perception in the field of automated driving by exploring human judgments concerning the costs of potential confusions between object classes.

In traffic psychology, surveys are an established method for uncovering humans' insights, e.g. into their evaluation and usage strategies of driver assistance systems [31]. Besides the obvious advantages of this method, self-reports also bear some limitations that must be considered when interpreting the results. In the following, we critically reflect on possible limitations of our conducted survey on different but strongly interacting dimensions (survey and person) in light of the survey participants' feedback. To this end, we asked participants at the end of the survey to rate the degree of perceived difficulty in completing it ("How complicated did you find the survey?", 1 = very simple, 2 = fairly easy, 3 = challenging, 4 = complicated, 5 = extremely complicated). On average, participants experienced the survey as challenging (*mean* = 3.04, *standard deviation* = 0.88, *number of participants* = 412). In addition, participants had the opportunity to leave feedback on the survey as free text. The following reflection on the survey builds upon these responses (which we have translated from German into English).

A. Survey-related Factors Influencing the Response Behavior of Survey Participants

The subject of our survey was particularly complex and challenging. Even though automated driving is an issue of high current relevance, it still appears as a black box for most people and is surrounded by uncertainty. This uncertainty is associated with every innovation and results from a lack of knowledge about the functionality and use of the new technology, see e.g. [32]. For AI technology, this is especially true because its underlying highly complex algorithms make it impossible to entirely see through and comprehend the AI's decision making processes, even for experts, see e.g. [33]. Due to the limited capacity of human working memory [34], at any point in time, only a certain amount of information can be processed simultaneously, so that some information decays. The intransparency of how an AI operates, however, counteracts the development of trust in the technology. Particularly in the field of automated driving, trust in AI's decision making is vital for acceptance and adoption of this technology in society, see e.g. [35]. Accordingly, there is an urgent need for increasing transparency of the underlying models in automated vehicles and, consequently, moving towards an explainable AI [36]. Yet, increased transparency ought not to come at the expense of AI's decision accuracy. To ensure a high level of decision accuracy, AI technology inevitably requires the processing of large amounts of data far beyond the scope of human comprehension. Thus, there is a fundamental tension between the functioning of the AI being based on a vast amount of data and sophisticated algorithmic models on the one hand, and, on the other hand, the constraints of human cognition, which is usually not able to decode such complex data and models. This causes a dilemma when asking people to partly empathize with an AI, as it was the case in our survey. When designing our survey, we had to find a balance between addressing participants in a way they could still comprehend and, at the same time, not obscuring, but sufficiently capturing the complexity inherent in the topic to obtain meaningful input to feed back into the AI. Thus, despite our best efforts, the survey inevitably involved a certain degree of complexity that might have caused a high cognitive intrinsic load [37]. Specifically, participants in our study had to make responsible decisions when assessing the level of fatality of confusion between objects in the given traffic scenarios. For every scenario, they had to consider a vast number of aspects and possible consequences for each of the five confusion cases. Hence, participants in our study may not have been able to recognize and process all relevant information sufficiently at a given point in time. Some

participants reported this high level of complexity along with substantial uncertainty and problematized it as part of their feedback.

> "There are many aspects that must be taken into account. It would be particularly fatal if a person is not recognized as a person and therefore the car does not swerve or brake. With different static or dynamic objects a confusion is less serious in my opinion, because the car should be programmed to swerve in any case. Nevertheless, the issue is very difficult, because it is not possible to consider every aspect and the topic itself is very complex. It is also likely that in retrospect you will notice criteria that you spontaneously did not consider when answering the questions."

> "I struggled to empathize with an AI. What information does it gain from dynamic objects? [...]"

> "too many details go into a decision [...]"

Besides the challenging task and content, the special design and technical features of our survey may have caused an increased extraneous cognitive load that is mental load produced by the layout and design of the material [37]. For example, some participants reported that they had issues determining the weighting of fatality using the sliders and unfamiliar units of value (1-1000000 times more fatal than the confusion car with bus). We chose this scale as it is coherent with the specific design of the used AI model and explicitly intended *not* concealing the actually used confusion costs, like it is the case in the standard Bayes decision rule, cf. Eq. (5).

> "For me, it was difficult to determine exactly which slider how to move so that it matched in relation to the other classes. [...]"

> "School grades or grading from 1-10 would be easier. As a non-specialist it is very complicated to understand the evaluation criteria..."

> "The reference values are difficult to assess and the wording of questions is rather complicated"

To sum up, especially due to the complex characteristics of the topic and task, as well as some design features of our survey, some participants presumably experienced increased intrinsic and extrinsic cognitive load while completing our survey. Consequently, if extraneous and intrinsic cognitive loads were too high, this may have exceeded the limited working memory capacity, so that sufficient cognitive resources that would have been necessary to master the task may not have been available [37]. In other words, some participants may have been more concerned with understanding the complex scenario and task as well as with becoming familiar with the interface and features of the survey, than with collecting, combining and, weighting relevant information to finally make a sound decision of how fatal they judge the confusion between classes. As a consequence, to save cognitive resources, participants who experienced such cognitive overload may have used less elaborate judgment heuristics, which could risk resulting in poor decision making [38].

The extent to which the potentially challenging aforementioned features of the survey may have affected participants' responses might be mediated by personal characteristics.

B. Personal Characteristics influencing the Response Behavior of Survey Participants

Various personal characteristics could have determined to what extent the specific features of the survey might have affected participants' response behavior. For instance, the degree of intrinsic cognitive load essentially depends on whether participants were already familiar with the topics of automated driving or AI in general, and whether they could have drawn on theoretical or even practical experiences. If this has been the case, participants will more likely have perceived the survey as less complex and with less uncertainty due to their existing knowledge than participants who were entirely new to the topic. This is also evident from some feedback reports, in which participants with little or no prior knowledge described their difficulties while completing the questionnaire.

> "Despite a short introduction to the topic, I think it is too complicated for non-experts. A simplification [...] might be necessary."

> "Perspectives very technical, more practical explanation for people who really have no idea at all (like me)"

Accordingly, we assume that the more experience and knowledge someone had about AI and automated driving in general, as well as about scenarios like those presented in our survey more specifically, the easier it was to cope with the decision problem of evaluating the fatality of confusions. Participants with relevant experience and prior knowledge likely required less cognitive resources to understand the task and become familiar with the scenarios, allowing them to devote more cognitive resources to the actual decision making process. This would also imply that there might have been some kind of training effect over the course of evaluating the scenarios while completing our survey. This assumption stands in line with some of the participants' feedback responses.

> "You don't understand some questions/situations until you answer the same question several times."

> "Only in the course of the survey did I get a sense of what I would rate stronger or weaker, because I couldn't think much about the consequences beforehand."

Furthermore, the perceived ease of use of the web-based survey with its technical features probably differed between participants depending on their digital literacy. To be more specific, we expect participants with pronounced digital literacy to have been less distracted and overwhelmed by the technical features of the survey than participants with low digital literacy. Relying on the Technology Acceptance Model, perceived ease of use is one of the two key factors (besides perceived usefulness), which shapes an individual's attitude towards a technology and therefore serves as a predictor for its adoption [39]. Accordingly, perceived ease of use, depending on the degree of digital literacy, could have determined the participants' attitude towards our survey and consequently their level of engagement. The extent to which the subjects were engaged in their participation in our study certainly also depended on their general interest in topics such as AI and automated driving. In terms of a positive selection, it should be taken into account that the participants who participated voluntarily in our study may have had a general interest in the topic. In turn, interest in and attitudes toward the special topics of AI and automated driving vary with other factors. For example, [40] pointed out, people's interests for and attitudes toward self-driving cars significantly differ between ages. In their study, they found that older people have, for example, less interest and less confidence in automated driving than young people. Furthermore, gender could have had an influence on both, the attitude towards the tech-centric topic and interface of our survey as well as on the participants' engagement and response behavior. For example, women, compared to men, tend to have higher computer anxiety and lower computer self-efficacy [39]

The list of factors that may have influenced participants in completing the questionnaire is long and cannot be discussed exhaustively in this paper. However, it should have become clear that both the specific nature of our survey and personal characteristics not analyzed in our study may have significantly influenced the participants' judgments of the fatality of confusion between objects of different classes. Thus, the results of our study should be interpreted in light of these limitations. Nevertheless, our study provides significant insights for the integration of human ethical judgments in the development of AI technology for automated driving and offers several starting points for further research in this field.

VI. THE CONFUSION COST MATRIX FROM AN ETHICAL POINT OF VIEW

The Eq. (2) in Section II provides the tradeoff space within which the Bayes decision rule amounts to a specific choice with moral significance. Its choice reflects the wish to maximize accuracy in the prediction process, cf. Eq. (5). A choice which prima facie seems to stand in conflict with our moral intuition, which e.g. would not set the cost of confusion between a human and a traffic light to be symmetric, as it is the case in the Bayes decision rule according to Eq. (4). But if the cost of confusion should be distinct from the choice in the Bayes decision rule, what should it be? What would justify a decision rule that would decrease predictive accuracy? Any specific

choice will amount to a decision about possible accidents happening or not happening. A feature that it shares with the infamous trolley problem [41].

We will now briefly consider the role of normative theories of ethics and criticisms set forth against the trolley problem and whether they similarly apply to the the choice of a confusion cost matrix. This will be followed by an application of ethical guidelines to the definition of a confusion cost matrix. Finally, we will consider the role surveys may play within an ethical analysis of the confusion cost matrix.

A. Normative Ethics and the Trolley Problem

Note that the tradeoff space set up by the confusion cost matrix in Eq. (2) is where a decision has to be made and where a conflict arises between a symmetric confusion cost matrix with our moral intuition. Normative theories of ethics provide various frameworks within which these choices can be ethically assessed. There are inter alia ethical theories relying on virtues, there are rule-based or deontological accounts and also normative theories that determine the ethically 'right' choice by relying on the consequences only [2]–[4]. However, since there is no agreement among ethicists about the "right" normative approach (see [42, p. 676]) the value of these accounts in individual cases remains contested.

The ethical analysis of the confusion cost matrix relies on two relevant assessment stages, which may or may not impact the moral deliberation. First, the choice of the confusion cost matrix itself, which, as indicated above, may amount to a morally significant choice with probabilistic consequences (stage 1 assessment). And second, the analysis of what any individual choice of the confusion cost matrix would actually amount to in concrete cases in terms of possible harms in non-probabilistic terms (stage 2 assessment). One difficulty may arise due to e.g. possible assessments of stage 1 choices on the basis of e.g. an overall reduction of harm of individuals, which then may be in conflict with individual stage 2 assessments, which may be concerned with the offsetting of victims – an ethically problematic procedure. The assessment of stage 2 is crucial for a consequentialist analysis of stage 1, while for other non-consequentialist normative theories the stages can in be assessed independently. Note that the assessment of stage 2 can only play a role in an ex post re-evaluation of the stage 1 choice of the confusion cost matrix. It is the stage 2 assessment, however, which shares features with the trolley problem. We therefore consider now briefly the trolley problem and discuss why issues set forth against its relevance, do not affect our current ethical analysis.

In brief, the traditional trolley case is concerned with a trolley that is heading directly towards five people that would die on impact. A lever would allow to switch the tracks of the trolley, such that it would head towards one person and kill that person instead. Should the lever be switched or not? The conceptualization of distinct conflicting moral intuitions regarding what the answer should be has been a much-discussed topic in moral philosophy [43]–[45]. However, these often abstract discussions are of more practical relevance in the context of autonomous vehicles, where these choices may need to be programmed in a certain way. This has led to an extensive debate about the right way to program the autonomous vehicle (see e.g. [46]–[48]). However, recently the practical relevance of the trolley problem has been challenged in [42], [49], [50].

Following Keeling [51] we consider two objections to the relevance of trolley cases for autonomous vehicles and whether they would similarly apply to the determination of a confusion cost matrix. The first objection puts doubt on whether the autonomous vehicle will actually encounter trolley cases and so objects to the relevance of moral deliberation for practical applications. Trolley cases like the one mentioned above will amount to extremely rare circumstances and will therefore not be of practical relevance or so the argument goes. Irrespective of whether this is the case for trolley cases, this objection does not apply to the above outlined choice of the confusion cost matrix, as any specific choice of decision rule is a precondition for the application of the algorithm in the first place. As such the decision sets in at an earlier stage in the process of development. Nevertheless, there is a possible analogous objection. Namely, whether the once reasonably restricted choices for the costs of confusions will in effect amount to statistically significant differences in the outcomes.

To assess this point we can now refer to the results of the survey in Section IV and the results of the consequentialist analysis in Section IV-D. The survey points to differences in the choice of the cost of confusion between gender (male, female) and perspective (passenger, external). Nevertheless, the corresponding segmentation masks only display marginal differences. So even though there are differences in opinions regarding the decision rules, these will not, at least prima facie, significantly change what the AI can "see". However, this prima facie irrelevance of the

survey-based confusion matrix does not remain irrelevant once one considers the full consequentialist assessment of these decision rules. Here it becomes apparent that the survey-based decision rules do, unsurprisingly, a better job in recognizing humans compared to the Bayes decision rule. So an analysis of possible variations of the decision rule do lead to significant differences of relevance for moral deliberation. Note that even if there would not have been a difference, the irrelevance objection would not apply, as for the determination of the decision rule, the irrelevance claim is a result of the assessment and not something that can be applied beforehand.

The second objection to trolley cases is the moral difference argument according to which the moral deliberation regarding the choice between pushing the lever or not, differs in a morally significant way from the deliberation in real world cases. In real world cases there are additional circumstances that would impact the moral deliberation of the tradeoff space. For instance the manufacturers may have certain obligations (e.g. protecting the safety of the passengers) or usually the outcomes are not known a priori, but may only be associated with probabilities. As has been argued in [42], [50], reasoning under uncertainty differs significantly from reasoning about known facts. This is a significant objection to trolley cases that similarly impacts the deliberation on the possible choices in our tradeoff space.

One may abstractly morally deliberate on certain choices of the confusion cost matrix, but these would be significantly impacted by additional information about the possible outcomes that would be associated with these choices. While we did consider the consequences of specific choices of the confusion cost matrix in Section IV-D, thereby mitigating this objection to some extent, it remains a possibly categorical objection to the significance of survey results, where limitations to the human capability to reason with probabilities and uncertainties will restrict the possible value of these results. This is nicely illustrated by the somewhat surprising result of the consequentialist analysis that assigning generically high costs in the survey has a mitigating effect in valuing human lives. A consequence that probably was not foreseen by the survey participants aiming at a more conservative driving attitude.

The moral difference argument therefore points to the possibly problematic nature of the already discussed difference between the stage 1 and 2 assessments. However, the consequentialist analysis in Section IV-D provides a tool that does address this problem head-on by making transparent the possible moral differences in the choices in the survey compared to an assessment of their corresponding consequences. A lesson one may learn from this, is that for more reliable survey results, i.e. survey results which more completely capture the individuals' choices, one may need to complement the determination of the cost confusion matrix with the respective birds-eye view pictures of the corresponding consequences.

B. Ethics Guidelines of the Ethics Commission on Automated and Connected Driving

In 2016 an ethics commission on automated and connected driving was appointed by the German federal minister of transport and digital infrastructure to develop ethical guidelines for automated and connected driving. In June 2017 they published a report with 20 "ethical rules for automated and connected vehicular traffic" [9]. Since then many more guidelines have appeared, significantly in the European context the recommendations of the "Ethics of connected and automated vehicles" report by an independent expert group implemented by the European commission (EC) [10]. We will now consider the relevant rules and what they imply for the choice of the confusion cost matrix and the inclusion of a survey-based approach.

Ethical guidelines provide recommendations that are often intentionally stated at an abstract level to ensure generality, while at the same time need to be concrete enough to lead to actionable advice. This presents a tension that is not always easy to implement. For instance, rule 5 of the German report states that the "[a]utomated and connected technology should prevent accidents wherever this is practically possible", and the design and programming of the vehicles "be such that they drive in a defensive and anticipatory manner, posing as little risk as possible to vulnerable road users". One difficulty in implementing this rule is the ambiguousness regarding what may be considered "practically possible" or what "as little risk as possible" mean. Both terms do not recognize the tradeoff situation present in these situations. The term "as little risk as possible" may suggest that it is a problem of minimization, where in fact it is an optimization problem, where various values and their respective moral considerations need to be balanced.

Being aware of this predicament, the report of the EC expert group recognizes, after listing the guiding ethical principles, that "[t]he above principles cannot be applied with a mechanical top-down procedure. They need to be

specified, discussed and redefined in-context.". They further note that this "is why the design and development of CAV (connected and automated vehicle) systems should be supportive of and resulting from inclusive deliberation processes involving relevant stakeholders and the wider public.". That is, the concretization and implementation of any individual recommendation should not be the decision of individual scientists but include the "wider public" (as it is aimed with the survey-based approach in Section III and Section IV) and other stakeholders (see industry perspective in Section VII) as well as being "supportive of" the inclusive deliberation process (which we implement by considering the limitations and capabilities of the survey-based approach in Section V).

The starting point of our analysis is in line with recommendation 14 of the EC expert group, which requires a reduction of opacity in algorithmic decisions, a recommendation necessitated by the fact that "algorithm-based CAV systems [...] may operate as "black-boxes" that do not allow cognitive access to how they have arrived at a particular output, or what input factors or a combination of input factors have contributed to the decision-making process or outcome" [10, p. 49]. By recognizing the implicit tradeoff space within which the Bayes decision rule is making a morally significant choice one opens up the "black-box" to scrutiny. This has led us to the two assessment stages, which in turn now need to be confronted with what the other ethical rules and recommendations imply for them.

So let us now turn to the relevant rules and recommendations of the two guidelines regarding the assessment of the two stages. Rule 7 of the German report states that "within the constraints of what is technologically feasible, the systems must be programmed to accept damage to animals or property in a conflict if this means that personal injury can be prevented". This statement addresses the stage 2 decision regarding the offsetting of various scenarios. This may be translatable, and therefore technologically feasible by setting strong asymmetric cost functions, valuing the human category significantly higher than other entities. For that one needs to consider the relation between choices of the confusion cost matrix and their consequences in specific scenarios. Something already considered in Section IV-D. This, however, remains in conflict with the general practicability of the autonomous vehicle (AV), cf. Section II-B, and thus needs to be put in balance with a justification for the practicability of the AV. An impracticable AV would not be introduced and therefore could not provide "promises to produce at least a diminution in harm compared with human driving, in other words a positive balance of risks" (rule 2). This is recognized as a necessary part of the justification of AVs in general.

In Rule 9 the ethics commission formulates a clear prohibition "to offset victims against one another". They provide two reasons for that. First, there is a simple practical issue mentioned in rule 8: "a decision between one human life and another, depend on the actual specific situation, incorporating "unpredictable" behaviour by parties affected. They can thus not be clearly standardized, nor can they be programmed such that they are ethically unquestionable". This practical reason is complemented with an underlying Kantian argument speaking against a pre-programmed stage 2 decision. From a Kantian perspective an individual enjoys the right of moral self-determination, a right that the person would be robbed off in case of an externally determined decision, which "in extremis, [would] be able to take correct ethical decisions on the demise of the individual human being". Rule 9 therefore amounts to a prohibition to base the confusion cost matrix on the basis of individual stage 2 decisions.

The ethics commission, however, continues in rule 9 by stating that a "[g]eneral programming to reduce the number of personal injuries may be justifiable". The reason for that is that a general programming of this kind would be abstract in the sense that it does not take the life of any specific individual into consideration, as these are not known at this level, and further "the programming reduce[s] the risk to every single road user in equal measure" [9, p. 18]. As such a stage 1 decision that would only be statistically determined by individual stage 2 assessments would be in line with the recommendations of the ethics commission, as long as overall personal injuries would be reduced.

The report by the EC expert group takes a more general approach regarding dilemma situations, by considering acceptable behaviour in dilemma situations as something that can "organically emerge from the adherence to the principles of risk distribution stated in Recommendation 5". Risk-distribution in recommendation 5 allows for a differential behaviour around a certain subset of road users in cases where these belong to certain vulnerable groups. A different behaviour of the CAV regarding these vulnerable groups then amounts to an opportunity to redress inequalities present in current traffic collisions. That is, it would allow one to adequately account for the ethical issues involved with the consideration of e.g. wheelchair users, cyclists and visually impaired users. This is of course only a viable option, if the algorithm has the capability to distinguish between these groups. In the case of the cost confusion matrix above, the corresponding classes do not distinguish between possible vulnerable

groups except the distinction between the classes "human" and "dynamic". At this stage the report again refers to the necessity to discuss the ethical and social acceptability "as a topic for inclusive deliberation" [10, p. 31]. This now leads us to the survey-based approach and its ethical ramifications.

C. Surveys, the Moral Machine Experiment, and Ethics

The moral machine experiment is a massive survey-based analysis of the trolley problem [6]. By including more than 40 million decisions by over two million participants from over 200 countries it provides an extensive analysis about the moral intuition of people around the world regarding the trolley problem and its variations. It has since generated a lot of discussion and criticism, largely due to the framing of the trolley setups. Providing scenarios where one may have to choose between athletes vs. overweight persons or executives vs. homeless persons may lead to information about the corresponding people and their cultures but less on what the ethics of the AI should be, as the introduction of these distinctions in the classification process could already be in violation of basic human rights (see e.g. [52]–[55] for this and other objections against the moral machine experiment).

Even in less problematic decision processes (e.g. decisions in dilemma setups involving human vs. non-human entities, which would be in agreement with rule 7 of the German report) there are limitations to drawing conclusions from the moral intuition of individuals to what ought to be the case. This is true both in the case of more individual implementations of the moral intuition (possibly regionally restricted) as well as the identification of moral intuitions shared by the majority. The former may lead to a moral relativism while the latter may result in the corresponding systematic discrimination of vulnerable minority groups.

So what role can surveys play in the ethical analysis of tradeoff situations? To consider one of these roles let us briefly introduce another variant of the trolley problem, namely the tunnel problem [56], [57]: Imagine being the passenger of an autonomous vehicle approaching a tunnel, where all of a sudden a child mistakenly runs in front of the tunnel. There are two options, either one hits the child or one swerves to either direction colliding with the wall of the tunnel. So either the child dies or the passenger of the vehicle dies. What this problem points to is the very personal nature of this choice. There is again no clear ethical solution to this case and for either decision there might be good reasons to give.

This has led Millar [57] to suggest an analogy to the case of end-of-life decisions in the healthcare context. Medical professionals are required to seek informed consent from the patients and are not simply allowed to decide for the patient. Similarly, Millar argues, one should incorporate the individuality of the people involved in the decision-making process and it should not be the decision of the engineer or manufacturer of the autonomous vehicle. The EC expert group report does explicitly incorporate in recommendation 8 that one should "[e]nable user choice, seek informed consent options and develop related best practice industry standards", where consent-based user agreements should "go beyond "take-it-or-leave-it" models of consent." This together with the already mentioned necessity for inclusive deliberation processes involving the wider public introduces an important ethical role to survey-based approaches.

VII. PARTICIPATORY APPROACH TO ETHICS FROM THE POINT OF VIEW OF THE AUTOMOTIVE INDUSTRY

One of the major reasons for assessing the quality of perception AI within industrial applications lies with the validation of safety goals. For safety-critical automotive software systems such as advanced driver assistance systems (ADAS), a *positive risk balance* is aimed at, i.e. the risk stemming from the usage of a system shall be less than the risk while not using the system. Furthermore, *unreasonable risk* is to be avoided. Here, risk is called *unreasonable*, if it is "unacceptable in a certain context according to valid societal moral concepts" [58]. Hence, driving notions behind arguing safety for ADAS compare the behavior of a system with the respective human behavior. Positive risk balance and avoidance of unreasonable risk are then argued by careful analysis, development of rigor as well as quick update processes in the case of detected safety issues, and extensive testing.

Different standards aiming at particular aspects of *safety* for automotive software and AI therein are being developed, see e.g. [59], ISO/AWI PAS 8800, and ISO/AWI TS 5083 that are under development (at the time of writing this work). Next to standardization institutions, also legislation is putting regulatory frameworks into place. One such example is the EU AI act, which is currently in preparation. It involves important guidelines, e.g. with respect to the constitution of test sets and AI behavior. This involves bias-freeness, fairness as well as statistical

representativity and diversity both in training and validation data. It also advises human oversight and assessment of the performance.

In this light, those outlined standards could represent some form of instructions that software engineers should follow when programming AI software. Therefore, on the one hand, standards need to take ethics guidelines developed by expert groups into consideration as described in Section VI-B. On the other hand, developing standards additionally also involves the need to understand the social expectation towards safe system behavior and its derived performance requirements on (AI-driven) perception systems. The approach examined within this work, cf. Section III, presents a possible way to understand the required AI-based perception performance as it allows for understanding social expectations towards the semantic perception interpretation in a democratized way. Thereby, this approach may help to raise public acceptance as well as trust for AI-based systems for automated driving.

The findings of this work indicate a clear discrepancy in the results when comparing the participatory approach with the robotistic approach. Relevantly in terms of safety, the perception system resulting from the participatory approach clearly detects more safety critical pedestrians than that from the robtistic approach, while keeping the false positive rate at almost the same level. This points towards the presented approach being a possible candidate for the design process of systems for automated driving (possibly even in both stage 1 and stage 2 assessment, cf. Section VI-A).

Nonetheless, to to be fully usable from an industrial point of view, some future work is still required:

1) Safety *always* relates to the overall system and not to particular sub-components like the perception. It also allows for inner-system resilience, e.g. unreasonably erroneous perception results can be mitigated by the trajectory planner. Hence, the desired perception behavior needs to be assessed within the context of the overall system for deriving safety requirements.

2) Clarity with respect to the given survey tasks has to be ensured. Otherwise the interpretations of the survey results might be misleading.

3) Alternative evaluation schemes such as comparison to human perception, e.g. "What do humans see?" rather than "What do humans expect the robot to see?", could also be taken into consideration to derive an improved understanding of the social expectation to perception performance.

VIII. CONCLUSION AND OUTLOOK

A. Conclusions

In this work we explored a problem of practical ethics in the context of automated driving, namely the definition of the cost of confusion of different categories, such as humans, dynamic objects, road, and others. As one solution strategy, we experimentally explored a participatory approach, i.e. we ask people to enter the values for the cost of confusions in an online survey. We developed statistical tests to prove that different groups of survey participants, e.g. by gender or perspective (passenger vs. pedestrian), differ significantly in their cost valuations. Moreover, we show that cost structures of all survey participants as unit differ even more significantly from the "robotistic" view, in which all confusions come with the same cost. We demonstrated that the different cost structures change the perception of an system of artificial intelligence (AI), which possibly results in individual human instances, within a safety critical distance relative to the self-driving car, to be poorly detected or even being entirely overlooked.

We also explored the limitations of the presented participatory approach from three different disciplinary perspectives. We therefore embed our approach in a broader discussion of practical ethics, investigate the psychological difficulties of taking such decisions on the cost structure, and include feedback from the automotive industry.

Whether participatory elements can become an element in an *ethics by design* approach to AI or whether it can be taken into consideration in regulatory frameworks for AI technology depends on the resolution of several issues that were described in this article.

First of all, further investigations are needed to better understand the potential role of participatory elements for the normative restriction of AI design approaches, let it be in the design step itself or in regulatory texts like technical norms. Here, a deeper understanding of the influence on the setup for participation, such as the survey design, is important. In particular, the difficulties for people to take a decision on the configuration of an AI system without the ability to be aware of all potential consequences needs further consideration.

B. Recommendations for Future Research

The survey conducted in the present study is the result of a weighing process between a design that is understandable for the participants on the one hand and still meets the necessary depth to sufficiently depict the confusion costs for the training of the AI model on the other hand. Maybe future studies can still further refine the questionnaire towards an increased usability without undermining the required level of data complexity. Special attention should then be paid on reducing participants' extraneous cognitive load as far as possible. For instance, this could be achieved by avoiding unfamiliar technical features such as the sliders. Moreover, in future work visualizations should depict the respective perspective from which the participants are asked to assess scenarios. In the present study, participants assigned to the group of external traffic participants were not given different visualizations, i.e. they also view the visualizations from the perspective of car passengers and were only prompted by a textual cue to imagine being outside the car as a pedestrian. This means that in order to evaluate the situation from their assigned role, they need to perform an extra step of mental elaboration, namely switching from the shown perspective to their assigned perspective. This is in contrast to to the participants in the group of car passengers, who viewed the visualizations from their assigned perspective.

The question whether an improved survey design changes the outcome of the survey merits an investigation. To fully capture the moral intuition of the involved participants in both the stages of decision – defining the cost structure and the resulting outcomes, respectively – should be part of any future survey process, to ensure that the participants are aware of the consequences of individual confusion cost determinations. Further, by making explicit the underlying confusion cost structure in decision rules, one introduces the opportunity not only to adapt and incorporate human ethical intuition but also go beyond it and improve it. Future research should consider the possibility to include more differentiated classes in the analysis to allow, following recommendation 5 of the EC expert group, to rectify existing inequalities where vulnerable groups are involved.

Besides the analysis of deviations from the Bayes decision rule, one should also incorporate further values in the analysis. Any choice of the confusion cost matrix based on a participatory approach needs to be systematically balanced with other important values like practicability (see Section IV-D for a first step in that direction) and speed. Note that this is not a balancing of ethical decisions with non-ethical ones. As already mentioned, practicability is a precondition for the introduction of self-driving cars in the first place, which in turn can lead to an overall reduction of harm and so is part and parcel of the moral calculus. The moral dimension of these often technical features is not always apparent from the get-go and interdisciplinary work may allow to reveal the moral aspects of these technical features. This kind of moral evaluation of technical features, the systematic comparison of them, and the incorporation of other value-laden features should be considered in future research.

In the present study, we used static traffic images to compare human with robotic perceptions based on the different confusion judgments. Building on this, future studies should investigate human confusion judgments under real-life conditions in a simulation-based environment, such as e.g. presented in [60]. The traffic scenarios to be evaluated are dynamic in nature, accordingly, a simulation-based environment could help participants to put themselves in the complex situation, to estimate consequences of confusions and to make immediate decisions authentic. We therefore recommend complementing the insights drawn from the self-reported data of this study with behavioral data from a simulation-based environment in future research.

In general, the understanding and the shaping of relations between normative ethics, technical norms, the public's opinion, and design decisions for AI in safety relevant applications, such as automated driving, pose interesting challenges for the future. This equally applies to research, academic, and political debate.

References

[1] D. Bundesregierung, "Strategy artificial intelligence of the federal government of germany (in german)," *Berlin, November*, 2018. [Online]. Available: https://www.bundesregierung.de/resource/blob/975226/1550276/3f7d3c41c6e05695741273e78b8039f2/2018-11-15-ki-strategie-data.pdf?download=1 1

[2] C. Allen, I. Smit, and W. Wallach, "Artificial morality: Top-down, bottom-up, and hybrid approaches," *Ethics and information technology*, vol. 7, no. 3, pp. 149–155, 2005. 1, 18

[3] W. Wallach and C. Allen, *Moral machines: Teaching robots right from wrong.* Oxford University Press, 2008. 1, 18

[4] S. Vallor, *Technology and the virtues: A philosophical guide to a future worth wanting.* Oxford University Press, 2016. 1, 18

[5] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on fairness, accountability and transparency.* PMLR, 2018, pp. 77–91. 2

[6] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, and I. Rahwan, "The moral machine experiment," *Nature*, vol. 563, no. 7729, pp. 59–64, 2018. 2, 21

[7] R. Chan, M. Rottmann, R. Dardashti, F. Hüger, P. Schlicht, and H. Gottschalk, "The ethical dilemma when (not) setting up cost-based decision rules in semantic segmentation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 6 2019. 2, 5, 7

[8] R. Chan, M. Rottmann, F. Hüger, P. Schlicht, and H. Gottschalk, "Application of Maximum Likelihood Decision Rules for Handling Class Imbalance in Semantic Segmentation," in *The 30th European Safety and Reliability Conference (ESREL)*, 2020. 2, 4, 6

[9] U. Di Fabio, M. Broy, R. Brüngger *et al.*, "Ethics commission automated and connected driving," *Federal Ministry of Transport and Digital Infrastructure of the Federal Republic of Germany*, 2017. [Online]. Available: https://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission.pdf?__blob=publicationFile 2, 6, 19, 20

[10] J.-F. Bonnefon, D. Černy, J. Danaher, N. Devillier, V. Johansson, T. Kovacikova, M. Martens, M. Mladenovic, P. Palade, N. Reed *et al.*, "Ethics of connected and automated vehicles: Recommendations on road safety, privacy, fairness, explainability and responsibility," 2020. [Online]. Available: https://op.europa.eu/en/publication-detail/-/publication/5014975b-fcb8-11ea-b44f-01aa75ed71a1/ 2, 5, 19, 20, 21

[11] U. D. of Transportation, "Automated driving systems - a vision for safety," 2017. [Online]. Available: https://www.nhtsa.gov/sites/nhtsa.gov/files/documents/13069a-ads2.0_090617_v9a_tag.pdf 2

[12] X. Zhu, Z. Gu, and Z. Wang, "Ethical challenges and countermeasures of autonomous vehicles," in *E3S Web of Conferences*, vol. 233. EDP Sciences, 2021, p. 04016. 2

[13] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 7, 11

[14] D. Brüggemann, R. Chan, H. Gottschalk, and S. Bracke, "Software architecture for human-centered reliability assessment for neural networks in autonomous driving," *11th IMA International Conference on Modelling in Industrial Maintenance and Reliability (MIMAR)*, 2021. 2, 11

[15] H. Gottschalk, M. Rottmann, and M. Saltagic, "Does redundancy in ai perception systems help to test for super-human automated driving performance?" *arXiv preprint arXiv:2112.04758*, 2021. 2

[16] L. Fahrmeir, A. Hamerle, and W. Häussler, *Multivariate statistische Verfahren (in German)*, 2nd ed. Walter De Gruyter, 1996. 4, 5, 6

[17] H. Douglas, *Science, policy, and the value-free ideal.* University of Pittsburgh Pre, 2009. 5

[18] H. E. Longino, *The fate of knowledge.* Princeton University Press, 2018. 5

[19] M. J. Brown, *Science and moral imagination: A new ideal for values in science.* University of Pittsburgh Press, 2020. 5

[20] J. B. Biddle, "On predicting recidivism: epistemic risk, tradeoffs, and values in machine learning," *Canadian Journal of Philosophy*, pp. 1–21, 2020. 5

[21] V. C. Müller, "Ethics of Artificial Intelligence and Robotics," in *The Stanford Encyclopedia of Philosophy*, Summer 2021 ed., E. N. Zalta, Ed. Metaphysics Research Lab, Stanford University, 2021. 5

[22] A. HLEG, "High-level expert group on artificial intelligence: Ethics guidelines for trustworthy ai"," *European Commission*, p. 1–39, 2019. 5

[23] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of ai ethics guidelines," *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, 2019. 5

[24] P. Lin. (2013) The ethics of saving lives with autonomous cars are far murkier than you think. 6

[25] A. Hevelke and J. Nida-Rümelin, "Responsibility for crashes of autonomous vehicles: An ethical analysis," *Science and engineering ethics*, vol. 21, no. 3, pp. 619–630, 2015. 6

[26] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning.* The MIT Press, 2012. 6

[27] M. Taylor, "Self-driving mercedes-benzes will prioritize occupant safety over pedestrians," *Car and Driver*, vol. Oct. 7, 2016. [Online]. Available: https://www.caranddriver.com/news/a15344706/self-driving-mercedes-will-prioritize-occupant-safety-over-pedestrians/ 6

[28] D. Moore *et al.*, *Introduction to the practice of statistics*, 8th ed. New York : W.H. Freeman and Company, a Macmillan Higher Education Company, 2014. 9

[29] Y. Zhu, K. Sapra, F. A. Reda, K. J. Shih, S. Newsam, A. Tao, and B. Catanzaro, "Improving Semantic Segmentation via Video Propagation and Label Relaxation," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 8856–8865. 11

[30] V. Dignum, *Responsible Artificial Intelligence - How to Develop and Use AI in a Responsible Way.* Springer International Publishing, 2019. 15

[31] R. Sprenger, "Empirische forschungsmethoden für die evaluation visueller fahrerassistenzsysteme im kraftfahrzeug (in german)," Ph.D. dissertation, Universität Paderborn, Fakultät für Kulturwissenschaften, 2008. 15

[32] K. Weick and K. Weick, *Sensemaking in Organizations*, ser. Foundations for Organizational Science. SAGE Publications, 1995. 15

[33] H. Hagras, "Towards human-understandable explainable ai," *Computer*, vol. 51, no. 9, pp. 28–36, 2018. 15

[34] A. D. Baddeley and G. Hitch, "Working memory," in *Pyschology of Learning and Motivation*, G. H. Bower, Ed. Academic Press, 1974, vol. 8, pp. 47–89. 15

[35] J. K. Choi and Y. G. ji, "Investigating the importance of trust on adopting an autonomous vehicle," *International Journal of Human–Computer Interaction*, vol. 31, no. 10, pp. 692–702, 2015. 15

[36] W. Samek and K.-R. Mueller, "Towards explainable artificial intelligence," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Lecture Notes in Computer Science*, W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Mueller, Eds. Springer, Cham, 2019, vol. 11700, pp. 5–22. 15

[37] J. Sweller, "Cognitive load theory, learning difficulty, and instructional design," *Learning and Instruction*, vol. 4, no. 4, pp. 295–312, 1994. 15, 16

[38] Y. Wang and G. Ruhe, "The cognitive process of decision making." *Int. J. Cogn. Informatics Nat. Intell.*, vol. 1, no. 2, pp. 73–85, 2007. 16

[39] V. Venkatesh and H. Bala, "Technology acceptance model 3 and a research agenda on interventions," *Decision Sciences*, vol. 39, pp. 273–315, 2008. 17

[40] C. Lee, C. Ward, M. Raue, L. D'Ambrosio, and J. Coughlin, "Age differences in acceptance of self-driving cars: A survey of perceptions and attitudes," in *Human Aspects of IT for the Aged Population. Aging, Design and User Experienc*, 2017. 17

[41] P. Foot, "The problem of abortion and the doctrine of the double effect," *Oxford Review*, vol. 5, pp. 5–15, 1967. 18

[42] J. Himmelreich, "Never mind the trolley: The ethics of autonomous vehicles in mundane situations," *Ethical Theory and Moral Practice*, vol. 21, no. 3, pp. 669–684, 2018. 18, 19

[43] P. Foot, "The problem of abortion and the doctrine of the double effect," *Oxford review*, vol. 5, 1967. 18

[44] J. J. Thomson, "Killing, letting die, and the trolley problem," *The Monist*, vol. 59, no. 2, pp. 204–217, 1976. 18

[45] F. M. Kamm, *The trolley problem mysteries*. Oxford University Press, 2015. 18

[46] P. Lin, "Why ethics matters for autonomous cars," in *Autonomous driving*. Springer, Berlin, Heidelberg, 2016, pp. 69–85. 18

[47] D. Leben, "A rawlsian algorithm for autonomous vehicles," *Ethics and Information Technology*, vol. 19, no. 2, pp. 107–115, 2017. 18

[48] P. Lin, "The robot car of tomorrow may just be programmed to hit you," in *Machine Ethics and Robot Ethics*. Routledge, 2020, pp. 507–509. 18

[49] N. J. Goodall, "Away from trolley problems and toward risk management," *Applied Artificial Intelligence*, vol. 30, no. 8, pp. 810–821, 2016. 18

[50] S. Nyholm and J. Smids, "The ethics of accident-algorithms for self-driving cars: An applied trolley problem?" *Ethical theory and moral practice*, vol. 19, no. 5, pp. 1275–1289, 2016. 18, 19

[51] G. Keeling, "Why trolley problems matter for the ethics of automated vehicles," *Science and engineering ethics*, vol. 26, no. 1, pp. 293–307, 2020. 18

[52] Y. E. Bigman and K. Gray, "Life and death decisions of autonomous vehicles," *Nature*, vol. 579, no. 7797, pp. E1–E2, 2020. 21

[53] M. Kochupillai, C. Lütge, and F. Poszler, "Programming away human rights and responsibilities?"the moral machine experiment" and the need for a more "humane" av future," *NanoEthics*, vol. 14, no. 3, pp. 285–299, 2020. 21

[54] A. E. Jaques, "Why the moral machine is a monster," *University of Miami School of Law*, vol. 10, 2019. 21

[55] H. Furey and S. Hill, "Mit's moral machine project is a psychological roadblock to self-driving cars," *AI and Ethics*, vol. 1, no. 2, pp. 151–155, 2021. 21

[56] J. Millar, "Technology as moral proxy: Autonomy and paternalism by design," *IEEE technology and Society Magazine*, vol. 34, no. 2, pp. 47–55, 2015. 21

[57] ——, "Ethics settings for autonomous vehicles," in *Robot ethics 2.0: From autonomous cars to artificial intelligence*, P. Lin, K. Abney, and G. Bekey, Eds. Oxford University Press Oxford, 2017, pp. 20–34. 21

[58] ISO, "Road vehicles – Functional safety," 2011. 21

[59] ——, "Road vehicles — Safety and cybersecurity for automated driving systems — Design, verification and validation," International Organization for Standardization, Standard, Dec 2020. 21

[60] K. Kowol, S. Bracke, and H. Gottschalk, "A-eye: Driving with the eyes of ai for corner case generation," 2022. 23

APPENDIX

A. Modified F-test to test for statistical Differences between two Confusion Cost Matrices

Let $x_{l,k,j,i} \in \{0,1,2,3,4,5,6\}$ denote a single entry from the survey data where the indices correspond to the following sets:

$$l \in \{1,2\} : \text{group index} \qquad\qquad k \in \{1,2,\ldots,6\} : \text{true target class index}$$
$$j \in \{1,2,\ldots,6\} : \text{confused class index} \qquad\qquad i \in \{1,2,\ldots,n_{l,k}\} : \text{provided answer index} ,$$

where $n_{l,k} \in \{1,2,\ldots,n\}$ denotes the number of provided answers from group l assessing confusions with target class k. Further, we denote the overall mean entry for the confusion of target class k with class j by

$$\bar{x}_{\cdot,k,j,\cdot} = \frac{1}{n}\sum_{i=1}^{n} x_{\cdot,k,j,i} \tag{8}$$

and the mean entry within group l for the confusion of target class k with class j by

$$\bar{x}_{l,k,j,\cdot} = \frac{1}{n_{l,k}}\sum_{i=1}^{n_{l,k}} x_{l,k,j,i} . \tag{9}$$

Hence, the mean sum of squared differences *between* groups is computed via

$$MS_B := \sum_{k=1}^{6}\sum_{j=1}^{6} n_{1,k}(\bar{x}_{1,k,j,\cdot} - \bar{x}_{\cdot,k,j,\cdot})^2 + \sum_{k=1}^{6}\sum_{j=1}^{6} n_{2,k}(\bar{x}_{2,k,j,\cdot} - \bar{x}_{\cdot,k,j,\cdot})^2 \tag{10}$$

while the mean sum of squared differences *within* groups is computed via

$$MS_W := \sum_{k=1}^{6}\sum_{j=1}^{6}\sum_{i=1}^{n} \frac{(x_{1,k,j,i} - \bar{x}_{1,k,j,\cdot})^2}{\sum_{k=1}^{6} n_{1,k} + n_{2,k} - 60} + \sum_{k=1}^{6}\sum_{j=1}^{6}\sum_{i=1}^{n} \frac{(x_{2,k,j,i} - \bar{x}_{2,k,j,\cdot})^2}{\sum_{k=1}^{6} n_{1,k} + n_{2,k} - 60} \tag{11}$$

with 60 being the degrees of freedom (the number of entries in the two 6×6 confusion cost matrices that are not fixed, i.e. the off main diagonal entries). Then, the F-statistic

$$F := \frac{MS_B}{MS_W} = \frac{\text{between-groups variance}}{\text{within-groups variance}} \in \mathbb{R}_{\geq 0} \tag{12}$$

is defined as the ratio of *between* groups variances to *within* groups variances. We apply a bootstrapping method in order to approximate a *p-value*

$$p = \frac{\#\{F_{random} > F\}}{\#\{F_{random}\}} = \frac{\#\{F_{random} > F\}}{S} \in [0,1] \tag{13}$$

with $F_{random} \in \mathbb{R}_{\geq 0}$ the random F-statistic (according to Eq. (12)) after shuffling the group affiliations of the survey data and $S \in \mathbb{N}$ the number of shuffling steps. The lower the p-value the more likely the investigated groups' means differ.