THE ISSUE OF TRUSTWORTHINESS IN THE HA-AV MULTI-AGENT SYSTEM

SILVIYA SERAFIMOVA

Institute of Philosophy and Sociology, Bulgarian Academy of Sciences

silvija_serafimova@yahoo.com

Abstract

In this article, I aim at exploring the reasons behind the preference to AV reliability over AV trustworthiness. The investigation is focused upon the justification of AV reliability within the framework of two types of trustors (government and manufacturer) and one type of trustee (AV driver (owner)). Based upon an analysis of the conditions under which the two types of trustors can build a relationship of mutual trust, as well as clarifying the requirements of making the trustee *sufficiently* trust the trustors that the AV is *sufficiently* reliable, I point out the moral challenges to the implementation of the so-called decide-to-delegate model. Specifically, AV reliability and associated distributed responsibilities and liabilities are refracted through the lens of the issue *who, how and under what circumstances* can take back control of an AV in the event of an accident.

Keywords: HA-AV multi-agent systems, mixed-trust relationships, AV trustworthiness and reliability, decide-to-delegate model

Introduction

Conceptual clarifications

Buechner and Tavani’s model of trust in multi-agent systems includes humans, groups of humans and artificial agents “such as intelligent software agents and physical robots” (Tavani, 2015: 79; Ryan, 2020: 2763). This “diffuse/default model of trust” may be applied to AI, since it allows a distribution of responsibility “over a diverse network of human agents *and* artificial agents” (Ryan, 2020: 2763). Specifically, the kinds and degrees of HA (human agent)-AA (artificial agent) trust relationships in the so-called zones of default trust and zones of diffuse, default trust [1] can vary depending upon the autonomy of the particular artificial agents (Buechner et al., 2014: 66). That is why mixed-trust relationships, or multi-agent trust relationships “may take the form of trusting groups of individuals, organizations, and perhaps AI technologies within that network of trust” (Ryan, 2020: 2763).

When by multi-agent systems one understands a broad network of human and artificial agents, one's strive for enriching the idea of trust necessitates the enrichment of that of responsibility in terms of its distribution. In other words, the impact of different agents encourages the establishment of some multilateral (*multi-agent*) relationships. [2] Considering that the relationships in question require the interaction of more complex types of agents, one cannot argue for one and the same level of trustworthiness, nor can one assume that agents are responsible in one and the same manner when becoming trustworthy.

As Buechner et al. point out, regardless of being ineligible for moral agents, the degree or level of trustworthiness for every AA can be evaluated in terms of its reliability and competence (Buechner et al., 2014: 78). While a person can be treated as responsible because their agency is determined by their own moral motivation, AI cannot be defined as such, since it does not have moral motivation at all. Therefore, while a person can be labeled as trustworthy, AI can be defined as merely reliable.

Consequently, arguing for responsible AI is possible only by analogy with human responsibility, namely, when one adopts the analogical principle of distributed responsibility. Extrapolating the debate to the relations between trust and blame, one can claim that while a person can breach trust and then be blamed for that, AI cannot. Specifically, AI cannot be blamed because it can only cause disappointment with its failure. [3] As Ryan cogently argues, "there is no reason to state that AI has the capacity to be trusted, simply because it is being used or is making decisions within a multi-agent system" (Ryan, 2020: 2764).

Objective

Looking for an exemplification of the moral challenges to AI trustworthiness, I tackle the issue of AV reliability, as underlain by a multi-agent system with two different types of trustors (government and manufacturer) and one type of trustee (AV driver (owner)). [4] First, I analyze the conditions under which the two types of trustors can build a relationship of mutual trust and second, I clarify those under which the trustee should *sufficiently* trust the trustors that the AV is *sufficiently* reliable.

The analysis of the distribution of trustworthiness/reliability follows that of clarifying the issue of *who, how and under what circumstances* can take back control of an AV. The control problem is refracted through the lens of the so-called (by Floridi et al.) meta-autonomy or “decide-to-delegate” model. The model itself sheds light upon the reasons behind the requirement that people should always retain the power “to *decide which decisions to take*” (Floridi et al., 2018: 698).

In this context, the search for AV reliability is coupled with the moral concerns about the HA-AV trust relationship, taking into account that the one who should decide and take the major responsibility is not necessarily the one who 1) is trusted by the trustee and 2) should be trusted by the trustee. Elaborating upon the outcomes of the aforementioned two scenarios, I aim at revealing not only the negative implications of misplaced trust, but also the positive outcomes regarding its disclosure. For the purposes of exemplifying the complex role of interpersonal trust for AI reliability, I compare and contrast the different expectations of AV reliability by analyzing the impact of the so-called operational responsibility and distributed trust.

The role of trustworthiness for the HA-AV multi-agent system

Similar to the problems with AI trustworthiness, AV reliability requires one to analyze the trustworthiness of each of the agents in the HA-AV multi-agent system. However, the main difference is that in the AV field, one should examine trustworthiness of more than one type of trustors—two types of trustors (government and manufacturer) having different responsibilities and associated liabilities with respect to the trustee (the AV driver (owner)). [5] Consequently, the trustee should build a complex vision of trust which is inevitably grounded in the successful or unsuccessful distribution of trustors’ responsibilities.

According to the ideal model of a HA-AV multi-agent system, the two types of agent-trustors (the government and the AV manufacturer) should be recognized as sufficiently trustworthy. This means that their trustworthiness should be non-contradictorily justified not only in rational, but also in affective and normative terms on an interpersonal level. [6] In addition, the government and the manufacturer should also trust each other in the process of guaranteeing AV reliability for the AV driver (owner) as a trustee. In turn, the next requirement concerns the role of the AV driver (owner) as a trustee. They should *sufficiently* trust the government and/or the manufacturer that the AV is *sufficiently* reliable, since the justification of an AV as reliable has an impact on one’s purchase decision. [7]

However, trustworthiness of the trustors and that of the trustee have not only mutually complementary embodiments, but also mutually exclusive manifestations. For instance, the AV user – as a citizen – may trust the government more than the manufacturer, believing that the former is more responsible, or vice versa. In turn, distributed responsibility, which assumes the elaboration of different degrees of trust on the trustee's part, determines their different expectations of liability. [8] The AV driver can make liability claims to the government rather than to the manufacturer (or vice versa) due to the implications of the responsibility of the given trustor.

Judging by these clarifications, one may argue that interhuman trust in such a complex multi-agent system as that of the HA-AV is a matter of distributed trust. The latter is a result of the distributed responsibilities of the trustors, as well as being underlain by the different degrees of the distributed expectations on the trustee's part regarding the aforementioned responsibilities and AV's operational responsibility. [9] In addition, distributed responsibilities, which necessitate differentiated trust on the trustee's part, affect the complication of the concept of AV reliability.

Specifically, AV reliability is defined in terms of the epistemic non-controversy of crash-optimization algorithms (when the AV is designed as optimally safe), as well as in terms of the moral biased preferences of the potential AV owner (when they buy it for themselves or provide a theoretical evaluation of AV reliability). [10]

Some moral implications of trust for the "decide-to-delegate" model

Practically speaking, the distribution of trust/reliability raises the question of *who, how and under what circumstances* can take back control of an AV. Theoretically speaking, the issue addresses the so-called (by Floridi et al.) meta-autonomy or "decide-to-delegate" model. According to this model, "people should always retain the power to *decide which decisions to take*", [11] "exercising the freedom to choose where necessary", and ceding it in cases where overriding reasons (such as efficacy) "may outweigh the loss of control over the decision-making process" (Floridi et al., 2018: 698). [12]

Taking back control of an AV requires one to adopt the "decide-to-delegate" model due to which the delegation of trust to the manufacturer and the driver is recognized as a matter of building an asymmetric relationship. The latter should be mediated by AV reliability itself. One should also keep in mind that the justification of the aforementioned reliability is not determined by demonstrating an AV's technologically facilitated safety alone. Specifically, the recognition of AV reliability is underlain by the fulfillment of some crucial moral conditions as well. As such a condition, I would point out people's attitudes of "attributing more permission and less fault to AIs (vs. humans)" (Schank et al., 2019; Giroux et al., 2022: 3). This means that the higher degree of moral permissibility of AV actions in the event of an accident is considered to be triggered by people's attribution of responsibility to AVs as non-humans (Gill, 2020; Giroux et al., 2022: 4). [13]

Theoretical concerns about AV reliability

Theoretically speaking again, the question is whether or not one's indirect responsibility for a given outcome, which is evaluated differently due to the different degrees of significance of its particular consequences, gives us a sufficiently moral reason to recognize the source of responsibility as reliable, when judging by the outcomes alone.

Such a concern can be exemplified by conducting a genealogical analysis of the major components of responsibility and trust/reliability. The analysis begins with the inflicted injuries as an outcome. The AV cannot decide by itself, regardless of its high level of autonomy. Therefore, it cannot be blamed even for its technical failure, since it is reliable, but not trustworthy in interhuman terms. How can we identify who is responsible in the event of an AV accident? A possible suggestion would be the one who *should decide/should have decided* that the risk of using the AV is morally acceptable. [14] In turn, such a specification requires the technical, moral and existential acceptability of AV use to be in line with the criteria of moral permissibility.

In the context of distributed responsibility, the answer to the question "Who should decide?" is closely tied with that to the question "Who should be trusted and what should be relied upon?". I would argue that the moral concerns about the HA-AV multi-agent system derive from the conditions under which the one who should decide and take the major responsibility is not necessarily the one who 1) is trusted by the trustee and 2) should be trusted by the trustee.

The two cases are of crucial importance for understanding the issue of misplaced trust. For instance, the manufacturer is the one who takes the major responsibility for the AV production and that is why they should be trusted by the AV owner as a trustee. [15] However, the AV owner may unconditionally trust the government and delegate their trust to the manufacturer just because the government encourages such a delegation by providing the manufacturer with the so-called safety protocols (the first case). [16] Thus, the AV driver may buy an unreliable AV just because they delegate their trust to the one who has the real major responsibility (the manufacturer) as a matter of derivative trust. Regardless of the fact that the manufacturer is not initially trusted, the delegated trust (also due to the role of authority bias [17] and reversed proximity bias [18]) may negatively affect one's purchase decision.

Consequently, decisions regarding the government-manufacturer trust relationship can be a result of some different factors such as overtrust in AV reliability on the government's part. The reason might be the good reputation of the manufacturer, some collaborations which benefit one or both trustors, etc. [19] The AV owner's conscious and unconscious biases should also be taken into consideration – for instance, one should analyze the AV owner's ungrounded trust in institutions underlain by the role of authority bias (again), the role of informed consent, etc.

The analysis of these specifications necessitates one to provide an exemplification of the aforementioned second case of misplaced trust. For instance, the misled AV owner trusts the manufacturer as having the major responsibility for the production of AVs, when taking for granted that the government serves the public interest in guaranteeing AV safety. [20] However, misplaced trust itself may result from different factors such as the AV owner's lack of awareness of the commercial agreements between the government and the manufacturer.

The examined two cases may have not only negative, but also positive hypothetical representations. For instance, the AV owner may trust the manufacturer with the AV reliability, although the government has the major responsibility for risk-based regulation in the field of AV manufacturing (the first case). The government may deliberately hide some significant risks from potential clients because it has already signed an international agreement on encouraging the sale of these unsafe AVs. However, the manufacturer is aware of the risk and either refuses to produce unsafe AVs or attempts to provide the potential client with all the necessary information in the strive for informed consent on the client's part. Certainly, this is a scenario which is highly improbable in practice, unless the manufacturer works in a philanthropic company which does not care about profit at all.

In turn, as a ‘positive’ exemplification of the second case, I would point out the disclosure of misplaced trust. For instance, when the AV driver realizes that the manufacturer deliberately hides some AV problems, their trust in the manufacturer can be ascribed a positive value merely by its disenchantment as a matter of mistrust. Practically speaking, when the AV driver realizes that they have wrongly trusted the manufacturer, they may refuse to buy not only a particular AV, but any of the provided AVs, and may even persuade other people to refrain from buying said AVs. Certainly, such trust in mistrust thoroughly reframes the vision of AV reliability. Specifically, AV reliability is recognized as a process of disclosure of its unreliability, whenever necessary.

Practical dilemmas for AV reliability

The issue of AV reliability has crucial practical implications in moral terms which concern operational responsibility and distributed trust. Frank et al. argue that future research should focus upon “the more practical and more likely” dilemmas that occur in everyday life, including the degree to which the AVs “should be able to be more aggressive when maneuvering in high traffic conditions or overtaking slow vehicles” (Frank et al., 2019: 17). [21] In addition, they should speed to a destination if the passenger is seriously ill or needs to catch a connection (Ibid.). However, the adaptation to the preferences or commands of its user (Ibid.) displays only one variant of the possible technological adaptations of AVs. [22] Regarding the case when the AV should notify the operator about the potential accident and the associated course the operator wants to perform, the owner has the following two options: the AV can be either preprogrammed or can it be reprogrammed through the use of ethical principles according to the beliefs of the owner (Gurney, 2016: 254). [23]

Examining the role of responsibility as refracted through the lens of trust, Gurney cogently points out that people do not have “perfectly utilitarian or Kantian” views (Ibid.), which means that choosing an ethical paradigm does not trigger a given framework of trust by default. This statement requires two main clarifications to be provided, at least. First, similar to moral pluralism regarding the prior choice of ethics of crash-optimization algorithms, one faces pluralism of trust principles which do not necessarily contribute to precising the HA-AV trust relationships. If one opts for ‘mixed’ ethical principles for their AV, it is difficult to introduce a risk-aware motion planning which makes room for permissible killing in the event of an accident. [24] Consequently, one cannot rely upon such an AV neither as a driver (they will not trust it if there is still a risk that the AV can sacrifice them, regardless of its being generally programmed with an ethics which requires passenger protection) nor as a consumer (they can choose a manually steered car instead of an AV assuming that they will have a constant control of the vehicle). [25] In addition, even if the two vehicles are identical in technical terms, the passengers experience different risks on the basis of demographics including their age and gender, whether or not the passengers travel alone, whether or not they are drunk or sober etc. (Goodall, 2014: 62).

Based upon this analysis, one may argue that the more control variables that are taken into careful consideration, the more specific the situation is. For instance, an AV should certainly speed up if its passenger is in a serious condition, rather than if they need to catch a connection. But what can happen if there is another vehicle on the road which transports another seriously ill passenger? Speaking in terms of autonomy, both vehicles should be sufficiently autonomous to increase their speed, but from a moral standpoint it is difficult to say which one should be given way. Furthermore, if we accept that one of the vehicles should give way, this would mean that both of them should be presented with information about the passengers' medical condition, as well as be able to compare and contrast the potential outcomes of their decisions depending on different control variables. Otherwise, the AVs cannot interpret the information properly (in minimum time) and then evaluate it in moral terms.

The gist of the challenge is how an AV, which is not a human agent, can share responsibility with another AV obtaining the full moral competence of a human being. Going back to the "binary fashion" of arguing that the AV "must share operational responsibility with a driver" (Pattinson et al., 2020: 2), [26] one should keep in mind that this responsibility does not meet the requirements of commonly shared responsibility in interhuman terms. The challenge is also strengthened by the understanding that intention "for moral behavior in human-machine interaction is lower than that in human-human interaction" (Giroux et al., 2022: 4). Tackling the issue in terms of the attribution of responsibility for harm means that drivers may be misled as moral agents or simply can be more predisposed to delegate responsibility to AVs. Therefore, the driver may delegate the responsibility to the AV as a right decision due to the decline of guilt which "mediates people's decreased morality toward AI" (Ibid.: 5). The debate is additionally complicated once we consider that if guilt stems from a violation of norms, its lack may disguise a certain human misbehavior as a matter of following rules of conduct.

Extrapolating Castelo et al.'s theory of algorithm aversion as underlain by AA's low ability in subjective tasks (Castelo et al., 2019), as well as Longoni et al.'s suggestion that AA neglects consumers' unique characteristics (Longoni et al., 2019; Gill, 2020: 287), I would argue that humans as moral agents feel a lack of guilt when delegating responsibility to AVs due to the lack of a moral feedback. In other words, the lower ethical requirements for the interaction between humans and AVs are grounded in the assumption that the latter cannot reciprocate in moral terms. Furthermore, the lack of moral reciprocation understood as a process of achieving moral mutuality raises the issue of how an AV can deal with the risk of diffusion of responsibility, considering that it does not have the potential to make moral decisions which only humans can.

Conclusion

Analyzing the specificities of the HA-AV multi-agent system, I outline the different embodiments of mutual trust between the government and the manufacturer as trustors and the AV driver (owner) as a trustee, including these of misplaced trust. Based on an investigation of the aforementioned specificities, I reach the conclusion that one can argue for AV reliability rather than AV trustworthiness. Consequently, the differences regarding the asymmetric relationships of trust in the multi-agent system in question are recognized as originating from the associated asymmetric relations of autonomy. Specifically, the higher the autonomy, the more asymmetric trust is in functional terms. However, this relation between trust and autonomy also affects the justification of the self-protective functions of trust in a negative sense. That is why the investigation of the HA-AV multi-agent system should also take into account the associated processes of distributing responsibilities and liabilities.

As an illuminative embodiment of distributed trust, I point out the case when the driver should take back control of an AV in the event of an accident. The decision is evaluated through the lens of what Floridi et al. call "decide-to-delegate" model. On a theoretical level, I reach the conclusion that the introduction of the "decide-to-delegate" model to AV scenarios requires one to trace back the implications of the mutual breach of interpersonal trust and technological failure of AV reliability. In such a complicated multi-agent system as that of HA-AV, Buechner et al.'s requirement of the trustee's trustworthiness should be coupled with that of the trustors' and trustee's guilt. If there is a breach of trust, the relation between the questions "Who should be (is) *responsible*?" and "Who should be (is) *trustworthy*?" is extrapolated to that of "Who should be (is) *blamed*?". If an agent is *untrustworthy*, this is probably a result of being *irresponsible*. However, it does not follow that such an agent *should be blamed* by default because being untrustworthy might not be their *fault*.

I also specify that arguing for shared operational responsibility in the AV discourse is a difficult task in moral terms. Taking into account the challenges in properly calculating the risks, and then providing criteria for reliability regarding safety, one cannot agree to call the respective calculations responsible. The assumption is that every single crash-optimization algorithm triggers a corresponding duty or duties which the agents should take up as a matter of trusting those who create the algorithms. However, even in operational terms, neither the driver nor the AV can take up such duties by default. The reason is the complexity, in the first case, and the lack, in the second one, of their own moral and existential motivation.

In this context, the major moral concern is how one can deal with different AVs having mixed ethical principles when an accident is unavoidable. The challenge is moral rather than purely computational, since it is triggered by the initial lack of moral and existential motivation on the AV's part. Specifically, the difficulty stems from the fact that one can argue for AV's operational responsibility only by analogy with human responsibility (with all of its imperfections).

NOTES

[1] Buechner et al. borrow Margaret Urban Walker's concepts of zones of default trust and zones of diffusely distributed trust. Elaborating upon these concepts, they argue that trust relationships between humans and artificial agents are possible "in virtue of various contexts or zones of trust" (Buchner et al., 2014: 66) in which they interact. Specifically, human agents are "capable of entering into trust relationships with a wide range of AAs", including multi-agent systems, that can be "diffusely distributed" across various elements in a system or network of agents (Ibid.).

[2] When arguing for multi-agent trust relationships, Ryan provides a distinction between mutually related interpersonal and institutional trust on the one hand, and reliance regarding AI and other technologies, on the other one (Ibid.: 2764).

[3] Cf. Nickel et al., 2010.

[4] The choice of agents is determined by the results of the Open Roboethics Initiative Survey (2015). See Gurney, 2016: 252-259; Serafimova, 2022: 202-231.

[5] For a detailed analysis of moral responsibility and legal liability of the driver, manufacturer and government see Gurney, 2016: 252-260; Serafimova, 2022: 208-236.

[6] Regarding the differences between interhuman trust and AI's lack of trustworthiness, I refer to Ryan's classification of the three accounts of trust. The classification involves 1) a rational account of AI (which displays a form of reliance) 2) an affective account of AI (lacking the motivation of AI to do something based on a goodwill towards the trustee) and 3) a normative account of AI (which lacks its commitment to the relationship with the particular trustee) (Ryan, 2020: 2752-2753). Considering that trust is built around the motives and intentions of the entrusted, misrecognizing trust in AVs as a matter of interpersonal trust "may involve the wrong type of prescriptive account" (Chilson, 2022: 230).

[7] For one's reasons behind welcoming "utilitarian self-sacrificing" AVs on the road "without actually wanting to buy one for themselves" see Bonnefon et al., 2016: 1573-1576. In addition, one's decreased willingness of buying can also be explained by the role of the so-called algorithm aversion. According to the latter, humans as consumers can be averse to an autonomous agent despite of its superior performance (Gill, 2020: 273).

[8] The issue of legal liability is not a prime objective of exploration in this article. However, its implications in the AV discourse should be examined as closely tied with the more general discussions about legal personhood for AI (Bertolini and Episcopo, 2022). More specifically, the potential ascription of legal personhood to AVs triggers moral concerns about the so-called liability management opportunities (Bryson et al., 2017: 286). These concerns address the way in which the distribution of operational responsibility can be misinterpreted as a process of diffusion of general responsibility. Thus, liability management opportunities can encourage humans to use AVs in order to "insulate themselves from liability" in the event of an accident (Ibid.: 285).

[9] Such an operational responsibility, underpinned by the "dualistic 'driver in charge' or 'vehicle in charge' dichotomy" (Pattison et al., 2020: 6), is additionally complicated by the fact that an AV cannot take responsibility by itself due to not meeting the requirements of being considered a moral agent. In turn, the most apparent concerns about commonly shared operational responsibility in legal terms can be found in the way in which it can encourage the introduction of the so-called liability shield, namely, when a "natural person" uses an "artificial person" to shield themselves from the outcomes of their own conduct (Bryson et al., 2017: 286).

[10] There is a correspondence between people's moral biases and the process of self-projection regarding one's self-preserving behavior. When valuing themselves and their loved ones, the participants in the AV thought experiments show preference towards a deontological moral doctrine that rejects the idea of sacrificing the passengers in their vehicles. However, the same participants seem to employ a utilitarian doctrine of saving the majority of potential victims in an AV accident when they do not identity with the affected minority which should be sacrificed

(Rand, 2016; Schariff et al., 2017; Frank et al., 2019). For the impact of imagined vs. real harm upon human decisions see Gill, 2020.

[11] The definition of "the decide-to-delegate model" partially meets Giubilini and Savulescu's definition of AI as a moral adviser. According to the latter definition, it is the agent who decides whether or not to accept the AI-based moral advice (Giubilini and Savulescu, 2018: 174). However, such an artificial adviser can function as an assistant rather than an adviser in moral matters. Extrapolating the debate to the AV discourse, it would mean that "the decide-to-delegate model" makes room for an AV as a facilitator in building interhuman trust due to the limited conditions of shared operational responsibility.

[12] "The decide-do-delegate model", as applied in the AV field, is characterized by what Shaefer et al. describe as a framework of some emotive factors, such as confidence in the automation, attitudes towards the automation, commitment to and comfort with automation (Ashleigh and Stanton, 2001; Shaefer et al., 2016). They point out that this type of confidence, which must not be confused with self-confidence, can be considered an antecedent of trust (Shaefer et al., 2016).

[13] The practical implications of this outcome can be revealed by referring to Gill's investigation. According to the latter, when driving a car with a manual steering system, people used to avoid harming a pedestrian; the assumption is that they have complete control over their actions (Gill, 2020: 273). On the other hand, when an AV is in control, a higher proportion of these people chose to harm the same pedestrian (Ibid.).

[14] The issue of risk is closely tied with the so-called blame problem regarding the role of responsibility in a multi-agent system when an accident is unavoidable. The problem in question can be described as triggered by the so-called responsibility gap, namely, when "humans are forced to compensate damages for which they have no or very limited control" (Bertolini and Episcopo, 2022: 5). The limited control itself is recognized as a necessary and sufficient condition of delegating responsibility whose major objective is the reduction of legal liability. That is why humans delegate responsibility to AVs even when they are aware that the process of delegation displays diffusion of responsibility.

[15] This assumption is grounded in the so-called trust-comparison index for AI (Floridi, 2018: 703), which indicates the establishment of trust-labels that should guarantee the coupling of AI products' technological reliability with the social benefits of their use. For the concerns about the moral reliability of trust-comparison index regarding AVs see Serafimova, 2022: 217-224.

[16] Safety protocols display an attempt at combining technological and moral regulations of AVs. Government safety protocol can implement ethical rules into algorithms so that the results can guarantee immunity for the manufacturer (Gurney, 2016: 258). Practically speaking, the government should also be restricted to some "general principles with the manufacturer gap filling through bottom-up approaches to ethics" (Ibid.: 259). The outcome is that "the benefits of bright-line rules" should be supplied with "the completeness of utilitarian approaches" (Ibid.). For the reasons behind the asymmetric minimization of manufacturer's liability, as

evaluated in terms of governmental liability, see Serafimova, 2022: 229-231. For the implications of the co-regulatory model of data governance, justifying the process of co-regulation as a "legal link" between top-down approaches and bottom-up solutions of self-regulation, see also Pagallo, 2022: 169-172.

[17] Trusting a government with AV operations just because it is a government and, as such, it should protect drivers' interests in the best possible way, may trigger misplaced trust as a result of a driver's authority bias. The latter derives from the driver's unrealistic or wrong beliefs that the government 'knows best' what is good for them.

[18] I would argue for the reversion of the so-called proximity bias according to which remote workers are less productive than team members. In the context of AVs, mistrust in government can be explained by the coupling of two biases–the reversed proximity bias according to which remote agents are evaluated as more productive than close agents, and the authority bias regarding one's trust in highly autonomous institutions.

[19] Another significant moral concern is raised by Bonnefon et. al. They argue that the manufacturer may engage in advertising and lobbying in order to influence the preferences of the consumer and the regulations that the government may pass. Then, "a critical collective problem consists of deciding whether governments should regulate the moral algorithms that manufacturers offer to consumers" (Bonnefon et al., 2016: 1575).

[20] For the moral challenges of implementing the triplet *trust in government-public trust-public interest* in the AV discourse see Serafimova, 2022: 225-229.

[21] The scenario of colliding AVs is one relatively simple scenario. What, however, deserves special attention is the justification of "a principle of fairness in the distribution of the unavoidable risks on the road" under circumstances involving hybrid traffic (Berkey, 2022: 211). The latter affects not only AVs, but also human-driven vehicles, motorcyclists, bicyclists and pedestrians.

[22] The idea is explicitly stated by Gurney who suggests that the AV driver may select "the ethics of her car" prior to using it in practice (Gurney, 2016: 254). This suggestion marks the beginning of a broad discussion about the implementation of different ethical principles to AVs, as well as tracing the practical outcomes of such a 'moral' design. For instance, the application of the principle of ethical selection underlies the so-called ethical knob assuming that consumers can "customize based on their own moral proclivity" (Contissa et al., 2017).

[23] For some potential unintended consequences of AVs see also Gurney, 2022: 147-152.

[24] For instance, Formosa examines a hybrid model which combines the benefits of the so-called PES (personal ethics setting) and MES (mandatory ethics setting) in the AV discourse. This model promises that people "*can* change their AV's ethical settings" (Formosa, 2022: 184), maintaining the freedom of choice which includes the option of a selfish choice. However, the latter type of choice can be reduced by adopting MES; specifically, people "will keep the

mandated default choice settings that we collectively judge to be best, thereby avoiding a race to the selfish bottom that might follow from a PES…" (Ibid.).

[25] For the different ways of building trust between manufacturers and consumers regarding the way in which the choice of crash-optimization algorithms is expected to bring about the best results of AV driving behavior, see Bennett, 2022; Berkey, 2022 and Chilson, 2022.

[26] The investigation is focused upon the legal implications of operational responsibility, namely, how shared operational responsibility delineates the legal liability of the vehicle and the driver in the event of an accident (Ibid.: 1). However, most issues regarding legal liability are also applicable to some purely moral aspects of the problem such as the role of informed consent, the risks of situational awareness, as well as the lack of trust.

REFERENCES

Ashleigh, M. J. and N. A. Stanton. (2001). Trust: Key Elements in Human Supervisory Control Domains. — In: *Cognition Technology and Work,* 3(2), 92-100. https//org. doi//10.1007/PL00011527.

Bennett, S. (2022). Algorithms of Life and Death: A Utilitarian Approach to the Ethics of Self-Driving Cars. — In: *Autonomous Vehicle Ethics: The Trolley Problem and Beyond*. Ryan Jenkins, David Černý and Tomáš Hříbek (eds.). Oxford, Oxford University Press, pp. 191-209.

Berkey, **B.** (2022). Autonomous Vehicles, Business Ethics, and Risk Distribution in Hybrid Traffic. — In: *Autonomous Vehicle Ethics: The Trolley Problem and Beyond*, Ryan Jenkins, David Černý and Tomáš Hříbek (eds.). Oxford, Oxford University Press, pp. 210-228.

Bertolini, A. and F. Episcopo. (2022). Robots and AI as Legal Subjects? Disentangling the Ontological and Functional Perspective. — In: *Front Robot AI,* 9, 842213. https://doi.org./10.3389/frobt.2022.842213.

Bonnefon, J.-F., Shariff, A. and I. Rahwan. (2016). The Social Dilemma of Autonomous Vehicles. — In: *Science*, 352, 1573-1576. https://doi.org/10.1126/science.aaf2654.

Bryson, J. J., Diamantis, M. E. and T. D. Grant. (2017). Of, for, and by the People: The Legal Lacuna of Synthetic Persons. — In: *Artificial Intelligence and Law*, 25, 273-291.

Buechner, J., Simon, J. and H. T. Tavani. (2014). Re-Thinking Trust and Trustworthiness in Digital Environments. — In: E. Buchanan et al. (eds.). *Autonomous Technologies: Philosophical Issues, Practical Solutions, Human Nature: Proceedings of the Tenth*

International Conference on Computer Ethics—Philosophical Inquiry: CEPE 2013. Menomonie, WI: INSEIT, pp. 65-79.

Castelo, N., Bos, M. W. and D. R. Lehmann. (2019). Task-Dependent Algorithm Aversion. — In: *Journal of Marketing Research*, 56(5), 809–825.

Chilson, **K.** (2022). An Epistemic Approach to Cultivating Appropriate Trust in Autonomous Vehicles. — In: *Autonomous Vehicle Ethics: The Trolley Problem and Beyond*. Ryan Jenkins, David Černý and Tomáš Hříbek (eds.). Oxford, Oxford University Press, pp. 229-242.

Contissa, G., Lagioia, F. and G. Sartor. (2017). The Ethical Knob: Ethically-Customizable Automated Vehicles and the Law. — In: *Artificial Intelligence and Law*, 25(3), 365-378.

Floridi, L., Cowls, J., Bertrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Shafer, B., Valcke, P. and E. Vayena. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. — In: *Mind and Machines*, 28, 689-707. https//doi.org/10.1007/s11023-018-9482-5.

Formosa, P. (2022). Autonomous Vehicles and Ethical Settings: Who Should Decide? — In: *Autonomous Vehicle Ethics: The Trolley Problem and Beyond*. Ryan Jenkins, David Černý and Tomáš Hříbek (eds.). Oxford, Oxford University Press, pp. 176-190.

Frank, D.-A., Chrysochou, P., Mitkidis, P. and D. Ariely. (2019). Human Decision-Making Biases in the Moral Dilemmas of Autonomous Vehicles. — In: *Sci Rep*, 9, 13080. https://doi.org/10.1038/s41598-019-49411-7.

Gill, T. (2020). Blame It on the Self-Driving Car: How Autonomous Vehicles Can Alter Consumer Morality. — In: *Journal of Consumer Research*, 47(2), 272-291.

Giroux, M., Kim, J., Lee, J. C. and J. Park. (2022). Artificial Intelligence and Declined Guilt: Retailing Morality. Comparison between Human and AI. — In: *Journal of Business Ethics*, 1-15. https://doi. org/10.1007/s10551-022-05056-7.

Giubilini, A. and J. Savulescu. (2018). The Artificial Moral Advisor. The "Ideal Observer" Meets Artificial Intelligence. — In: *Philosophy & Technology*, 31(2), 169–188. https://doi.org/10.1007/s13347-017-0285-z.

Goodall, N. (2014). Ethical Decision Making during Automated Vehicle Crashes. — In: *Transp. Res. Record J., Transp. Res. Board*, 2424, 1, 58-65. https://doi.org/10.3141/2424-07.

Gurney, J. K. (2016). Crashing into the Unknown: An Examination of Crash-Optimization Algorithms through the Two Lanes of Ethics and Law. — In: *Albany Law Review,* 79(1), 183-267.

Gurney, J. K. (2022). Unintended Externalities of Highly Automated Vehicles. — In: *Autonomous Vehicle Ethics: The Trolley Problem and Beyond*. Ryan Jenkins, David Černý and Tomáš Hříbek (eds.). Oxford, Oxford University Press, pp. 147-158.

Longoni, C., Bonezzi, A. and C. Morewedge. (2019). Resistance to Medical Artificial Intelligence. — In: *Journal of Consumer Research*, 46(4), 629–650.

Nickel, P. J., Franssen, M. and P. Kroes. (2010). Can We Make Sense of the Notion of Trustworthy Technology? — In: *Know Tech Pol*, 23, 429-444.

Pagallo, U. (2022). The Politics of Self-Driving Cars: Soft Ethics, Hard Law, Big Business, Social Norms. — In: *Autonomous Vehicle Ethics: The Trolley Problem and Beyond*. Ryan Jenkins, David Černý and Tomáš Hříbek (eds.). Oxford, Oxford University Press, pp. 159-175.

Pattinson, J.-A., Chen, H. and S. Basu. (2020). Legal Issues in Automated Vehicles: Critically Considering the Potential Role of Consent and Interactive Digital Interfaces. — In: *Humanities and Social Sciences Communication*, 7, 153. https://doi.org/10.1057/s41599-020-00644-2.

Rand, D. G. (2016). Cooperation, Fast and Slow: Meta-Analytic Evidence for a Theory of Social Heuristics and Self-Interested Deliberation. — In: *Psychological Science*, 27(9), 1192-1206. https://doi.org/10.1177/0956797616654455.

Ryan, M. (2020). In AI We Trust: Ethics, Artificial Intelligence, and Reliability. — In: *Science and Engineering Ethics*, 26, 2749-2767. https//doi.org/10.1007/s11948-020-00228-y.

Serafimova, S. (2022). *Perspectives on Moral AI. Some Pitfalls of Making Machines Moral.* Sofia: Avangard Prima.

Shaefer, K. E., Chen, J. Y., Szalma, J. L. and P. A. Hancock. (2016). A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems. — In: *Human Factors,* 58(3), 377-400. https://doi.org/10.1177/0018720816634228.

Shank, D. B., De Santi, A. and T. Maninger. (2019). When Are Artificial Intelligence versus Human Agents Faulted for Wrongdoing? Moral Attributions after Individual and Joint Decisions. — In: *Information, Communication & Society*, 22(5), 648-663.

Shariff, A., Bonnefon, J.-F. and I. Rahwan. (2017). Psychological Roadblocks to the Adoption of Self-Driving Vehicles. — In: *Nature Human Behavior*, 1, 694-696. https://doi.org./10.1038/s41562-017-0202-6.

Tavani, H. T. (2015). Levels of Trust in the Context of Machine Ethics. — In: *Philosophy & Technology*, 28(1), 75-90.