Please notice that this is a DRAFT of a paper the final version of which has been published in the

International Journal of Technoethics
13(1), 2022, pp. 1-20
DOI: 10.4018/IJT.291553

Operationalising the Ethics of Connected and Automated Vehicles.
An Engineering Perspective.

Authors: Fabio Fossa, Stefano Arrigoni, Giandomenico Caruso, Federico Cheli, Hafeez Husain Cholakkal, Pragyan Dahal, Matteo Matteucci.

Introduction

Connected and Automated Vehicles [CAVs] are arguably one of the most researched and discussed applications of Artificial Intelligence [AI] technologies. Advances in design and development fuel the anticipation of a future where our roads will be populated by both regular and automated vehicles. Concurrently, social, ethical, and legal issues surrounding the impacts of CAV technologies have been raised (Nyholm, 2018a, 2018b; Taeihagh & Lim, 2018). This kindled a lively interdisciplinary debate and highlighted the necessity of shared normative frameworks to steer innovation towards ethically desirable and socially sustainable directions. In line with this trend, the European Union [EU] has recently presented its ethical framework to promote responsible innovation in CAV technology (Horizon, 2020) and asked stakeholders to contribute to its operationalization.

This paper responds to the call by presenting some methodological suggestions on how to ease the translation of the EU recommendations into practice from the viewpoint of engineering.[1] In what follows, we elaborate on a bottom-up, *function-based working approach* for the development of flowcharts, checklists, and similar methodological tools supporting the exercise of moral judgment aimed at aligning CAV design to the EU normative framework[2]. By focusing on given functions, determining which ethical challenges they pose vis-à-vis the EU framework, and devising *ad hoc* methodological tools to discuss them, the gap between principles and design practices can be narrowed down and the need for further conceptual refinements of the normative framework can be better specified.

The paper is structured as follows. In Section I we present the EU ethical framework, while in Section II we further clarify our aims and sketch the main features of the function-based working approach.

[1] This paper explores issues in AI and robot ethics from the viewpoint of integrating ethical values to technologies through design. However, we do not endorse technological solutionism, i.e., the idea that ethical design will suffice in the effort of aligning technologies to our moral values and minimizing social risks. Design is a powerful tool in this respect, but will yield tangible results only as part of a much wider commitment that involves organizational cultures, regulative frameworks, political and institutional efforts, and social initiatives. This is the background against which we would like our proposals to be discussed.

[2] We are aware that the ethical effort to minimize risks and maximize social benefits in the design of technologies is not reducible to the unreflective application of methodological tools. Ethics cannot be substituted by checklists or flowcharts. However, methodological tools offer guidance and structure to moral analysis and judgment, thus promoting discussion and helping identify and manage relevant concerns. This is the intention at the back of our work: not to bypass moral judgment through procedures, but to provide frameworks to support analysis, discussion, and creative thinking.

In the remaining sections we show how this approach can be applied to outline tools for bridging gaps between recommendations and practice. In particular, Section III focuses on problems revolving around the principle of personal autonomy; Section IV considers challenges posed by explainability; finally, Section V takes a closer look to privacy issues. In each case we map a methodological tool aimed at further operationalizing the EU guidelines. However, due to the preliminary stage of our research, the suggestions we advance are to be read more as evidence in support of the function-based approach than as refined tools for inquiry. Therefore, for each tool we underline what aspects are still in need of further clarification, thus setting the agenda for future research. Notwithstanding this limitation, we believe that our work might already demonstrate the productivity of the function-based approach and foster its adoption in the CAV scientific community.

§ I: Ethics of Connected and Automated Driving

In September 2020, the European Commission released its first systematic document on the ethics of autonomous driving. The report, entitled *Ethics of Connected and Automated Vehicles. Recommendations on Road Safety, Privacy, Fairness, Explainability and Responsibility* (ECAV from now on) is authored by the 'Horizon 2020 Commission Expert Group to advise on specific ethical issues raised by driverless mobility', an independent task force of 14 experts – mostly academics: philosophers, law scholars, engineers – led by Jean François Bonnefon and Filippo Santoni de Sio (Horizon, 2020).

Following closely the European approach to ethical AI put forward in *Ethics Guidelines for Trustworthy AI* (AIHLEG 2019) and analogous documents, the report aims at developing a coherent framework to analyse, assess, and manage ethical issues proper to CAVs, thus ensuring a "safe and responsible transition" (Horizon, 2020, p. 15) to driverless mobility in the EU. By applying a Responsible Research and Innovation approach (Owen *et al.*, 2012), the authors consider both risks and benefits related to CAVs, stressing the need for legal and ethical guidelines to steer technological advancements towards socially desirable outcomes. Based on this, the report presents 20 recommendations to promote alignment of CAV technologies to the EU fundamental values and, thus, justified social trust in innovation.

As such, the report is directed to all stakeholders and is intended to serve as a basis for a widely participated debate on how to face the ethical challenges posed by CAVs. Particular attention is indeed dedicated to the main stakeholders – manufacturers and deployers, policymakers, and researchers – whose roles in the accomplishment of each recommendation is carefully outlined (Horizon 2020, pp. 65-69). As we will see, this is supposed to help go over a common hurdle of similar enterprises, i.e., the necessity of filling the gap between statements and practice.

As in the case of AIHLEG (2019), the report opens with a section where the "fundamental ethical and legal principles" (Horizon, 2020, p. 21) are listed and briefly commented. The principles, which serve as normative cornerstones of the whole framework, are 1) Non-Maleficence, 2) Beneficence, 3) Dignity, 4) Personal Autonomy, 5) Responsibility, 6) Justice, 7) Solidarity, and 8) Inclusive Deliberation. While principles 1-6 are well established in the AI Ethics community (Floridi & Cowls 2019; Jobin *et al.*, 2019), principles 7 and 8 are less common but reflect the Expert Group's attention to vulnerable categories and stakeholder involvement, which are both commonly acknowledged as important values in the EU perspective. As such, the normative component of the framework is in line with the EU approach to the ethics of AI, of which the report manifestly represents a specification in the direction of CAV technologies.

The 8 principles are the backbone of 20 recommendations organized in three main sub-sets: Safety, Data and algorithm ethics (privacy, fairness, and explainability), and Responsibility. Each recommendation is presented in its content and bearings on different stakeholders. Moreover, in an attempt to offer more concrete insight, every recommendation is accompanied by a discussion box. Ranging from improving road safety to risk distribution and crash avoidance algorithms, from privacy to fairness, explainability, and finally to the many facets of responsibility, the 20 recommendations cover a large territory of ethically relevant issues in CAV development, deployment, use, and regulation, offering a rich guide to reflection and action. Although, as the authors readily recognize, some important problems remain unaddressed – "such as the connection between CAVs and environmental sustainability, the future of employment, and transport accessibility" (Horizon, 2020, p. 19) –, the proposed guidelines appear as a valuable starting point to promote a responsible attitude towards innovation and legislative efforts in autonomous mobility.

§II: A function-based working approach as a further step towards practice

Providing a general framework to clarify which values (and why) are to be complied with and a set of recommendations on how to do so comes a long way in the integration of ethics and CAV innovation, but does not come all the way down (Floridi, 2019; Morley *et al.*, 2021). To bring about changes effectively, frameworks and recommendations must find their way to the right people working in the right venues (Adamson *et al.*, 2019; Morley *at al.*, 2020). Principles and recommendations, no matter how refined, will always need an extra effort to be translated in good practices and ethically adequate technologies. This is a well-established truth in the field of applied ethics (Beauchamp, 1984; van den Hoven, 2008), and one the authors of the report are well aware of. As they explain, "researchers, policymakers, manufacturers and deployers of CAVs will sometimes have to take the extra step of *bringing the recommendations to their specific policy or industry domains*, and thus identifying the specific tools needed to translate them into living policies and practices" (Horizon 2020, p. 19). Frameworks such as this one, in fact, are elaborated "to support not to replace the work of stakeholders engaged in the design, development, and regulation of CAVs", so that "stakeholders should further collaborate with experts in the operationalization and translation into practice of the general principles and recommendations identified in the report, based on their professional expertise" (Horizon 2020, p. 70). A similar effort is integral to what the principle of responsibility demands, in that it requires "providing different actors (CAV users, but also CAV manufacturers and deployers) with sufficient knowledge, capacity, motivation and opportunities to comply with these standards" (Horizon 2020, p. 22).

In line with these suggestions, we present in what follows the initial sketch of some methodological tools intended to bring recommendations and practice closer. These admittedly preliminary results, which we hope to further refine in future works, were elaborated during interdisciplinary meetings between philosophers of technology and engineers working on CAV technologies at the Politecnico di Milano (Italy), during which ECAV was presented in detail and discussed. The need for philosophers and engineers to work together in such endeavours is tangible. While engineers add the necessary technical knowledge to the table, philosophers help clarify the conceptuality and issues involved in ethical frameworks. Both ingredients are required to merge theory and practice together and turn ethical principles into actual efforts (Morley *et al.*, 2019). As such, our work intends to contribute to the process of specification and 'translation into practice' of principles and recommendations, which the authors of the report also envisioned as the next stage of the process.

The cornerstone of our work is the realization that a productive way to circumvent the inevitable ambiguity of general guidelines is to specify them through a *function-based working approach*. By 'function-based working approach' we intend an operational framework which is expected to offer support in the application of general ethical guidelines to concrete cases. It endorses a bottom-up methodology which moves from specific technological functions or operations – as, for example, an instance of data processing or an automated task – and, depending on their features, helps connect them to relevant recommendations through *ad hoc* methodological tools.[3] The resulting discussion will in turn impact not only on potential design choices, but also on the assumed ethical frameworks, thus providing support both to design practices and conceptual refinement (figure 1).

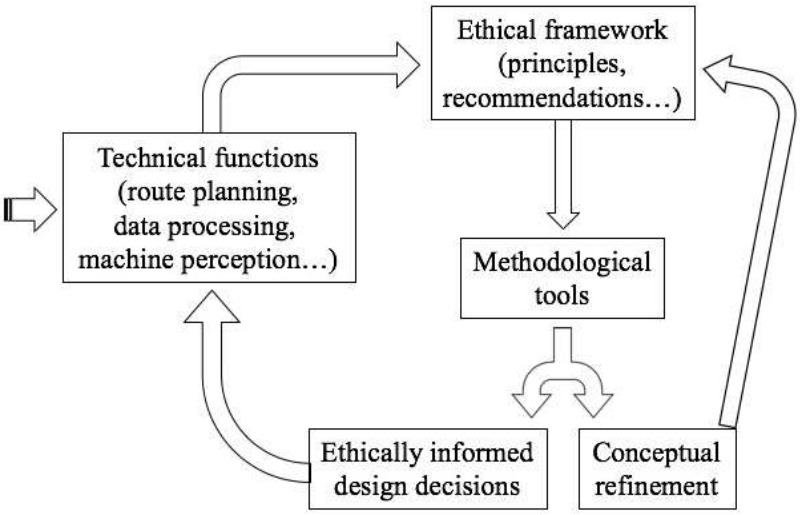

Figure 1 The function-based working approach

Indeed, assessing technological functions against the background of ethical frameworks and vice versa is useful not only to integrate moral values to design, but also to refine ethical concepts as they apply to given technical contexts. That said, our approach is not supposed to offer anything more than an ECAV-based methodology to raise awareness on potential ethical issues and to structure ethical discussion in interdisciplinary groups working on CAV technologies. The function-based working approach (along with the related tools) is neither a problem-solving procedure nor a source of regulatory outcomes. To an extent, it presupposes both. On the one hand, solutions cannot but be case-specific and heavily rely on rational justification, critical discussion and, more broadly, the exercise of moral judgment – which is precisely what the approach tries to foster. On the other hand, it has the aim of helping manage situations, frame trade-offs, and evaluate value-conflicts for which no standard procedure or regulation have already been established,[4] thus hopefully producing useful evidence for regulatory purposes. Its intended task, however, is of an ethical nature: to foster fine-

[3] One possible weakness of the approach might be that considering functions separately will make it harder to detect ethical issues raised by function interdependency or, more generally, pertaining to the CAV as a whole. Nonetheless, focusing on one function to assess its ethical significance against the background of ECAV does not necessarily imply a compartmentalized view of the technology as a whole. Actually, interdependency-related aspects belong to the technical description of a function and, thus, should be accounted for since the outset. However, it is likely that some of these interdependency-related aspects could pass unnoticed. Since our approach is not supposed to be exhaustive, a broader, system-level ethical analysis might complement our methodology and help cope with this possible blind spot.

[4] Of course, analysis inspired by the function-based working approach can also be carried out to evaluate the ethical content of a given regulatory proposal or law. However, the methodology we present primarily pertains to situations that are not already regulated by law where technologies have relevant ethical impacts.

grained discussion on the ethics of autonomous driving and support the exercise of moral judgment in design contexts, so to bring relevant ethical values closer to emerging technologies.
Although the development of similar methodologies lies outside the scope of ECAV, our perspective seems to be in line with some hints provided in the report, as for example the suggestion to approach the definition of safety metrics on a function-by-function basis (Horizon 2020, p. 26). Accordingly, a working approach that puts functions at the centre of the stage provides the granularity needed to bridge gaps between general ethical statements and particular cases, thus offering a mediation most needed in CAV ethics (Danvall, 2020) as in any other field of applied (Bayles, 1984) or engineering[5] ethics.
In this paper, we offer a demonstration of the potentialities of the function-based methodology by focusing on some principles and guidelines that struck us as in need of further specification or discussion to support an implementation process as clear and unambiguous as possible. Namely, we focus our attention on three issues – personal autonomy, explainability, and privacy – that, in our opinion, most urgently require fine-grained guidance due to the associated ethical risks[6]. By doing this, our purpose is not to oppose the EU framework, which on the contrary is timely and generally acceptable. Rather, we wish to provide further guidance on how to move beyond some ambiguity that can be found there and offer practical tools to the application of the recommendations advanced in the document. We believe that efforts such as this one are pivotal to truly integrate ethics to technological innovation and hope to foster analogous attempts in the research community.

§ III: Personal Autonomy

The Principle of Personal Autonomy (PPA) is the first element that might benefit from the application of a function-based approach. PPA figures as one of the 8 ethical principles that inspire the whole framework and, as an ethical value, enjoys wide acknowledgment in the EU moral landscape (AIHLEG, 2019). As the authors clarify, PPA states that human beings are to be conceived as "free moral agents" (Horizon, 2020, p. 22) and, as such, their right to self-determination ought to be respected. In relation to autonomous driving, PPA demands that CAVs are designed so to "protect

[5] Our approach shares its general aim with other engineering ethics methodology – in particular, with Value Sensitive Design (VSD; see Friedman, 1996; Friedman, Kahn, Borning, 2008; Friedman, Hendry, 2019). However, its scope is much more limited, in accordance with the much more limited purpose it is intended to serve in this paper. Indeed, VSD threefold iterative structure covers the entire ethical work that must be carried out to adequately implement values in technology design – from determining which ethical values are relevant and how they should be ordered to focusing on different stakeholders and more technical aspects. On the contrary, our approach is introduced exclusively to support the application of a given ethical framework, providing suggestions on how to assess technological functions in its light. Since the function-based working approach is specifically aimed at addressing application issues, it necessarily presupposes a wider normative framework. So, should the function-based working approach be used independently from a given ethical framework (in our case, the one presented in ECAV) it would need to be complemented by a normative effort – as happens with VSD. Therefore, it would be a mistake to put VSD and the function-based working approach on the same level: they stem from the very same need but their scope is different. Rather, our approach could be integrated into VSD as a useful resource to carry out the third stage of the iterative methodology, which focuses on how relevant values impinge on design. For an application of VSD to autonomous driving, see Thornton (*et al.*, 2018) and Umbrello, Yampolskyi (2021).
[6] The choice of the issues to further analyze has been inspired by considerations concerning both their relevance and their current state of elaboration and discussion. We do not intend in any sense to imply that problems connected to personal autonomy, explainability, and privacy are more severe than issues related to, e.g., safety or fair responsibility allocation. We decided to put our approach to test in relation to these three issues and not other, equally relevant problems hoping to provide inspiration and suggestions to adequately deal with concerns that are evidently important but somewhat less discussed in the current ethics debate.

and promote human beings' capacity to decide about their movements and, more generally, to set their own standards and ends for accommodating a variety of conceptions of a 'good life'" (Horizon, 2020, p. 22). As such, PPA is involved in many recommendations, ranging from the protection of privacy rights and the promotion of user choice to reducing opacity and enhancing explainability.

From a general perspective, there seems to be little reason to doubt the relevance of PPA and the associated recommendations. First of all, bypassing personal autonomy through technical means paves the way to technocracy, i.e., a state of affair where decisions concerning the well-being of users are taken by designers or engineers with no right or particular competence to do so and without users being aware of it or having any chance to partake in the decision-making process (Habermas, 1971). This state of affair is evidently incompatible with the EU culture and its stress on the right to self-determination on individual matters, but could occur when the way in which technologies operate bypasses human judgment – as, e.g., in the case of a speed control system that could not be overridden by human intervention even in case of emergency (Schoonmaker, 2016). Moreover, considering human beings as free moral agents by principle means, at the same time, considering them responsible agents as well, to the extent that they can exercise such freedom. This is an important presupposition to establishing who is responsible, and why, when harmful consequences follow from the use of autonomous technologies (Matthias, 2004; Nyholm, 2018c; Chiodo, 2021). So, PPA is vital to the necessity of distributing responsibility in a clear and fair way, of encouraging responsible behaviour, and of respecting other people's dignity – which are all fundamental objectives of the EU approach to trustworthy AI.

That being said, technologies such as CAVs put great pressure on personal autonomy, to the point that it becomes difficult to understand how to comply with such principle and, at the same time, keep developing them. Innovation in automated driving points at high-level automation and is fuelled by the opportunity for passengers to move around without paying any attention to the road. This seems to align with PPA (Glancy, 2012), particularly if we consider the second part of the explanation provided in the report, according to which we shall "protect and promote human beings' capacity to (…) set their own standards and ends for accommodating a variety of conceptions of a 'good life'" (Horizon, 2020, p. 22). Personal autonomy seems to be conceived here as the possibility of having access to the widest set of resources (be it free time or transportation) to pursue one's own individual interests and well-being. Automated mobility offers important opportunities in this regard. In fact, such understanding of 'personal autonomy' is often instantiated in narratives surrounding CAVs. For instance, it plays a central role in the common commercial narrative that presents the technology as a means to free oneself from the burden of driving so to gain time for one's own personal activities.[7] Other narratives as well, however, lend themselves to strengthening this trend towards full automation. For example, from a safety and traffic management perspective, human intervention might be said to get in the way of making roads safer, less polluted and less congested due to its unpredictability and slow reaction time. Finally, this facet of PPA seems to lie also at the heart of the narrative based on inclusiveness, according to which CAVs will empower users who are now excluded from regular driving due to various cognitive disabilities, as also demanded by the principles of beneficence and solidarity (Goggin, 2019). Accordingly, human intervention is increasingly disappearing – or, at least, such is the final objective.

This trend, however, seems to be in contrast with a second meaning associated to the idea of PPA, i.e., that we must protect and promote "human beings' capacity to decide about their movements"

[7] For an example, see https://www.bmw.com/en/innovation/value-of-time-via-autonomous-driving.html.

(Horizon, 2020, p. 22). The focus seems now to be less on the general accomplishment of self-determined ends and more on the meaningful control of what happens when CAVs are used. Indeed, the exercise of personal autonomy is variously impacted by the way in which CAVs automate transport (Xu, 2021). The experience of driving is a complex one, composed by a myriad of decisions to be made, some of which are moral decisions or might have a considerable impact on personal well-being. Consequently, it seems reasonable to claim that some decisions should remain in the hands of human users and that full automation should not become a threat to this form of limited supervision or intervention (Nunes *et al.*, 2018; Fridman, 2018). This is why frameworks that study how to keep users of CAVs in control of morally relevant decisions, like the Meaningful Human Control approach (Santoni de Sio & van den Hoven, 2018), are getting more and more traction. However, it is hard to figure out, in the present state of affair at least, how protecting the exercise of user autonomy can go hand in hand with the other narratives we have briefly referred to. In sum, it seems that PPA steers in directions that are difficult to harmonize, since it supports both partial and full automation. This leaves engineers with the puzzling task of figuring out in what sense personal autonomy – intended as human control over relevant driving functions – can be a value to embed in high-level vehicle automation, or how to embed it.

This ambiguity, that stems from the complexity of PPA and competing narratives about CAVs, represents a barrier towards designing CAVs that protect and promote personal autonomy. Further research must tackle this obscurity and provide less ambiguous accounts of personal autonomy in the context of autonomous driving. However, until new frameworks concerning personal autonomy in driving automation emerge, ambiguity must be assumed as a given. Learning how to deal with it becomes then of utmost importance. The most urgent task is to raise awareness on possible threats to personal autonomy – whatever ill-defined the concept might be – and devise solutions to minimize potential harm. Such preliminary, applied ethics work might help specify what it means to develop CAVs compliant with the respect of user autonomy and raise awareness on the associated challenges. To this aim, and with the scope and purpose of our approach in mind, we believe that resorting to a function-based methodology might contribute to handling at least some of the ambiguity surrounding the PPA. It is possible to elaborate two similar tools depending on the analytical task that needs to be carried out. A first concern in this respect might be to carry out an ethical assessment of all system functions that might impact negatively on personal autonomy. In this sense, it might be useful to define more clearly:

a. which driving functions – if any – should remain under user control for personal autonomy to be respected in high-level automation;
b. what would it mean "to be in control of a driving function", i.e., what conditions should be satisfied for users to exercise meaningful human control over those driving functions;
c. how a. and b. would impact on the commercial, safety-traffic management, inclusiveness and other narratives on CAVs;
d. the feasibility of a. and b. vis-à-vis high-level automation as a technological objective.

In some situations, however, the task at hand might not be to assess the ethical significance of a set of given functions, but to evaluate the ethical impact of just one particular function (F). Suppose, for example, that a new functionality is being tested which in some way bypasses human intervention under a specific respect – e.g., route planning. In similar cases the following methodological tool might be applied:

a. Should F remain under user control for personal autonomy to be respected in high-level automation?
b. What would it mean "to be in control" of F, i.e., what conditions should be satisfied for users to exercise meaningful control over it?
c. how would a. and b. impact on the commercial, safety-traffic management, inclusiveness and other narratives on CAVs?
d. Would a. and b. be feasible vis-à-vis high-level automation as a technological objective?

An inquiry aimed at clarifying points a-d might help manage issues related to the protection and promotion of personal autonomy in CAV technologies, thus helping specify the notion of personal autonomy as it applies to autonomous driving. Such specifications might in turn be of support in moving from ethical analysis and discussion to standards and regulation.

An example: route planning. Value conflicts that originate from the tension between user autonomy and high-level automation are many and diverse. A first example of a function that might require specific guidelines is route planning (Schoonmaker, 2016). Suppose that user P wants to drive from point A to point B. Of course, setting the destination is a task that will remain under the purview of human users even in high-level automation. However, setting the best route is arguably a task for the system to carry out. Delegating this task to CAVs has relevant collective advantages. For example, it increases predictability. If CAVs will be able to share their routes with one another, letting the system perform mission planning and dynamic path planning fully autonomously and in real time will contribute to minimizing uncertainties, thus increasing the ability to predict the behaviour of other vehicles. This, in turn, will have a positive impact on the efficiency and safety of short time path planning while, at the same time, reducing the computational load otherwise necessary to carry out planning functions (Mozaffari *et al.*, 2020). Moreover, automating road planning makes it easier to manage traffic processing on a wider scale (Friedrich, 2016). In a scenario where vehicles are fully autonomous and interconnected, managing the whole traffic flow would become a centralized optimization problem with fairly predictable multiple agents, which would make it easier to handle. In such an environment, in fact, traffic control mechanisms would benefit from real time, less uncertain knowledge of the vehicle flux, thus enabling the control system to deviate traffic as per the need. Similarly, if need be, each participating vehicle could plan its own path to avoid congested zones, ultimately improving traffic efficiency and ensuring optimal use of the available road infrastructure. Sustainability will be improved as well. In fact, mission paths and dynamic paths planning programs will be able to select optimal paths with minimum energy consumption, which will impact positively on the environment (Barth *et al.*, 2014). Lastly, information processes with limited or no human intervention work with less constraints and are computationally cheaper, thus requiring less energy to run.
For all these reasons, the case for categorizing "road planning" as a function that should not remain under the purview of human users, but should rather be delegated to the CAV system, appears to be a strong one. However, when road planning is delegated to the CAV system, user autonomy is evidently bypassed – at least partially. Although the advantages of automating road planning are considerable, choosing which road to take might be perceived as an exercise of autonomy (perhaps even moral autonomy) – and, therefore, a function that should not be delegated to the CAV system insofar as user autonomy should be promoted and protected [a].

Going back to our case, once user P sets her destination, her CAV system develops a mission plan: the planner computes multiple strategies and chooses the most convenient – i.e., the one that maximizes the parameters that programmers have selected to represent various constraints and costs. Suppose, for instance, that P would prefer to take another way, since the one computed by the planner would take P closer to her ex's house and P does not want to get anywhere near it. Or suppose, as considered in ECAV, that an automatedly computed route will have P's CAV drive through a non-public area in which data are collected without passengers being aware or consenting to this (Horizon, 2020, pp. 40-41). Many personal reasons, even ethically relevant ones, might impinge on the roads we decide to take. Bypassing these decisions by delegating route planning to CAVs might have a relevant impact on the exercise of user autonomy.

Although there is no explicit discussion of this issue, some suggestions in the report seems to support the claim that "choosing routes" (Horizon, 2020, p. 41) is a function that should be meaningfully controlled and exercised by users. In light of this, one might wonder, providing users with various options to choose from, even though they may not choose the optimal alternative, might seem the only solution truly compliant with PPA. However, this solution comes with costs. Delegating road planning to users might have many disadvantages, both technical – as unpredictable traffic processing and algorithm complexity – and in terms of usability – increased cognitive load, interface complexity, and so on. In this respect, answering [b] by specifying which degree of autonomy should be left to users when it comes to road planning will help clarify which solutions should be integrated to the technology in order for it to satisfy the demands of PPA. Finally, measuring the impact of such potential solutions on other relevant narratives or values [c] and assessing their actual feasibility [d] might help go over, or at least handle, some of the ambiguity of PPA. By following the steps of this function-based methodological tool it might be possible to structure reasoning and envision potential design solutions to improve the promotion and protection of personal autonomy in CAV technology.

§ IV: Explainability

In the context of the European approach to the ethics of algorithms, explainability has emerged as a specific principle to be added to the list of more general ethical principles such as non-maleficence, beneficence, dignity, justice, and so on. In AIHLEG (2019, p. 13) it is stated that automated decision-making needs to be "– to the extent possible – explainable to those directly or indirectly affected" as part of a more general commitment to explicability, which encompasses process transparency and open communication concerning the capabilities and purposes of AI systems as well. The main reason for concern is the opacity of black-box algorithms, which is to be reduced proportionally to the harmful consequences that might follow from errors or inaccuracies. Moreover, explanations must be "timely and adapted to the expertise of the stakeholder concerned" (AIHLEG, 2019, p. 18). So, the results of techniques such as those developed in the field of Explainable AI (Barredo Arrieta *et al.*, 2020) must be paired with an effort to translate technical information into a language that users and laypeople can understand. Finally, explainability concerns not only computational decision-making processes, but also human decisions concerning the design and deployment of a given technology in a given context.

In ECAV, the two recommendations dedicated to explainability build on this approach. Recommendation 14 encourages to reduce opacity in algorithmic decision-making through user-centred interfaces and methods, stressing the importance of recurring to Explainable AI techniques, accessible and transparent vocabulary, and intelligible system explanations and justifications.

Recommendation 15 shifts the attention to education and public participation, highlighting the need of providing the public with the knowledge, opportunities, and skills necessary to adequately understand their interactions with CAVs, be aware of the risks involved, and be able of fully exercising their rights. As in AIHLEG (2019), explainability is closely associated with other principles and is acknowledged as a necessary condition to social trust: "Without adequate means of access, the role of human agency and oversight is severely weakened or hindered and risks undermining the principles of human dignity and autonomy, with the consequence of critically eroding public trust in these fast-developing technologies" (Horizon, 2020, p. 50).

Although the general framework is clear and acceptable, the move to more concrete action has a challenge to face. In fact, the report argues that explainability is to be provided of "relevant CAV applications of algorithm and/or machine learning based operational requirements and decision making" (Horizon, 2020, p. 48). However, little guidance is offered on what makes a function "relevant" in this sense. Such specification task seems a good starting point for applying our function-based working approach to explainability issues. Moreover, the identification of relevant functions will help determine which further aspects require consideration. In fact, in addition to providing guidance in distinguishing relevant form irrelevant functions vis-à-vis explainability, the methodology helps frame discussion concerning which means are most apt to the task and which competences and skills are necessary for users to understand relevant CAVs operations. In turn, discussion and experimentation surrounding these aspects will help elaborate more fine-grained guidelines to the implementation of explainability.

The first step of the methodological tool concerns the kinds of functions that are to be considered as "relevant" vis-à-vis explainability, i.e., that should be made accessible to users. Surely, it would be counterproductive to provide users with indiscriminate access to all processes carried out by a CAV. This is the case firstly because some processes will raise no ethical concerns if users will be unable to access or understand them. For instance, it seems unnecessary to show and explain in detail to users how different data originating from internal sensors (e.g., measuring acceleration and velocity or monitoring the state of various components) contribute to determining the AV status and the optimization of its functioning.

An analogy to traditional driving education seems relevant here. Even in traditional driving many processes are arguably obscure to ordinary drivers, without this exposing them to significant risk, and thus are not covered in educational programs. In order to get a license, one must not study mechanical engineering and vehicle dynamics, but learn just what one might need to understand that assistance is needed. Similarly, it seems that explainability to passengers must be provided not for all processes and operations, but just for a subset of them that exhibit some determinate features. Therefore, some guidelines should be elaborated to help engineers distinguish between functions that should be made accessible and explainable to users through interface design and functions that could, or perhaps should, remain hidden. In this sense, it might be useful to determine:

a. which algorithms could remain opaque to users, and why;
b. which algorithms should be explainable to users and why;

To carry out this analysis it is necessary to clearly spell out under which conditions a CAV process should be designed to be explainable to users, and under which conditions this requirement is unnecessary (Rosenfeld & Richardson, 2019; Setchi *et al.*, 2020). Raising awareness of risks or potential harm might be a promising starting point to figure out which criteria must be considered to

decide whether a function requires explanation. In addition to this, it will also be important to take into due consideration the attention and cognitive capabilities of users, which also seem relevant factors particularly when explicability must be provided during operation time (Miller, 2019). Cognitive stress and information overload are both possible outcome of a poorly balanced explanatory effort. Therefore, due care should be exercised not to overwhelm users, thus impairing their ability to exercise responsible and considered judgment. In this respect, the way to elaborate and synthesize various data in order to provide adequate explanations to users could differ depending on the roles and the motivations of the explanation, so that a taxonomy of data types, user rights to explanation, and motivations might also be very useful to minimize contrasts with other principles such as personal autonomy, dignity, and justice. All these aspects, the clarification of which must be postponed to future work, clearly underline the necessity of determining the rationale behind the distinction between relevant and irrelevant functions in relation to explainability.

The second step concerns the means of communication between human users and CAV systems. In fact, as already stressed, explainability is not to be obtained exclusively on a technical level, but mostly on a HMI level. This means that machine functions must become accessible to users independently from their technical competence, so processes and data must be presented in a generally understandable fashion (Confalonieri *et al.*, 2021). This poses huge practical issues not just of a technical nature, but also of a communicative kind. How to provide access to, or display, the collected data concerning the system state and operations in ways that effectively convey the right amount of information to users at the right moment is an important aspect that will affect future interaction modalities. Nonetheless, it remains rather unclear as of now what actions should be taken in order to provide users with the right degree of explainability.

The modalities through which these data are displayed and access is provided to them are extremely important for designing technologies capable of being perceived as reliable and 'trustworthy', as the EU approach envisions. Therefore, it is fundamental to determine rules – or at least necessary requirements or benchmarks – for designing and developing effective interfaces for explainability to be integrated in CAVs.

Direct explanation of automated decisions is surely an effective modality in this regard. However, also more indirect modalities might be equally effective. For instance, informational feedback might impact positively on user trust in autonomous systems like CAVs. Informational feedback allows users to continuously monitor the status of the vehicle and better understand its behaviour, thus enhancing explainability and promoting the development of adequate mental models of the technology. Trust is a key factor in many aspects of the autonomous vehicle HMI design and profoundly influence user acceptance (Liu *et al.*, 2020). Moreover, user trust is important for safe, efficient, comfortable and enjoyable driving in general. This does not mean that the more users trust the technology, the better it is. Rather, the aim is to generate in users a level of trust that is adequate to the technical reliability of the system. In fact, overtrusting is just as problematic as distrusting (Dzindolet *et al.*, 2003). Overtrust occurs when user trust exceeds system reliability. Overtrust commonly causes misuse and might even lead to accidents, as it is arguably the case in the 2016 and 2018 Tesla mortal accidents (NTSB, 2020). This is particularly relevant in high automation, where passengers can avoid paying continuous attention to the road ('hands off', 'mind off') but might be required to retake control due to severe system failures or changes in the Operational Design Domains. On the contrary, when user trust falls short of system reliability, a situation of distrust and disuse arises, which hinders the introduction of technologies that might have significantly positive impacts – in the case of CAVs, on road safety, inclusiveness, sustainability, traffic management, and

so on. The ideal or adequate level of trust, then, is reached when user trust and system reliability are well balanced and aligned. Indirect explainability through informational feedback might be a useful tool to accomplish this alignment. A taxonomy of different modalities to provide users with understandable and effective explanations of relevant functions, paired with information concerning their strengths and weaknesses, might be of great help in the effort of operationalizing explainability. In line with these considerations, it would be helpful to clarify:

c. which direct and indirect modalities can be implemented to offer effective explanations;

Such preliminary work on the means to concretize explainability in CAV technologies would help structure research concerning strengths and weaknesses of each modality, and thus concur to pairing each function to the modality best suited to offer effective explanations. For instance, in the case of functions that require case-by-case explanation feasibility might be a concern: it might be impossible, for example, to present users with justifications supporting decisions to be taken in situations where time is of the essence. Also, the computational costs might be difficult to handle. Inclusiveness could be a further concern for some modalities of explanations, just as potential difficulties for cultural habits or health disparity must be considered. In these cases, indirect modalities could be better suited for the task. In light of this, the following points should be discussed:

d. which strengths and weaknesses are associated which each modality;
e. given a function F, which modality is best suited to provide effective explanation of F.

Finally, the implementation of modalities suggested by e) might shed light on which competences and skills are to be presupposed in users for them to be able to adequately respond to system explanations. This information would help clarify the kind of knowledge that should be transferred to users and the public in general in order for them to interact with CAVs in a responsible and informed way. Both engineering design and educational programs would benefit from such a close cooperation and users would access to training tailored to the actual features of the technology (Liu *et al.*, 2020). The last question of our tool would then be:

f. given modality M, what competences and skills should users be equipped with?

Put together, the methodological tool would look as follows:

a. which algorithms could remain opaque to users, and why?
b. Which algorithms should be explainable to users and why?
c. Which direct and indirect modalities could be implemented to offer effective explanations?
d. Which strengths and weaknesses are associated which each modality?
e. Given a function F (from b.), which modality (given c. and d.) is best suited to provide effective explanation of F?
f. Given a modality M, what competences and skills should users be equipped with?

We believe that taking these questions into consideration from the perspective of CAV design might be of support both in the effort of operationalizing explainability and of setting benchmarks, standards, and best practices.

§ V: Privacy

In line with existing regulation (GDPR, 2016; Constantini *et al.*, 2020), in the context of ECAV privacy is understood as the individual right to control access to and use of information that pertain to one's own personal sphere of existence. In order for CAVs to properly function[8] – but also to seize new advantageous opportunities[9] –, personal data concerning individuals such as end-users or road users must be gathered and processed, if not also shared and stored. Sensors that collect biometric data for monitoring the users' states and facial recognition technology for personal identification are two examples of possible CAV functions that involve sensible data. Personal data– in particular, data from road users – will also be very useful in personalizing the services which CAV users can access. Suppose a CAV user wants to be alerted once the vehicle passes in front of a certain store, or when the shop window of that store is renewed. If this information is not on the map, data about the vehicle surroundings need to be collected and processed in order to provide this service. This might raise serious concerns from the GDPR perspective, as no informed consent can be a-priori filed for the use of such data. Similar situations expose data subjects to several privacy-related risks such as threats to personal autonomy, disclosure of personal sensitive information, and surveillance (Glancy, 2012; Schoonmaker, 2016). Thus, it becomes necessary to enforce the right to privacy through ethical recommendation and regulation.

Generally, individuals must be put in condition of giving informed and unambiguous permission for collection of personal data. Moreover, data subjects must be able to easily control personal information sharing and enjoy protection from surveillance, data leaks or thefts. From the perspective of design, it follows that particular care must be taken to provide "reliable and sufficient protection against manipulation, misuse or unauthorized access to either the technical infrastructure or the associated data processes" (Horizon, 2020, p. 36), which would ideally lead to the development of "industry standards that offer robust protection without relying solely on consent" (Horizon, 2020, p. 37). To sum up, it is a design duty to minimize the risk of data infringement and to promote the protection of private information from tampering, leaks, and thefts.

Although the ethical framework surrounding privacy is sufficiently clear, it is more difficult to figure out how to apply it to concrete cases, especially when functionality is involved (Liu *et al.*, 2020). In some occasion, operations that are critical for the efficiency of CAVs require sensible data to be carried out, thus leading to a value conflict (functionality vs. privacy) of which the report does not say much. However, the nature of the problem, being already centred on the ethical impacts of given functions, lends itself to be discussed through our function-based working approach. In order to sort out similar issues, and in line with the necessity of "identify(ing) alternative and CAV-specific solutions to protect informational privacy and informed consent, and establish best practices for industry" (Horizon, 2020, p. 38), we thus propose a methodological tool which aims at reducing privacy risks posed by data processes that are necessary for the proper functioning of CAVs but, at the same time, involve sensible data.

[8] The report covers also privacy issues arising from the collection and sharing of data for purposes that extend beyond the proper functioning of CAVs, such as advertising, profiling, and marketing. In what follows, we will focus the attention on privacy issues related to data processes that are necessary for CAV technologies to function properly.

[9] For instance, "help determine liability in accidents, streamline insurance pricing, motivate better driving practices, and improve vehicle safety" (Dhar, 2016, p. 82).

Our tool is based on a useful distinction that is not entirely acknowledged in ECAV, i.e., the distinction between on-board and off-board data processing. Several privacy guidelines proposed in the document, in particular those revolving around Recommendation 9, appear in fact to incorporate a presupposition according to which a large number data processes must inevitably involve external infrastructure – so that, for CAVs to work properly, (sensible) data are to be sent outside the vehicle. For example, in the Discussion of Recommendation 9 (Horizon, 2020, pp. 39-40) the focus falls immediately on V2V (vehicle-to-vehicle), V2I (vehicle-to-infrastructure), and V2X (vehicle-to-everything) scenarios, as if such situation were an unavoidable condition for CAVs to properly function. Although connectivity surely offers great opportunities in terms of efficiency, this presupposition could, and should, be challenged. By doing this, it would be possible to get a more fine-grained and technologically informative approach to data management in CAVs. For this reason, we believe it might be useful to provide the approach with further details in order to move towards more effective and viable guidelines. In this section we present some preliminary ideas on how to accomplish this result.
As a first step, we introduce a framework constituted by two Principles, one Aim, and one Maxim.

a. Principle 1 (Ethics): Privacy is a value relevant to automated and connected driving.
b. Principle 2 (Technology): On-board data processing is safer, privacy-wise, than data processing that involves off-board infrastructure.
c. Aim: Responsible engineers should mitigate or minimize privacy risks.
d. Maxim: To minimize privacy risks, data processing that involves sensible information and that *can* be executed exclusively on-board, *should* be executed exclusively on-board.

The first Principle is of an ethically normative nature and states that *privacy* is a value relevant to automated and connected driving. As such, the principle – which is widely acknowledged in literature – is directly assumed from ECAV, to which we also refer for its justification. The second Principle is more of a technological nature and states that on-board data processing is generally safer, privacy-wise, than data processing that involves off-board infrastructure. This principle is intended to challenge the presupposition according to which all data processing involves off-board infrastructure and, as such, to add the granularity needed to provide more easily applicable guidelines. In this case, however, a justification is needed.
We think that there are strong reasons to claim that on-board data processing presents less privacy risks than data processing involving off-board infrastructure (Glancy, 2012; Rannenberg, 2016). Indeed, in the latter case, the number of devices and algorithms that need to appropriately handle data privacy is higher and, thus, it requires more effort to be secured. If the data coming out of a vehicle is already anonymized the remaining part of the processing has less constraints from a privacy preserving perspective. Also, we believe it is equally reasonable to claim that on-board data processing represents a viable technical solution for some automated driving functions that are commonly associated with off-board data processing. For instance, there is no technical need to perform off-board sensor data processing for functions such as obstacle detection and collision avoidance. Despite the recent V2X communication infrastructure will allow to send data streams off vehicle and get remote commands in real time, this could be obtained, at least partially, also through on-board processing. Some mixed solutions exist and could be envisaged for striking trade-offs between full on-board processing (i.e., with a high demand in hardware cost) and full processing offloading (e.g., what is done nowadays with the cloud paradigm). By mixing the two approaches

and offloading pre-processed anonymized data, it could be possible to preserve privacy and leverage on external resources too. In light of this, in order to associate risks with data collection and processing in a clear way, it might indeed be useful to distinguish between cases where it is necessary for data to leave the vehicle and cases where all data processing could be handled inside the vehicle, thus limiting risks of breaches and leaks to the on-board systems.

Now that the principles have been adequately justified, let's consider the Aim and the Maxim. Firstly, and obviously, the relevance of the two principles pairs with the Aim of mitigating or minimizing the privacy risks in the development and use of automated vehicles. The Aim directly follows from Principle 1 when it is assumed in a context of responsible engineering, as it is now. As such, it does not require any further justification. Finally, all these elements support the Maxim according to which, in order to minimize privacy risks, data processing that involves sensible information and that *can* be executed exclusively on-board, *should* be executed exclusively on-board. The Maxim is the main practical guideline of our proposal and is sufficiently fine-grained to provide applicable guidance in reasoning concerning privacy and the design of automated vehicles. Its coherent and extensive application would approximate a condition where sensible data (e.g., about faces, license plate numbers, vehicle owners, passengers, etc.) are handled exclusively on-board, while only privacy-neutral or low risk data (e.g., position[10], speed, trajectory) leave the vehicle and are shared with off-board infrastructure.

To operationalize the framework, it could be helpful to elaborate a methodological tool to guide applied reasoning concerning the relation between on/off board data processing, privacy risks, and recommended action. To this aim, we elaborated a flowchart which may offer support in determining how to mitigate privacy risks by minimizing liabilities due to off-board data processing and how to single out impacts on privacy that require further consideration and assessment. This is an outlook of the flowchart:

1. In the automated and connected driving system under examination, which data processing pose risks to privacy?

2. Of 1, which data processing involve exclusively on-board systems?
   2.1. Of 2, could privacy risks be appropriately handled?
      - If Yes, then privacy risks will have been mitigated.
      - If No, then sensible data should not be included in the data processing.

3. Of 1, which data processing also involve off-board infrastructure?
   3.1. Of 3, which data processing could be executed exclusively on-board?
      3.1.1. Would they still pose risks to privacy?
      - If Yes, then proceed to 2.1.

[10] Data related to the position of an AV are actually problematic from a privacy standpoint. On the one hand, it might seem that they do not qualify as personal data – provided that they lack any link to the individuals who own or use the vehicle, which might be obtain through, e.g., anonymization techniques or aggregation. Moreover, it might be argued that data on speed and trajectory would be rather pointless without information about the location of the vehicle. On the other hand, should these data be continuously collected, de-anonymization through data merging would make it possible to infer sensible information about users (Rannenberg, 2016). For example, matching data concerning departure and arrival points of a commuting vehicle with data concerning home and work addresses of potential users might make it possible to infer who travels on a particular vehicle and, so, to track their past and present movements or to build a model to predict their future movements.

- If No, then privacy risks will have been mitigated.

3.2. Of the data processing that could not be executed exclusively on-board, could the privacy risks they pose be appropriately handled?
   - If Yes, then privacy risks will have been mitigated.
   - If No, then sensible data should not be included in the data processing (> Impact assessment and trade-offs).

The flowchart, as formulated above, is intended to serve as a tool for carrying out an analysis of all data processing impacting on privacy. In Figure 1, the same flowchart is graphically represented as a tool for assessing a single instance of data processing. No essential difference distinguishes the two, but since different situations might require the use of one or the other formulation, for clarity's sake we thought it best to spell out both of them.

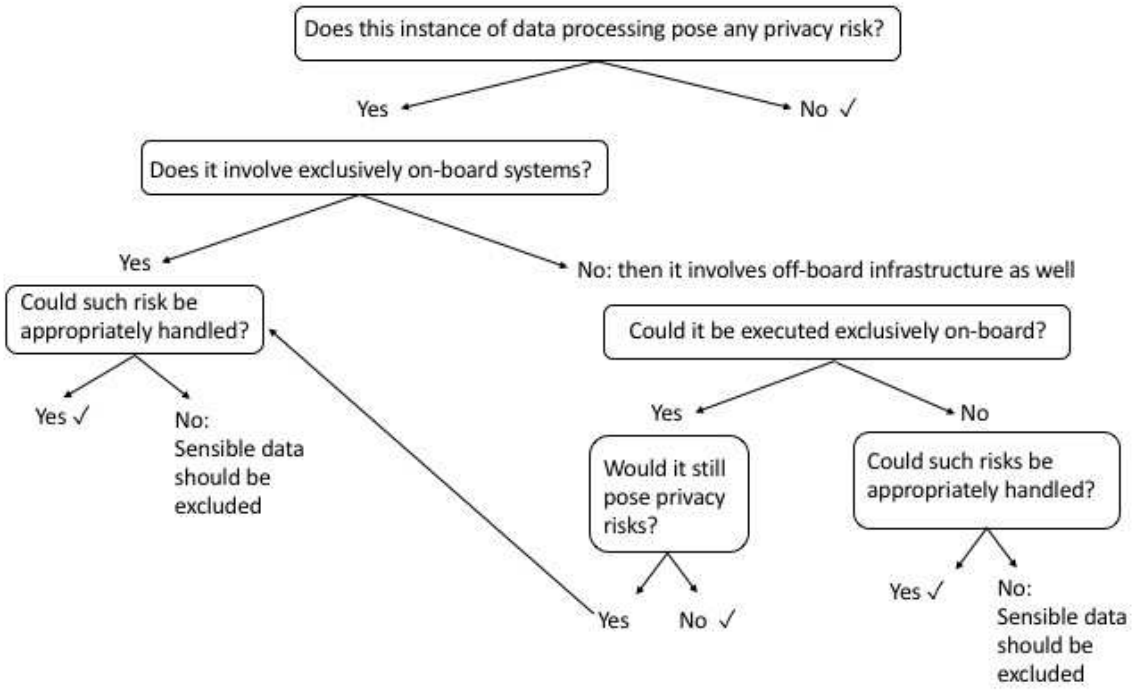

Figure 2: Privacy flowchart

Although we believe that such tool might already prove useful, many other aspects are in need of further inquiry.[11] For example, it evidently presupposes a clear notion of which data are sensible and thus need protection in the context of automated driving (Schoonmaker, 2016; Rannenberg, 2016; Krontiris *et al.*, 2020). Also, it is important to notice that the flowchart can be used to analyse privacy

[11] Moreover, privacy issues are not raised just by data processing. Other functions are also relevant in this respect – such as, for instance, data storing. Even though further research will have to be carried out, we believe that the flowchart to operationalize privacy in data processing could be adapted to data storing too. The main assumption would be that on-board data storing should be avoided unless the expected benefits would evidently trump the risks, which in any case should be identified and minimized. For instance, storing data on board could be useful for legal and technical reasons – e.g., so to enable ex-post analysis of accidents or near misses. In these cases, strictly necessary data could be stored in black boxes ensuring security, privacy, and proportionality – e.g., for a limited amount of time and providing access exclusively to legally authorised subjects for well-defined purposes (as happens in the case of video surveillance: see, for example, EDPS, 2010). Soft and hard regulatory frameworks are key to provide clear guidance to practitioners in this context as well.

issues at both development time and run time, since the two situations are different and pose different privacy threats.[12] Moreover, the results of the inquiry might also prove useful to elaborate a research roadmap, since they help realize situations where current technologies lead to hard privacy issues and innovative solutions might be required. Indeed, the analysis might help direct future efforts towards devising new methods to translate off-board into on-board operations and to provide extra protection in cases where data must necessarily leave the vehicle. In addition to this, it also important to stress that negative results must always be accompanied by further analysis in order to assess the impact of the exclusion of sensible data on related aspects of automated driving. For example, it seems necessary to assess how a privacy-enforcing solution would impact on the system overall functionality in order to evaluate potential trade-offs between privacy protection and system efficiency. Similarly, and subsequently, ethical trade-offs might ensue between privacy protection and other relevant ethical values such as safety, inclusiveness, autonomy, and so on (Dhar, 2016).
Future research is necessary to inquire into these many aspects with adequate attention. At this moment, we believe it is best just to provide a possible example where on-board and off-board processing could be questioned. Consider distributed sensing in V2V and V2I scenarios. According to some possible realizations, sensor data (including camera and lidar information) could be exchanged between several vehicles and between vehicles and infrastructure to perform sensor fusion. Image processing could be performed at the infrastructure level where images from several vehicles would be collected and fused; or at the level of each single vehicle, where images would be processed and high-level, anonymized information is exchanged with the infrastructure post processing. This high-level information could be the presence of a generic pedestrian in a given position or simply an anonymized image where faces and licence plates have been removed. Rather than discussing what might be the most proper solution, what is interesting for our purposes is to note that, in a computing continuum scenario, distributing computation between the "edge" (i.e., the CAV) and the "cloud" (i.e., the infrastructure) might result in privacy-enhancing applications.

Conclusion

In this paper we have presented the results of an interdisciplinary research concerning the recent EU ethical recommendations on connected and automated vehicles aimed at its operationalization. Ethical frameworks that clarify principles and recommendations play an important role in organizing social action and promote responsible and sustainable innovation. However, frameworks are ineffective unless stakeholders commit to their operationalization and help translate guidelines into best practices, benchmarks, standards, and regulations. To this purpose, we answer the call to operationalization by introducing a bottom-up, function-based working approach for the development of methodological tools that aim at simplifying the practical application of the EU normative framework. The productivity of the function-based approach is explored in connection to personal autonomy, explainability, and privacy in the context of CAV design and development. Each methodological tool is intended to bring further clarity and granularity in its respective context, thus supporting value alignment and responsible innovation. Although many aspects still need to be inquired and refined for these tools to properly serve their function, we believe that they offer – even

[12] For instance, at development time it might be necessary to collect huge amount of data for machine learning and high-density maps – in this case, data need to be stored and processing necessarily happens offboard. At runtime, the storage and processing of such an amount of data is not required and thus privacy concerns are different.

in this preliminary state – substantial support to the claim that experimenting with a function-based working approach in interdisciplinary research contexts might come a long way in the effort towards the operationalization of the framework proposed in ECAV and, consequently, towards responsible innovation in CAV engineering.

References

Adamson, G., Havens, J. C., & Chatila, R. (2019). Designing a value-driven future for ethical autonomous and intelligent systems. *Proceedings of the IEEE*, *107*(3), 518-525. https://doi.org/10.1109/JPROC.2018.2884923.

AIHLEG (2019), *Ethics Guidelines for Trustworthy AI*, https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence (accessed March 22nd, 2021).

Barth, M., Boriboonsomsin, K., & Wu, G. (2014). Vehicle Automation and Its Potential Impacts on Energy and Emissions. In G. Meyer & S. Beiker (Eds.), *Road Vehicle Automation. Lecture Notes in Mobility* (103-112), Springer. https://doi.org/10.1007/978-3-319-05990-7_10.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012.

Bayles, M. (1984). Moral Theory and Application. *Social Theory and Practice, 10*, 97-120. https://doi.org/10.5840/soctheorpract19841015.

Beauchamp, T.L. (1984). On Eliminating the Distinction Between Applied Ethics and Ethical Theory. *The Monist, 67*(4), 514-531. https://doi.org/10.5840/monist198467430.

Chiodo, S. (2021). Human autonomy, technological automation (and reverse). *AI & Society*, 1-10, https://doi.org/10.1007/s00146-021-01149-5.

Confalonieri, R., Coba, L., Wagner, B., & Besold, T. R. (2021). A historical perspective of explainable Artificial Intelligence. *WIREs Data Mining & Knowledge Discovery, 11*(1), 1-21. https://doi.org/10.1002/widm.1391.

Costantini, F., Thomopoulos, N., Steibel, F., Curl, A., Lugano, G., & Kováčiková, T. (2020). Autonomous vehicles in a GDPR era: An international comparison. In D. Milakis, N. Thomopoulos, & B. van Wee (Eds.), *Advances in Transport Policy and Planning, Volume 5* (191-213), Academic Press. https://doi.org/10.1016/bs.atpp.2020.02.005.

Davnall, R. (2020). Solving the Single-Vehicle Self-Driving Car Trolley Problem Using Risk Theory and Vehicle Dynamics. *Science & Engineering Ethics, 26*, 431-449. https://doi.org/10.1007/s11948-019-00102-6.

Dhar, V. (2016). Equity, Safety, and Privacy in the Autonomous Vehicle Era. *Computer*, *49*(11), 80-83. https://doi.org/10.1109/MC.2016.326.

Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., Pierce, L.G., & Beck, H.P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*, 697-718, https://doi.org/10.1016/S1071-5819(03)00038-7.

EDPS – European Data Protection Supervisor (2010). *The EDPS Video Surveillance Guidelines,* https://edps.europa.eu/sites/default/files/publication/10-03-17_video-surveillance_guidelines_en.pdf

Floridi, L. (2019). Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical. *Philosophy of Technology, 32*, 185-193. https://doi.org/10.1007/s13347-019-00354-x.

Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review, 1*(1). https://doi.org/10.1162/99608f92.8cd550d1.

Fridman, L. (2018). *Human-Centered Autonomous Vehicle Systems: Principles of Effective Shared Autonomy*, https://arxiv.org/abs/1810.01835 (accessed March 23rd, 2021).

Friedman, B. (1996). Value Sensitive Design. *Interactions* (November-December, 17-23).

Friedman, B., Kahn, P. H. Jr., Borning, A. (2008). *Value Sensitive Design and Information Systems*. In K.E. Himma, & H.T. Tavani (Eds.), *The Handbook of Information and Computer Ethics* (69-101), John Wiley & Sons. https://doi.org/10.1002/9780470281819.ch4.

Friedman, B., & Hendry, D. G. (2019). *Value sensitive design: Shaping technology with moral imagination*. MIT Press.

Friedrich, B. (2016). *The Effect of Autonomous Vehicles on Traffic*. In M. Maurer, J. Gerdes, B. Lenz, & H. Winner (Eds.), *Autonomous Driving* (317-334), Springer. https://doi.org/10.1007/978-3-662-48847-8_16.

GDPR – General Data Protection Regulation, (2016). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN (accessed March 23rd, 2021).

Glancy, D.J. (2012). Privacy in Autonomous Vehicles. *Santa Clara Law Review, 52*(4), 1171-1239. https://digitalcommons.law.scu.edu/lawreview/vol52/iss4/3.

Goggin, G. (2019). Disability, Connected Cars, and Communication. *International Journal of Communication*, *13*, 2748-2773. https://ijoc.org/index.php/ijoc/article/view/9021.

Habermas, J. (1971). Knowledge and human interests: a general perspective. In Id., *Knowledge and Human Interests* (301-17), trans. by J. J. Shapiro, Beacon Press.

Horizon 2020 Commission Expert Group to advise on specific ethical issues raised by driverless mobility (E03659), (2020). *Ethics of Connected and Automated Vehicles: recommendations on road safety, privacy, fairness, explainability and responsibility*, https://op.europa.eu/en/publication-detail/-/publication/89624e2c-f98c-11ea-b44f-01aa75ed71a1/language-en (accessed March 22nd, 2021).

van den Hoven, J. (2008). Moral Methodology and Information Technology. In K.E. Himma, & H.T. Tavani (Eds.), *The Handbook of Information and Computer Ethics* (49-67), John Wiley & Sons. https://doi.org/10.1002/9780470281819.ch3.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*, 389-399. https://doi.org/10.1038/s42256-019-0088-2.

Krontiris, I., Kalliroi, G., Kalliopi, T., Zacharopoulou, M., Tsinkitikou, M., Baladima, F., Sakellari, C., & Kaouras, K. (2020). Autonomous Vehicles: Data Protection and Ethical Considerations. In *Computer Science in Cars Symposium (CSCS '20), December 2, 2020, Feldkirchen, Germany*, ACM. https://doi.org/10.1145/3385958.3430481.

Liu, N., Nikitas, A., & Parkinson, S. (2020). Exploring expert perceptions about the cyber security and privacy of Connected and Autonomous Vehicles: A thematic analysis approach. *Transportation Research*, *75*, 66-86. https://doi.org/10.1016/j.trf.2020.09.019.

Matthias, A. (2004). The responsibility gap: ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*, 175-183, https://doi.org/10.1007/s10676-004-3422-1.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1-38, https://doi.org/10.1016/j.artint.2018.07.007.

Morley J., Floridi L., Kinsey L., Elhalal A. (2020). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Science & Engineering Ethics, 26*, 2141-2168. https://doi.org/10.1007/s11948-019-00165-5.

Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mokander, J., & Floridi L., (2021). *Ethics as a service: a pragmatic operationalisation of AI Ethics,* https://ssrn.com/abstract=3784238 (accessed March 22nd, 2020).

Mozaffari, S., Al-Jarrah, O.Y., Dianati, M., Jennings, P., Mouzakitis, A. (2020). Deep Learning-Based Vehicle Behavior Prediction for Autonomous Driving Applications: A Review. *IEEE Transactions on Intelligent Transportation Systems*, 1-15. https://doi.org/10.1109/TITS.2020.3012034.

NTSB – National Transportation Safety Board (2020). *Tesla crash investigation yields 9 NTSB safety recommendations*, https://www.ntsb.gov/news/press-releases/Pages/NR20200225.aspx (accessed March 22nd, 2020).

Nunes, A., Reimer, B., Coughflin, J. F. (2018). People must retain control over autonomous vehicles. *Nature*, *556*, 169-171. https://doi.org/10.1038/d41586-018-04158-5.

Nyholm, S. (2018a). The ethics of crashes with self-driving cars: A roadmap, I. *Philosophy Compass*. https://doi.org/10.1111/phc3.12507 (accessed March 22nd, 2020).

Nyholm, S. (2018b). The ethics of crashes with self-driving cars: A roadmap, II. *Philosophy Compass*. https://doi.org/10.1111/phc3.12506 (accessed March 22nd, 2020).

Nyholm, S. (2018c). Attributing Agency to Automated Systems: Reflections on Human–Robot Collaborations and Responsibility-Loci. *Science and Engineering Ethics, 24*, 1201-1219. https://doi.org/10.1007/s11948-017-9943-x.

Owen, R., Macnaghten, O., & Stilgoe, J., (2012). Responsible research and innovation: From science in society to science for society, with society. *Science and Public Policy, 39*(6), 751-760. https://doi.org/10.1093/scipol/scs093.

Rannenberg, K. (2016). Opportunities and Risks Associated with Collecting and Making Usable Additional Data. In M. Maurer, J. Gerdes, B. Lenz, & H. Winner (Eds.), *Autonomous Driving* (497-517), Springer. https://doi.org/10.1007/978-3-662-48847-8_24.

Rosenfeld, A., & Richardson, A. (2019). Explainability in human–agent systems. *Autonomous Agents and Multi-Agent Systems*, *33***,** 673–705. https://doi.org/10.1007/s10458-019-09408-y.

Santoni De Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers of Robotics and AI*, *5*(15), 1-14. https://doi.org/10.3389/frobt.2018.00015.

Schoonmaker, J. (2016). Proactive privacy for a driverless age. *Information & Communications Technology Law*, 1-33. https://doi.org/10.1080/13600834.2016.1184456

Setchi, R., Dehkordi, M. B., & Khan, J. S. (2020). Explainable Robotics in Human-Robot Interactions. *Procedia Computer Science*, *176*, 3057-3066. https://doi.org/10.1016/j.procs.2020.09.198.

Taeihagh, A., & Lim, H. S. M. (2019). Governing autonomous vehicles: emerging responses for safety, liability, privacy, cybersecurity, and industry risks. *Transport Reviews*, *39*(1), 103-128. https://doi.org/10.1080/01441647.2018.1494640.

Thornton, S. M., Lewis, F. E., Zhang, V., Kochenderfer, M. J., & Gerdes, J. C. (2018). Value sensitive design for autonomous vehicle motion planning. In *2018 IEEE Intelligent Vehicles Symposium (IV)* (1157-1162). IEEE.

Umbrello, S., & Yampolskiy, R. V. (2021). Designing AI for explainability and verifiability: a value sensitive design approach to avoid artificial stupidity in autonomous vehicles. *International Journal of Social Robotics*, 1-10. https://link.springer.com/article/10.1007/s12369-021-00790-w.

Xu, W. (2021). From automation to autonomy and autonomous vehicles. Challenges and opportunities for human-computer interactions. *Interactions*, *28*(1), 49-53. https://doi.org/10.1145/3434580.