CONNECT - Trustworthiness

Autonomous vehicles, and the broader socio-technical systems that they will be a part of, are likely to have a deep, lasting impact on our societies. Trustworthiness and trust are key values that will play a role in shaping the development and deployment of autonomous driving systems. There has been significant recent attention to use of trustworthiness and trust as operational concepts in technological design as a means to minimize risk and ethical harms while maximizing the benefits. This is reflected, for example, in the calls and assessment criteria for trustworthy AI by the EU High-level Expert Group on AI (HLEG, 2019) as well as in calls for trustworthy AV systems (Fernandez Llorca & Gomez, 2021).

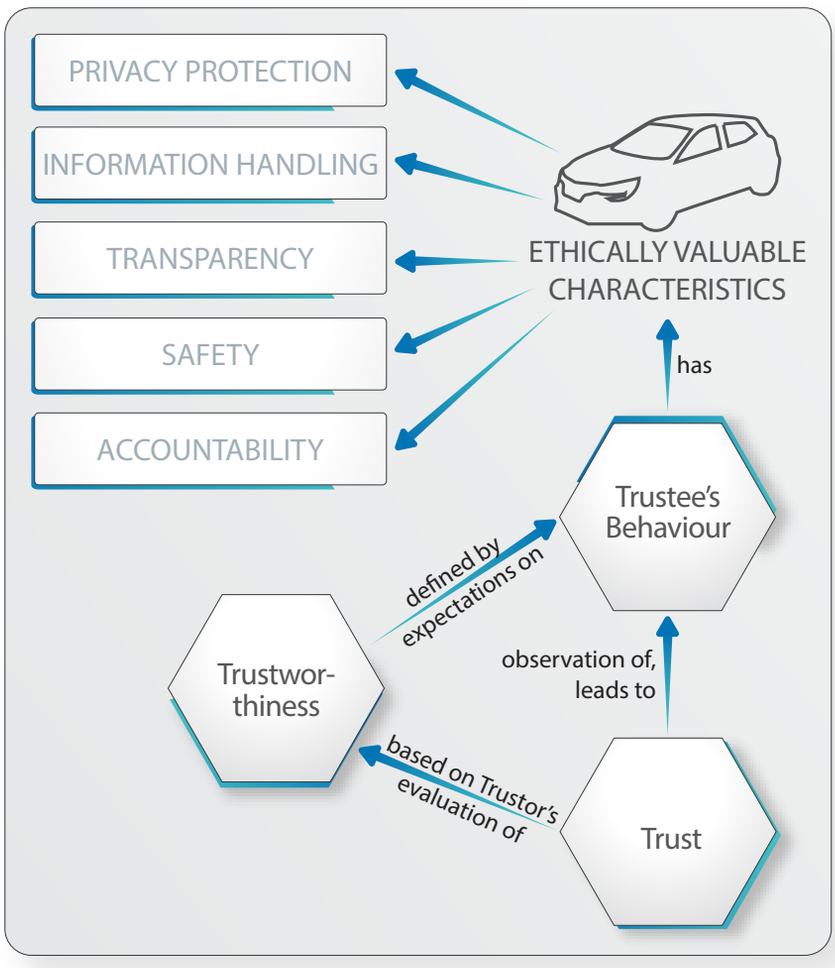
In CONNECT, a key aim is to enable a dynamic and continuous assessment of trust in a

CCAM (connected, cooperative, and automated mobility) system and to investigate mechanisms that provide increased trust assurances compared to today's systems. To achieve this, it is important to have a clear and precise definitional account of trustworthiness and trust that can be applied to, and is and practically useful in, the context of autonomous technological systems. The context of an autonomous system would demand that the account of trust and trustworthiness followed here is applicable to a diverse set of relationships – for example, relationships involving two users, one user and a part of the autonomous system, and two technological parts of the autonomous system (or two components). The diversity of trust relationships potentially involved in such a socio-technical system requires an account of trust that can go beyond anthropocentrism and describe, for example, trust relationships involving technical components.

One of the important aims here is, therefore, to provide an account that is applicable on a specific level and at a general level. At the specific level, the aim in developing this definitional account is to facilitate operationalization of the concepts of "trustworthiness and trust" for design of autonomous systems and components therein. At the general level it is to offer broad accounts of trustworthiness and trust that can be used across a large range of technical and practical situations relating to these autonomous systems.

Trustworthiness

In general, trust can be conceived as of a three-place relation involving a trustor (one who trusts), a trustee (one who is trusted), and the entrusted task or domain (Baier, 1986). Trustworthiness can be broadly conceived as a measure of the trustee's ability to achieve the entrusted task and respond to the trust placed in it by the trustor. Further, in some cases, the trust relationship may depend on the "entrusted task" to be conceived more broadly than just a performance outcome. Lee & See (2004), for example, advance a 3P model of trust in automation, signifying that the trustor's expectations from the trustee are a function of not just the performance (outcome) of the entrusted task, but also the process (through which the entrusted task was carried out), as well as the purpose for which the entrusted task was chosen and fits into the overall scheme of the technological system in consideration. It should not be noted though that for trust relationships to work successfully, trustor's expectations need to be appropriate or reasonable, otherwise there may be a threat for misuse and disuse (Lee & See, 2004). It is critical to avoid, for example, overtrust, where the trustor's expectations exceed trustee's capabilities.



In the case of autonomous systems, trustworthiness of the trustee can be further broken down into two components: competence and integrity (Kate Devitt, 2018). Here, competence implies technical aspects of the performance, such as reliability, accuracy, and so on. Integrity refers to the motives, goals, intentions of the trustee. For a non-human trustee, such as a technical component, a measure of integrity is the degree to which the expected behaviour (or past behaviour) of the trustee aligns with the goals of trustor. Together, the two components of competence and integrity, give a sense of the trustee's likelihood to meet the trustor's expectations.

In order to evaluate trustworthiness, it is critical that due attention is paid to the contextual conditions. For example, a sensor that is solely responsible for detecting an object in the path of the vehicle will need a higher degree of reliability than a sensor that is part of a system with some built-in redundancy through the use of series of such sensors that are utilized in conjunction to detect such an object.

Given this discussion, Trustworthiness can be defined as **the likelihood of the trustee to fulfil trustor's reasonable expectations in a given context,** where such expectations can be a function of the entrusted task, the process through which it was achieved, and the purpose for which the task was chosen. Further, the context within which the trust relationship operates also plays a key role in determining what "reasonable expectations" would amount to. Another factor that determines the scope of "reasonable expectations" are the ethical concerns and values in play in that context.

Ethical Values and designing for Trustworthy Autonomous Vehicle Systems

In the context of autonomous vehicle systems, even if a user does not have expectations regarding particular values, we might require that minimum standards are met. For instance, a user may not value protection of his/her private information, but a system that fails to protect privacy related concerns would still fail to be trustworthy. Privacy protection and appropriate information flow is central in addressing key ethical concerns, for example, of human autonomy and non-discrimination. Since autonomous vehicle systems depend on gathering and processing a significant amount of data of the vehicle a passenger is in, the environment, and even potential pedestrians on the road, it is imperative that such data is accessible only to those with authorized access and not accessible to potentially malicious or third parties without authorization or consent from the users. The aim in CONNECT is to design and develop privacy enhancing technologies that prevent such unauthorized access to collected data, and that also ensure controlled and restricted linkability of personal information or of data that may pose significant privacy risks for a user.

Similarly, there are other key ethical principles that determine the properties an autonomous vehicle system must exhibit in order to meet reasonable expectations, and consequently, be deemed worthy of trust. Potential examples of such key ethical principles include: Respect for human autonomy, prevention of harm, fairness, and explainability (Fernandez Llorca & Gomez,2021). Explainability, for example, is a key principle in consideration in CONNECT, through which we aim to ensure that there is transparency for the user regarding the purposes and capabilities of the system, particularly in relation to key decisions regarding handling of information and deployment of privacy enhancing technologies.

These key ethical principles can then be translated into properties or requirements for AV systems to be deemed trustworthy, such as: Human oversight, transparency, accountability, privacy protection, non-discrimination, technical robustness, safety, societal and environmental well-being.

In CONNECT, our aim is to provide more concrete and a further refined list of such key requirements and properties of a trustworthy CCAM AV system, as well as its components. Further, through our research within the CONNECT program, we also aim to demonstrate the methodology for identifying properties that enable both the design of trustworthy mobility systems as well as evaluation of trustworthiness of existing and/or future mobility systems and their components.

References

[1] Baier, A. (1986). Trust and Antitrust. Ethics, 96(2), 231–260. https://doi.org/10.1086/292745

[2] Fernandez Llorca, D., & Gomez, E. (2021). Trustworthy Autonomous Vehicles (JRC Research Reports No. JRC127051). Joint Research Centre (Seville site). https://econpapers.repec.org/paper/iptiptwpa/jrc127051.html

[3] HLEG, A. (2019, April 8). Ethics guidelines for trustworthy AI | Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[4] Kate Devitt, S. (2018). Trustworthiness of Autonomous Systems. In H. A. Abbass, J. Scholz, & D. J. Reid (Eds.), Foundations of Trusted Autonomy (pp. 161–184). Springer International Publishing. https://doi.org/10.1007/978-3-319-64816-3_9

[5] Kelp, C., & Simion, M. (2023). What Is Trustworthiness? Noûs, n/a(n/a). https://doi.org/10.1111/nous.12448

[6] Lee, J. D., & See, K. A. (2004). TRUST IN AUTOMATION: DESIGNING FOR APPROPRIATE RELIANCE. Human Factors, 46(1). https://trid.trb.org/view/755423

Budget

€ 5.7 Million

100% EU-funded

Consortium

17 Partners

9 Countries

Duration

36 Months

09/2022 - 08/2025

Partners

TECHNIKON · UBITECH · HUAWEI · ICCS · universität ulm · Red Hat · Trialog · DENSO Crafting the Core · intel · Suite5 · uni.systems · UNIVERSITY OF TWENTE. · FSCOM · STELLANTIS · Politecnico di Torino · System× · UNIVERSITY OF SURREY

Funded by the European Union

Funded by the European Union under grant agreement no. 101069688. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Find out more about CONNECT:

https://horizon-connect.eu

@connect_horizon

CONNECT Horizon Europe project 101069688