Received: 12 July 2023 | Revised: 21 December 2023 | Accepted: 28 December 2023

DOI: 10.1002/sys.21744

REVIEW ARTICLE

WILEY

Challenges for trustworthy autonomous vehicles: Let us learn from life

Imke Hoppe[1,2] | Willem Hagemann[1] | Ingo Stierand[1] | Axel Hahn[1,3] | André Bolles[1]

[1]German Aerospace Center, Institute of Systems Engineering for Future Mobility, Oldenburg, Germany

[2]Ludwig-Maximilians-Universität München, Munchen, Germany

[3]Department of Computing Science, University of Oldenburg, Oldenburg, Germany

Correspondence
Ingo Stierand, German Aerospace Center, Institute of Systems Engineering for Future Mobility, Oldenburg, Germany.
Email: ingo.stierand@dlr.de

Abstract

Current surveys indicate limited public and individual trust in autonomous vehicles despite a long tradition to ensure their (technical) trustworthiness in informatics and systems engineering. To address this trust gap, this article explores the underlying reasons. The article elaborates on the gap between trust understood as a social phenomenon and, in contrast, the research tradition aimed at guaranteeing (technical) trustworthiness. It discusses to what extent those research traditions in the social sciences and humanities have been recognized and reflected in systems engineering research to date. Trust, according to the current state of research in the social sciences and humanities, heavily relies on individual assessments of an autonomous vehicle's abilities, benevolence and integrity. By contrast, technical trustworthiness is defined as the sum of intersubjective, measurable, technical parameters. They describe certain abilities or properties of a system, often according to respective technical standards and norms. This article places the "explainability" of autonomous systems in a bridging role. Explainability can help to conceptualize an integrative trust layer to communicate a system's abilities, benevolence and integrity. As such, explainability should respect the individual and situational needs of users, and should therefore be responsive. In conclusion, the results demonstrate that "learning from life" requires extensive interdisciplinary collaboration with neighboring research fields. This novel perspective on trustworthiness aligns existing research areas. It delves deeper into the conceptual "how", dives into the intricacies and showcases (missing) interconnectedness in the state of research.

KEYWORDS
autonomous vehicles, interdisciplinarity, trust, trustworthiness

1 | INTRODUCTION

Why is it widely accepted that people get their driver's license after a very limited set of driving lessons—for example, 9 h[a] in Germany?[2] Is it the human intelligence and the related behavior we trust in, so that the corresponding risks? The situation is quite different for trust in autonomous vehicles: Empirical studies demonstrate rather low individual trust in autonomous driving[3–7] and "indicate […] that a

substantial proportion of people remain very sceptical about the idea of travelling in a fully automated vehicle [3, p. 52]". In Germany, for example, around half of the participants in a quantitative market survey ($n = 1000$) indicated that they would not use an autonomous vehicle (e.g., for public transport) at all, or were unsure.[8] In the United States, the American Automobile Association found that 85% of surveyed Americans ($n = 1107$) were "fearful or unsure of self-driving technology". This finding is independent of the car model tested, and the level of non-acceptance has remained steady over the past few years.[9]

1.1 | Relevance of trust and trustworthiness for systems engineering

The empirical studies cited above call for a deeper and more grounded understanding of the current and future role of trust and trustworthiness in systems engineering, especially since autonomous vehicles are just one representative of a completely new class of autonomous systems that will cause substantial changes in our daily lives. Since only a small proportion of the surveyed persons will have had their own experiences with autonomous vehicles, it is all the more important to note the basic mindset on which first experiences with the technology will resonate.

Why is it important to create more trust? From an economic point of view, this is perhaps motivated by achieving higher sales. However, deeper knowledge about the evolution of trust should not be misused as a powerful manipulative force ("social engineering") driven by the desire for higher and faster market penetration of this technology. From a normative scientific viewpoint, trust in the context of artificial intelligence (AI) and transportation[b] [10] needs to consider the individual "AI literacy" of users and relevant societal actors. "AI literacy" is a term developed for the broader scope of AI-based systems in general, as "AI is becoming increasingly integrated in user-facing technologies" and challenges the users' abilities to interact with the respective systems [11, p. 1]. Thus, users should be enabled to make informed choices about whether and how to use autonomous vehicles,[11] and not be persuaded to trust a magical black box. Accordingly, we stress here that the lack of trust is not a matter of persuasion, but of communication.

Ongoing discussions inside the technological sphere name possible reasons for the surveyed mistrust towards or non-acceptance of autonomous vehicles. These discourses often assume that deficiencies in the existing technical functionality (such as sensor errors) and their lack of robustness in real-world settings are the main causations for mistrust.[10,12] The multidisciplinary and mostly social science-based research on the individual acceptance of autonomous driving agrees on the importance of robust technical functionality. It is seen as a precondition for trust, often evaluated in empirical studies as "perceived technical performance"[3] or "perceived usefulness".[7] However, empirical findings also indicate that negative framings (pre-attitudes) people have about the technology (e.g., "Autonomous cars are dangerous") and their overall mistrust towards these systems (e.g., "Overall, I would trust autonomous vehicle technology" with a respective Likert scale; see ref. [7]) overshadow the perception of usefulness[13] and tolerance of a certain rate of (minor) errors or unexpected behavior. In light of the Covid-19 pandemic, a comparison with vaccinations illustrates these cognitive "overshadowing" mechanisms (the related theoretical foundation is the cognitive dissonance approach; see ref. [14]): People who deem vaccination to be less useful often avoid or depreciate (scientific) information about the trustworthiness of vaccinations because they mistrust the actors who promote vaccination in general (governments, media, etc.).[15–17]

On the other hand, it is assumed that "overtrust" [18, p. 40] is just as apt to lead to mistakes during use as a lack of trust is. The case of "overtrust" becomes prominent in media publics particularly for fatal errors involving automated cars and assisted driving—say, when drivers fully rely on the car while using their mobile.[18] A scandalizing media debate triggered by events like these can then negatively impact societal trust in the technology.[19] To combat the loss of societal trust in light of single events like these, systems engineering has started to integrate results from machine ethics, law and political science from an early stage,[20] with the goal of ensuring a high level of harmonization of the technology with existing norms, values and ethical standards (especially with respect to safety and security; see ref. [21]).

On the very first brink of a new era in transportation, that is, the transition from automated to fully autonomous vehicles,[22] individual trust can serve as a bridge for overcoming psychological barriers. The same goes for societal trust, for example, for political decision makers who set the course for legal frameworks allowing autonomous driving[23] and for clarifying legal consequences in the case of wrong decisions or accidents for which autonomous vehicles are partly or fully responsible ("liability").[24] However, individual and societal trust are key not only for the very first stage of technology acceptance,[25] as more process-oriented and long-term models demonstrate.[13] Users will be confronted with continuing uncertainties even after they have fully decided to use autonomous vehicles (such as buses, trains and cars), due to the inherent nature of AI technology.[26] Autonomous vehicles will be able to make their own decisions. For this purpose, they should cooperate with humans in a trustful but informed way, including a certain rate of failure tolerance, especially in safety-critical situations. The more autonomous vehicles cross the threshold from being "technological tools"[27] with limited functionality to being "autonomous teammates" with their own strengths and weaknesses, that is, from automated to fully autonomous vehicles, the more that dialogue between humans and machines should help each to get to know and trust the other.[28]

1.2 | Scope, motivation and structure of the paper

Thus, the role of trust is increasingly considered to be an important "human factor" in systems engineering.[29] However, there is often a considerable gap between trust understood as an individual and social phenomenon, and "technical" trustworthiness. The latter is often understood—at least in the engineering disciplines—as a technical term. It is thus preferably defined such as to be measurable by technical parameters, or at least as a constraint that can serve as a driving factor in the design process. There is often the implicit assumption that these "technical" trustworthiness attributes have an impact on the

trust users accord to the system. The question as to what this impact looks like, and even more importantly, of how to empirically validate this assumption, often remains unclear. Therefore, it is important to broaden out our view on the many meanings and definitions of trust which are relevant in the context of autonomous vehicles, and to ascertain to what extent these different meanings have been recognized and reflected in systems engineering research so far. Beforehand, it is important to acknowledge that one cannot come up with a single silver bullet definition of trust and trustworthiness. Especially, when looking at trust as a social phenomenon, it is vital to appreciate that the construct is not dichotomous but multi-dimensional with many shades.

Informatics and computer science can look back on a very long and successful tradition of interdisciplinary research since their inception.[31,32] Facing complex challenges, systems engineering broadened its view early on and learned from other fields of research, for example, from physics in the case of positioning with inert sensors. Autonomous systems also impose new types of complex and challenging research objectives. And so, this article asks how systems engineering can design systems such that humans and autonomous systems cooperate in a trustworthy way. Consequently, this article develops a broad view on the various scientific disciplines working on the subject of trust and trustworthiness, and discusses the extent to which their results, theoretical models and methods can be integrated into systems engineering more comprehensively than at present. By employing two fictitious examples, the paper illustrates how this integrative perspective reveals current shortcomings and generates added value for research and interdisciplinary education. A holistic view, based on interdisciplinarity, of the disciplines engaged in the research field shows up new challenges for future engineering and can thus inspire innovations in the respective fields, including for education in companies and universities.

To accomplish this vision, we propose that autonomous systems be understood as "living entities" with an integrated role in our societies. The term "living entities" serves here as a metaphor to focus on the long-term perspective of such systems and their increasing capabilities to fulfil tasks and interact with humans and society. It points to several learning processes that both autonomous systems and humans will have to undergo within their respective "ecosystems", including the physical environment, other autonomous systems and human actors (for embeddedness into larger systems and "systems of systems," see, e.g., ref. [33]). Because of this embeddedness of autonomous vehicles—as living entities—and the key role of trust as an objective, this article is a call for interdisciplinary research which goes beyond the existing state of research. We would like to draw attention to the central feature and characteristic of this technology, namely the autonomy of these systems,[34] and illustrate new research routes to determine their trustworthiness.

2 | STATE OF RESEARCH

In this chapter, we will discuss the state of research on the role of individual trust in autonomous vehicles, and the extent to which these results have been integrated into systems engineering to date. As a wide variety of disciplines are working on trust and trustworthiness, our literature review starts with a clarification of the central terminology.

A systematic literature review was carried out to map definitions and aspects generally considered to be important for autonomous vehicles in relation to trust and trustworthiness. As the search string "trust" and "trustworthiness" turned out to be too general in combination with "autonomous", we used a third search term ("transportation," "vehicles," "cars," "bus," "public transport") in several databases (ISI Web of Science, Google Scholar, Springer Link, Scopus, ResearchGate). We manually selected relevant articles, with an emphasis on (a) the perspective of systems engineering and to what extent trust and trustworthiness have been reflected or integrated to date; and (b) on publications deemed to be classics in their field (e.g., the sociologist Niklas Luhmann's "Trust and Power" or Baier's essay "Trust and Antitrust"). All in all, we considered over 200 publications as the starting point for the analysis, while exploring certain aspects in greater depth later on (e.g., related ISO standards). It turned out that the majority of relevant articles with reference to systems engineering were published via Springer Link. Meanwhile, books and articles from the social sciences were often to be found in publications available via Web of Science. For the most part, they have a more general scope in relation to AI-based systems and are rarely focused on autonomous (transportation) systems. A major exception to this rule is research articles from psychology, which provide empirical evidence for trust-related constructs and are often published in technology-oriented journals. Humanities research is rarely incorporated in systems engineering publications, with the exception of philosophy and ethics (machine ethics). In Chapter 2, we summarize the literature and develop a structure to provide a systematic assessment in light of the research interest.

As a general lesson from the literature analysis, trust needs to be understood as something that develops over time, changes with experience and indeed shapes our experiences in turn. It has emotional and cognitive facets and fulfils—moreover—important normative and legal functions and has corresponding implications.[35] Thus, it has very different meanings in different disciplines and research fields.[13,30]

2.1 | Technical trustworthiness defined from an "engineering" viewpoint

The system property *trustworthiness* has been of particular interest in the context of developing safety-critical autonomous systems.[36] In the engineering disciplines, it is commonly agreed that trustworthiness is a collection of other well-established technical properties. The German Association for Electrical, Electronic & Information Technologies (VDE), for instance, defines it as a combination of reliability, availability, maintainability, safety, security, privacy, usability, explainability, ethics and robustness.[37] Many of these properties have been investigated for decades in engineering. Avizienis, Laprie, Randell and Landwehr[38] provide a comprehensive definition and survey of related work for most of these properties. For example, safety is defined as the "absence of catastrophic consequences [of using the system] on the user(s) and the environment". This definition is widely accepted in the engineering

community, and many norms and standards in various application domains (automotive, avionics, industry) refer to it. For example, ISO 26262[39] defines processes and methods that must be followed in the system engineering in order to demonstrate the safety of newly developed vehicles to certification authorities.

Security is also a well-established research and engineering discipline, probably best known by the encryption algorithms employed in many of today's applications. According to Anderson,[40] however, security engineering "is about building systems to remain dependable in the face of malice, error or mischance" and concerns methods, processes and tools to ensure confidentiality (no unauthorized parties get access), integrity (changes can be performed only by authorized parties) and availability (can be used by any authorized party) (see, also ref. [41]). As an interesting side note, Anderson states that "a trusted system […] is one whose failure can break the security policy, while a trustworthy system […] is one that won't fail".

Integrity is generally understood as the *absence of improper system alterations*.[38] Aside from its meaning as mainly malicious alteration in the context of security, integrity is also often considered as part of safety, with a focus on erroneous alterations: A degraded system state due to a failure can be seen as an improper alteration. Data bases are a classical domain where integrity plays a role. Similar integrity issues come into play for the more recently popularized blockchains. As in these examples, integrity issues are often related to concurrent access to data and other resources by multiple agents, with mutual exclusion[42] and Byzantine faults[43] as well-known problem classes. The SafeTRANS Roadmap[44] has identified increasingly complex notions of integrity, such as plan and learning integrity, as a main research challenge on the road to autonomous vehicles.

The property of robustness is closely related to integrity but focuses on the environment in which a system is functioning. The IEEE Standard Glossary of Software Engineering Terminology defines it as "the degree to which a system or component can function correctly in the presence of invalid inputs or stressful environmental conditions".[45] The term has a strong similarity to the term resilience. An overview of existing notions and principles can be found, for example, in ref. [46]. A large body of work exists in various subdisciplines such as programming and machine learning.[47]

As we will see later, the application of the above-mentioned system properties to some aspects of autonomous systems is not yet well understood in the engineering domain. This particularly concerns explainability and ethics. The former property is commonly understood as the ability of a system to provide information on how it produces results, or more generally, about its behavior. A main focus of current research is to make the information the system produces available and accessible to a potential "user" of explanations. The authors of ref. [48], for example, provide a conceptual framework of explanation using "explanation patterns" and discuss how it can be established in system design. Although it is known that such technical accounts of explainability might not be sufficient,[49] these aspects are left as an open research question.

Also concerning ethics, we observe some loose ends left by the engineering disciplines, such as in the safety standard ISO 26262 mentioned above. There, we find the fundamental concept of *unreasonable risk*. It is defined as a "risk judged to be unacceptable in a certain context according to valid societal moral concepts".[39] While the authors obviously acknowledge that the acceptability of a risk must be based on societal and moral standards, they left the important question open as to how this should be operationalized in the development process. In recent work, Koopman emphasizes an interesting direction for tackling this problem through the introduction of metrics and indicators in order to give means of measuring *acceptable risk*.[50]

Another key question in relation to research from the engineering perspective is whether autonomous systems should have ethical decisions implemented in them. The well-known trolley problem captures a series of ethical dilemma situations where any of two undesired choices must be taken, and has recently attracted renewed attention through the Moral Machine platform.[c] The report of the Ethics Commission of the German Federal Government stresses that systems should be implemented such that human life gets highest priority in critical situations, compared to, say, animals and property.[51] Referring to the legal system in the US, D'Amato et al.[53] argue that "if human harm is imminent and breaking the traffic code is necessary to reduce that harm […] then the ADS [automated driving system; authors' note] may legally be programmed to break traffic code". Their aim is to resolve the ethical dilemmas by proposing principles for handling exceptional driving situations that can be translated into engineering requirements. On the other hand, ethical "programming" is considered dangerous. For example, algorithms that qualify human characteristics (like gender, age, etc.) should not be developed. Instead, it should be ensured that ethically relevant situations are avoided.[51] As a consequence, the recently published German regulation on the approval and operation of motor vehicles with autonomous driving functions[52] specifies well-defined operating domains with zero or minimal risk for such situations as a fundamental condition for obtaining permission to put autonomous vehicles into operation. Finally, in his article "Can you program ethics into a self-driving car?",[53] Goodall bridges the gap between ethics and explanation, arguing that the ethical dilemmas of vehicle automation are a solvable problem as long as there is a rational justification for the action of a vehicle that also takes into account the ethical implications.

2.2 | "Nontechnical" perspectives on trust and trustworthiness

The *social sciences* and the *humanities* have each established their own research fields and key terms on the overarching issue of the role and meaning of AI, which by their nature have an interdisciplinary structure. Examples of such fields in the social sciences include "robopsychology",[54] "Sozionik",[55] and "deep mediatization".[56] Often, these research fields are embedded within broader research traditions, as is the case, for instance, with the "sociology of technology"[57] or the "psychology of technology"[58] in the social sciences. Within these fields, autonomous systems are often seen as very specific research objects within existing research traditions. The humanities have also explored the subject of AI from an early stage (e.g., within the research field of "digital humanities"[59]). These fields offer critical

views on societal hopes and fears connected to the technology (e.g., on dystopia vs. utopia in relation to AI; see ref. [60]), as well as discussions on epistemological questions connected to AI technologies (e.g., definitions of what "autonomous" is). The term "digital humanities" stands for a shared use of methods between the humanities and informatics.[d] A major strand of humanities engagement with AI is subsumed as "ethics of artificial intelligence"[61] and "machine ethics".[62] In the following, we will explore both perspectives—the humanities and social sciences—to develop a general idea of how the basic terms (trust and trustworthiness) are understood in these domains, and how far they differ from the technical account.

2.3 | Trust and trustworthiness in the humanities

In ancient Greek philosophy and in Christian moral theory, trust was primarily examined in the context of faith and fidelity. In the modern period, moral and political philosophers such as John Locke or Thomas Hobbes looked at trust in government, contracts and contractors. The prelude to the contemporary debate on trust and trustworthiness was probably provided by A. Baier's essay "Trust and Antitrust" in 1986.[63] Despite the diverse positions and theories among philosophers, there is widespread agreement that trust is a relation between two parties, the trustor *A* and the trustee *B*, where the scope of trust is often restricted, for example, to certain actions, domains or valued things *C*. That is to say, trust is a three-part relation: "trustor *A* trusts trustee *B* with respect to *C*".

It is interesting how the current philosophical debate on trust and trustworthiness shifts from relatively simple models of trust towards an increasingly in-depth analysis of the mutual relationship between trustor and trustee in terms of trustworthiness (see, ref. [64] for an excellent overview). To start with, trustor-centric accounts of trust primarily focus on the trustor's attitude towards the trustee and define trust as a hope that the trustee will prove to be trustworthy.[64] Already this account allows us to discuss some characteristics of trust that can be located on the side of the trustor. The trustor has epistemic uncertainty as to whether the trustee really acts as hoped.[65,66] Typical representatives of such trustor-centric theories are risk-assessment theories, where trust is the assumption of a low risk of being betrayed.[64] The problem with such theories is that they cannot appropriately explain the difference between mere reliability and trust (see ref. [64] for references for this claim).

Trust is generally considered to be stronger than reliance: While a failure of reliance is considered as a disappointment, "trusting can be betrayed, or at least let down" [63, p. 235]. Motives-based accounts of trust shift the focus to the trustee and their motives to prove themselves trustworthy. Hence, trustworthiness comes into focus. What characteristics must be present on the side of the trustee to be considered trustworthy? First of all, the trustee must have the necessary competencies in the domain of trust *C*.[63] A trustee who is not capable of acting in a certain domain as expected is not to be trusted.[63] In addition, the trustee must have reasons to act as expected. Baier[63] sees goodwill as the most important motive governing the trustee's action. Other motives can be benevolence, honesty, conscientiousness and integrity, but also moral obligations or virtues.[67]

Even more elaborate accounts draw back again from the trustee to the trustor. According to Jones [67, p. 67], a trustee takes the fact that the trustor counts on them as a reason in their "motivationally efficacious practical deliberation" to act as if they were being counted on. From this perspective, persons are trustworthy if they, as soon as they are counted on, take this as a compelling reason to act as if they were counted on. Jones' approach is called a trust-responsive account because the reasons for the trustee's actions arise from the trust relationship rather than from motives like benevolence. According to this approach, a person remains trustworthy even if they are unable to act as expected, provided they "have some excusing explanation for why they did not" [67, p. 71].

2.4 | Trust and trustworthiness in the social sciences

"How does a trustor recognize whether she or he can trust?" Communication scholars like Blöbaum [30, p. 10] point to both interaction partners: what counts as trustworthy depends on the expectations of the trustor, and at the same time on the characteristics of the trustee. One of the most prominent definitions of trust from the social sciences applied to the field of autonomous vehicles[18,28,29,68] comes from Mayer et al.[69] in the area of organizational research. Their definition states that trust *"is the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other part"* [69, p. 712]. Trustworthiness is—for example, according to Mayer et al.—a related concept and describes characteristics of the trustee seen from the perspective of the trustor—either from a rather subjective viewpoint or from a more objective "instance" (as in the case of an autonomous driving certification authority). The definition is thus close to the above-mentioned definitions from philosophy. Trustworthiness is always accorded with respect to certain objectives, according to Mayer et al., in relation to the abilities, the benevolence and the integrity of an automated vehicle.[69] Thus, trustworthiness is attributed when a trustor assesses that an autonomous vehicle has these three characteristics.[13] Here, again, the focus on benevolence and integrity is close to the understanding of the philosophical author cited above.[67] One of the main differences between a humanities perspective and a social science perspective is that the social sciences are interested in empirically testing the named causalities, such as: Which factors cause the attribution of trustworthiness? This places emphasis on the perception of humans, and to what extent these perceptions are related to the factual given properties.[35] The empirically tested "micro-causalities" are integrated into broader theoretical models (see Figure 1 for an overview), which stem here from communication studies, working at the intersection between psychology and sociology,[30] and continue studies on the technology acceptance model (TAM).[7] Hoff and Bashir[13] stress the cyclical nature of the relevant processes. Blöbaum[30] points to the major role of reputation markers, which are

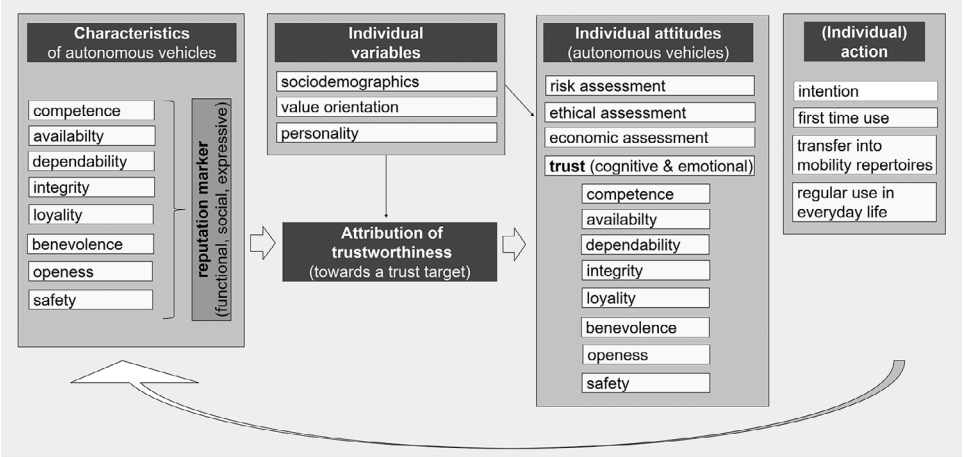

FIGURE 1 Integration of empirically informed assumptions into theoretical models of trust and trustworthiness (own visualization).

key for digital communication (also named human-computer interaction) and have been found to be a major predictor for the emergence of trust. Figure 1 summarizes empirically informed assumptions about the basic "paths of trust" on an individual level. It begins on the left with an object (an autonomous vehicle), which has certain (technical) characteristics (e.g., the ability to detect street signs correctly). Based on these symbolic reputation markers, a human user deems the object (e.g., the autonomous vehicle) to have these characteristics or not (e.g., it will detect street signs correctly) and assesses its trustworthiness accordingly. However, individual variables (personality, personal values, etc.) will also influence the extent to which someone will assess an object as trustworthy.

The attribution of trustworthiness can lead to feelings and cognitions of trust ("I trust the autonomous bus."). Once again, individual attitudes (understood as stable cognitions towards broader and more general objectives, e.g., privacy) moderate the path from "attribution of trustworthiness" to factual and experienced trust. Trust, alongside other individual variables (e.g., attitudes to risk), can lead to behavioral consequences, such as using an autonomous bus for the first time. For autonomous buses, there is empirical evidence that trust is the major predictor for developing a positive attitude towards the technology in general, which in turn influences the behavioral intention to use autonomous buses.[70] As soon as first-hand experiences are available, trust is probably founded less on reputation markers than on "real" experiences. From a meta-study by,[35] there is evidence that users deeming autonomous vehicles to have the relevant capabilities is key for humans to accord trust, meaning that the most important thing an autonomous vehicle can do to foster trust is to have convincing capabilities. The attribution of abilities by users is empirically measured by the reliability of the system and the error rate. However, the meta-study reveals the importance of individual traits and states as well, especially emotive factors such as the general attitude towards the technology as such. For example, if someone has problems with focused attention, this will influence the interaction between the vehicle and person and thus the evolution of trust.

3 | KEY INGREDIENTS FOR TRUST

To summarize, trust plays an outstanding role in the humanities and in the social sciences, which includes groundbreaking work from nearly all classics in the field as well as more specific and applied research, aiming to better understand the role of trust in the context of autonomous vehicles beyond a purely technical meaning. The humanities, especially philosophy, ask, for example, *when trust is justified*, the social sciences investigate *when and why humans grant trust* to others, and engineering is interested in *how autonomous systems are to be developed so that stakeholders justifiably attribute trust*. When the cited authors from philosophy—in contrast to the social science models discussed—are interested in trustworthiness as a property and less in the attribution of trustworthiness,[64] this does not reveal an incompatible difference of definitions, but reflects the different ambitions and interests of the two epistemic traditions. Whereas social science often empirically explores the extent to which an actor attributes trustworthiness to the other, philosophy looks for necessary and sufficient characteristics of trustworthiness on the basis of theoretical and logical considerations.

Bringing together both perspectives, individual trust and trustworthiness can basically be described as the process-oriented, mutual exchange of information between a minimum of two agents (e.g., a passenger/certification auditor and an autonomous vehicle) towards a certain scope ("trust target").[30] Thus, trust depends deeply on interaction,[13] in which—according to Mayer et al.,[69] but also in line with[63]—the ability, benevolence and integrity of the trustee (e.g., the autonomous vehicle) is positively assessed by the trustor (e.g., the passenger or the auditor of a certification authority). Technical trustworthiness is defined within the engineering tradition as the sum of intersubjective, measurable, technical parameters of a system, which describe certain properties of it, often according to respective technical standards and norms. From the viewpoint of social science models—as discussed above—the main question is to what extent these properties are perceived as such by (lay) humans, and to what extent they lead to deeper or more trust. There is broad consensus in the

social science literature on the acceptance of autonomous driving that technical parameters (functionalities) are no guarantee for the emergence of trust. In the humanities as well, trust depends deeply on the relationship of trustor and trustee, their motives and expectations. In the following sections, we discuss which "ingredients of trust" seen as central in humanities and social science research have been recognized by systems engineering and to what extent they have been integrated into systems engineering.

3.1 | Abilities, benevolence and integrity

The perspectives of the humanities and of social science have stressed the importance of the abilities, the benevolence and the integrity of an object (such as an autonomous vehicle) for the emergence of trust.[69] Will people develop trust towards the object itself, towards a vehicle like a car, bus, train or ship? Or will trust be granted to the science and/or technology behind the object, the companies who are producing and operating it, or the authorities who are certifying it? Presumably, the scope of trust will shift alongside the diffusion of the technology and will be strongly impacted by the attendant societal discourse about it.[19] Coming back to the technical understanding of trustworthiness, the following sections ask to what extent systems engineering models, methods and research have addressed these three fundamental ingredients of trust so far.

3.1.1 | Trust in the ability of an autonomous vehicle

Mayer et al. [69, p. 717] describe ability as a "group of skills, competencies, and characteristics that enable a party to have influence within some specific domain". They stress the task- and situation-specific nature of trust, while Blöbaum [30, p. 11] clarifies that competencies are the "ability of the relevant parties to fulfil their tasks" with respect to a certain situation. Systems engineering faces the challenge of defining the major and minor skills and competences (functionalities) a system should have. The well-established routine for doing this is a requirements analysis, in which functional requirements from manifold stakeholder perspectives (e.g., lay users) are formalized and quantified. The implemented functionalities can thus be verified and validated by formal methods. In seeing trust as related to the core technical functionality, it is a logical step to infer that a main aspect of *ability* is *reliability*,[12] which means, for example, that a driver trusts (relies on) the vehicle's brakes, as they guarantee the correct functionality of an important component of a software system[12] and thus must be engineered properly in order to minimize the risk of any harm caused to people. The engineering disciplines have well-established tool boxes to ensure reliability of the products they are building. For example, there are standards such as IEC 61508,[71] which provide guidance on how to address the safety-related aspects in the development process, such as the assessment of risks, derivation of corresponding safety requirements, and adequate design steps to ensure that the developed system fulfils these requirements. These standards have been adopted in various domain-specific standards such as ISO 26262 for the automotive domain[39] and ISO 17894 for the maritime domain,[72] where they define the state of practice.

Nevertheless, there are many open challenges in extending these "tool boxes" to incorporate support for the development of autonomous systems (see, e.g., the treatment of "automation risks"[73]), considering the wider objective of research on autonomous systems—namely their autonomy and their complexity as consisting of many units with specific functionalities. For example, brakes are used no longer by humans, but by a decision-making, intelligent and autonomous unit within the broader system.[12] So even if someone trusts the functionality of one system component, it is only a prerequisite for the trust given to the entire system like a car, train, or bus, which decides autonomously when and how the system's components are used and how well and reliably they work together.[12]

Returning to the social science perspective, it remains an open question which of the implemented abilities are perceived as such by human users (what Mayer et al. [69, p. 718]) term "perceived abilities": Which are the core functionalities—from the viewpoint of users/auditors? A large body of literature and related empirical studies can pinpoint individual design-related aspects. Kohn et al.,[28] for example, provide an overview of the TiA (Trust in Automation) construct and, for example, the extent to which embodied feedback, such as avatars, could help. However, the interplay of the various hardware and software components is rarely considered. The existing studies which take up this challenge and undertake a holistic, qualitative approach mostly have very small sample sizes and are thus rarely generalizable (e.g., $n = 12$).[68,74]

3.1.2 | Trust in the benevolence of an autonomous vehicle

According to Mayer et al.'s[69] definition, benevolence asks for the "good will" of the other party to an interaction. The authors state that "[b]enevolence is the extent to which a trustee is believed to want to do good *to the trustor*, aside from an egocentric profit motive" [69, p. 718]. As it is not always possible to draw black-and-white distinctions between good and bad, Blöbaum[30] suggest focusing on "intention" or "objective" instead, meaning to what extent the trustor evaluates the intention or objective of the trustee as matching his/her own ("Does the trustee want the same as me?"). When dealing with the benevolence of a technical system, more objectifiable characteristics are needed for systems engineering, and so ethics and legal studies move to the fore of interdisciplinary research on autonomous vehicles. In the context of implementing AI into autonomous systems with a physical backbone, the IEEE uses similar terminology ("beneficial to people and the environment") and emphasis: *"As the use and impact of autonomous and intelligent systems (A/IS) become pervasive, we need to establish societal and policy guidelines in order for such systems to remain human-centric, serving humanity's values and ethical principles. These systems must be developed and should operate in a way that is beneficial to people and the environment, beyond simply reaching functional goals and*

addressing technical problems. This approach will foster the heightened level of trust between people and technology that is needed for its fruitful use in our daily lives."[75] In order to make "benevolence" a fruitful term for systems engineering, ethical and legal research topics become crucial in their ability to provide specifications on which behavior counts as "good" or "bad" and thus furthers informed and well-calibrated trust.

3.1.3 | Trust in the integrity of an autonomous vehicle

In this context, integrity means that the system acts according to norms, standards and principles that are important for the trustor and defined beforehand[69]: "The relationship between integrity and trust involves the trustor's perception that the trustee adheres to a set of principles that the trustor finds acceptable." Integrity is a concept with a strong connection to time: the perception of integrity assumes a pre-post evaluation between norms, standards and principles defined beforehand and the factual behavior shown in specific contexts. Accordingly, trust is conceptualized as predictability,[12] assuming that (positive) user experiences[26] in prior situations and/or pre-existing attitudes (e.g., avoiding "causeless stops" is important on the road) create user expectations about future decisions of autonomous vehicles, which should be fulfilled by the system.

Although this meaning differs from the technical one as stated above, ensuring that the system behaves according to norms and standards is nonetheless a central—and well-established—aspect of systems engineering. As the discussion about technical trustworthiness above has shown, however, particularly in relation to ethics, it is outside the scope of the engineering disciplines to define these norms and standards in the context of autonomous vehicles.

3.2 | Trust and explainability

Even though engineers have succeeded in technically integrating parts of the three key ingredients—abilities, benevolence and integrity—into an autonomous system, this does not necessarily imply that the user or auditor will attribute trust to the autonomous system. One of the major reasons for this could be that users and even auditors are not always put in a situation in which they are able to perceive and observe all the sophisticated engineering hidden behind the surface and therefore unable to assess the degree of benevolence, integrity and functional excellence with which a system is acting. Moreover, current research on safety verification and explainability of AI systems[76,77] shows that system engineers themselves face the "black box problem" of AI-based system components.[78] The following examples demonstrate the need for an additional conceptual and integrative layer—namely, explainability—to communicate technical measures undertaken to make the system trustworthy with respect to its ability, benevolence and integrity.

Example 1. Imagine a person sitting in an autonomous car. She or he looks out of the window and sees that the car is driving through a busy area with many children and elderly people on the sidewalks. The skin conductance measurement indicates to the car that the passenger is experiencing increased stress. The car reports back that the vehicle has not identified any risk and has been observing the speed limit. However, the skin conductance measurement still indicates stress.

Example 2. Imagine an autonomous bus driving down a country road. The bus slows down and passes a car that has been involved in a bad accident with several injured people. The passengers of the bus are informed via loudspeaker that the bus has sent an automatic emergency call. However, some passengers are nevertheless very worried about how to deal with the situation and what to do next.

In both examples, the systems portrayed demonstrate a high degree of abilities, benevolence and integrity. Both vehicles act within their capabilities (abilities), behave in accordance with laws and social norms (integrity), and even show a certain degree of benevolence towards the passengers and other road users. Furthermore, the vehicles produce explanations for their behavior that refer to the three key ingredients: The vehicles explain remaining within their respective capabilities, for example, according to the risk assessment as in the first example, or explain their actions in relation to social and legal norms, such as making an emergency call in the second example. Nevertheless, the people involved are not convinced and do not feel comfortable with the action taken by the autonomous vehicle, even though they all perceive the same situation.

How can we provide more helpful explanations for Examples 1 and 2? Various explanations are conceivable here, depending on the factual decision-making of the system. Should the vehicle stop and ask the passengers to help? What if the situation takes place at night and puts the passengers in a dangerous situation? Another variation is that the vehicle passes by and simply make an emergency call. How will the vehicle explain its decision to the passengers, especially in a case where it is not clear that the passengers agree with that decision?

These examples raise important questions for systems engineering while also illustrating the potential and the limitations of the "explainability approach", which we subsume into two important cross-cutting questions: (a) How can the system detect whether a situation needs explanations, or whether the situation is so socially and/or ethically complex that it requires the intervention of human responsibility instead; and (b) How can the system incrementally reduce its autonomy to assistance (e.g., as a dialogue system), helping the humans involved to make informed decisions? The following ideas illustrate the need for and interplay of explanations and/or dialogues.

- In Example 1, the passenger could communicate to the vehicle that she or he has an increased need for safety (that may not even be entirely rational), so that the vehicle can then either meet this need or justify why driving slower may not increase safety in the current situation.

- In Example 2, the dialogue could extend to legal and ethical aspects. While it is clear that the vehicle itself cannot provide first aid (capabilities), this does not generally apply to the passengers. The passengers could be legally obliged to provide first aid, so that the vehicle would have to stop in this case. However, this certainly does not apply if the passengers are unable to do so (for example an autonomous bus filled with schoolchildren) or if intervening would put them in danger. On the other hand, stopping the bus on a narrow road could delay the arrival of a doctor rushing to the scene. An autonomous vehicle would have to be able to understand all these factors, evaluate them and determine to what extent responsibility needs to be handed over to humans.

Apparently, the explainability of the system and the mutual exchange of explanations in dialogue situations play a central role in such a system. Engineering science has already identified explanations and the explainability of systems as an important tool for proving trustworthiness to humans. However, as the examples show, such an explanation can fail or remain insufficient due to a lack of shared knowledge and the individual expectations of humans. Eliminating this lack of knowledge requires a process-oriented mutual exchange of information—that is to say, a dialogue—in terms of explanations, as discussed beforehand. This leads us to the twin questions: What makes an explanation, and how can one build explainable systems? Here, once again, it is worth taking a look at the different disciplines in turn.

The *scientific explanation*[79] is a central object of research in the philosophy of science. Beyond the intrinsic motivation of understanding and analyzing the *scientific* concept of explanation, researchers in this field have also discussed and analyzed a rich fund of everyday examples of explanation. Since Aristotle, there has been broadscale agreement that an explanation is a type of argument consisting of logically connected true statements. The central question is what properties such an argument must have in order to be explanatory (explaining *why*) rather than purely descriptive (explaining *what*). The modern treatment of this question dates back on the epoch-making work of Hempel and Oppenheim. According to their deductive-nomological (DN) model,[80] an explanation is a logically sound argument that consists of a true *explanans* (the explaining statements) and an *explanandum* (the phenomenon to be explained) such that the explanandum is a consequence of the explanans. The explanans includes lawlike statements (laws) and conditioning statements that characterize the given circumstances under which the laws have to be applied.

The following decades were marked by an ongoing debate as to what exactly characterizes such lawlike statements. For example, statistical laws were considered and it was pointed out that lawlike statements must also possess the ability to support counterfactual and modal import[81]—shifting the focus to causal mechanical (CM) models of explanation.[79] Finally, the unificationist theory investigated the deductive component of the DN model and proposed that explanations be viewed as inferences along argument patterns rather than logical arguments.[82] In summary, a scientific explanation is a rule-guided act of reducing the explanandum to the explanans that requires the validity of both, the reduction rules and the statements of the explanandum.

A similar reduction of the explanandum to its constituting explanantia can often be found in engineering disciplines. However, a technical definition reveals a central problem: Explainability introduces an addressee, for whom the generated explanations should be understandable and relevant. This leads immediately to the social sciences and humanities. Köhl et al.[83] try to formalize (at least some of) the relevant aspects that occur in these disciplines in order to make them available for an engineering-centric discourse.

When social sciences have asked why individuals need trust, they have identified one of the main individual motivations: to give trust is to counteract the overwhelmingly complexity[84] of our social and technical world—as the ambition to understand everything would otherwise lead to inaction and/or serious turbulence in the routines and functionalities of (social) systems. This is of special importance in the field of systems engineering for autonomous AI-based systems, as these systems need to act on their own in this complex world. Satisfying this motivation, explainability can help to foster an "informed trust" and enable humans to experience both cognitive relief and better understanding of the abilities, benevolence and integrity of an AI-based system.

4 | SUMMARY AND OUTLOOK

As a general lesson learned for systems engineering, trustworthy (autonomous) systems need to integrate the three "ingredients of trust"—abilities, benevolence and integrity. Although there is some overlap between these ingredients and the technical understanding of these terms in the engineering disciplines, it is not enough to consider only the latter. It was the aim of this article to provide evidence that these pillars alone are not sufficient because the system has to prove its factual trustworthiness in various specific situations. To do this, explainability is needed.

Or conversely, a system can only be trustworthy if

- it has the requisite abilities *and*
- it has the will (motive) to be trustworthy (benevolence) *and*
- it behaves according to standards and norms (integrity) *and*
- it can relate its decisions to these properties in specific situations in the form of an explanation that is understandable and satisfying to humans.
- Additionally, it may even be able to revise decisions and consider potential alternatives in accordance with the above properties in dialogue with the human user (responsiveness).

So far, systems engineering can only provide such systems for special use cases that can be (more or less comprehensively) defined in advance. However, in the case of autonomous systems—systems that will be allowed to define their actions in several complex but more or less undefined situations—this is not yet possible. Current approaches (such as simulation-based verification and validation) try to overcome this situation by exploring the entire parameter space and thus seeking to cover all possible situations in advance. This is where the idea

of a more intensive interdisciplinary approach comes into play. The humanities and social sciences can help engineering science to identify key elements of behavior evaluation so that such models can be transferred to systems engineering and combined with classical assurance approaches to create trustworthy systems more effectively and efficiently. Initial steps have already been taken to structure and manage such interdisciplinary development processes.[85] In return, the social sciences and humanities can deepen their understanding of autonomous systems and develop new theoretical and empirical models to grasp their specifics.

Unfortunately, the overview presented of the different interdisciplinary approaches to trust and trustworthiness is far from being complete. For example, it is not only the passengers' perspective that is relevant but also that of the auditors at a certification authority who give safety approval to newly developed autonomous vehicles. Building trust is essential, as formal safety verification and validation for autonomous vehicles with regard to unknown issues would seem to be impossible. Moreover, there are various other disciplines that would no doubt make a valuable contribution to the topic of this article. In addition to economics, a look at sustainability studies, law and religious studies, to name just a few, would be necessary to get a more complete overview of all facets of the phenomenon of trust and trustworthiness. Through this article, we hope nevertheless to have worked out three essential ingredients for trust and to have shown that explainability and responsiveness play a central role as the unifying catalysts that ultimately help to establish trust towards autonomous systems.

ACKNOWLEDGEMENTS

Open access funding enabled and organized by Projekt DEAL.

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

ORCID

Imke Hoppe https://orcid.org/0000-0002-2035-7715
Willem Hagemann https://orcid.org/0000-0002-8259-6084
Ingo Stierand https://orcid.org/0009-0000-7936-6969
Axel Hahn https://orcid.org/0000-0003-2240-5351

ENDNOTE

[a] Nine hours correspond to 12 driving lessons of 45 min each; see ref. [1, §5. Para. 3 Appendix 4].

[b] Autonomy often connotes AI. In the following, the term *autonomous vehicle* is understood as *fully automated vehicle* according to the SAE J3016 standard, which is to say, a vehicle that masters all driving tasks like a human driver. AI shall refer to a "machine to perform tasks commonly associated with intelligent beings" (Encyclopedia Britannica). In this sense, an autonomous vehicle may be considered an AI, independently of the particular technologies incorporated.

[3] See https://www.moralmachine.net/

[4] See, for example, the statutes of the Association for Digital Humanities in the German Speaking Areas (DHd e.V.), available at https://dig-hum.de/dhd-satzung.

REFERENCES

1. Bundesministerium der Justiz. Fahrschüler-Ausbildungsordnung vom 19. Juni 2012 (BGBl. I S. 1318), die zuletzt durch Artikel 2 der Verordnung vom 18. März 2022 (BGBl. I S. 498) geändert worden ist: FahrschAusbO. 2012. Accessed July 3, 2023. [Online]. https://www.gesetze-im-internet.de/fahrschausbo_2012/index.html
2. Bolles A, Hahn A, Hoppe I. *Trustworthy Autonomous Systems: Systems Engineering Challenges for Trustworthiness*. DLR; 2023. Accessed June 20, 2023. [Online]. https://a.storyblok.com/f/74249/x/a5f4eb6868/s3-1-2022-07-09_bolles_hoppe_hahn_trustworthy_autonomous_systems_final.pdf
3. Mara M, Meyer K. Acceptance of autonomous vehicles: an overview of user-specific, car-specific and contextual determinants. In: Riener A, Jeon M, Alvarez I, eds. *Springer eBook Collection*. 1st ed. Springer International Publishing, Imprint Springer; 2022:51-83. User Experience Design in the Era of Automated Driving.
4. Fleischer T, Schippl J, Puhe M. Autonomes Fahren und soziale Akzeptanz: konzeptionelle Überlegungen und empirische Einsichten. *J Mob Verk*. 2022(12):9-23. https://journals.qucosa.de/jmv/article/download/80/69. [Online].
5. Kröller T, Schwarz J. Kovarianzstrukturanalyse: analyse der Einflussfaktoren auf die Akzeptanz autonomer Automobile in Deutschland. In: Boßow-Thies S, Krol B, eds. *Lehrbuch, Quantitative Forschung in Masterarbeiten: Best-Practice-Beispiele wirtschaftswissenschaftlicher Studienrichtungen*. Springer Gabler; 2022:379-417.
6. Lenz B, Fraedrich E. Neue Mobilitätskonzepte und autonomes Fahren: potenziale der Veränderung. In: Maurer M, Gerdes JC. Lenz B, Winner H, eds. *Springer Open, Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte*. Springer; 2015:175-195.
7. Nastjuk I, Herrenkind B, Marrone M, Brendel AB, olbe LM. What drives the acceptance of autonomous driving? An investigation of acceptance factors from an end-user's perspective. *Technol Forecast Soc Change*. 2020;161:120319. doi:10.1016/j.techfore.2020.120319
8. TÜV Verband. TÜV Mobility Studie 2022. 2022. Accessed June 20, 2023. [Online]. https://www.tuev-verband.de/?tx_epxelo_file[id]=878103
9. Gross A. Consumer Skepticism Toward Autonomous Driving Features Justified. 2022. Accessed June 20, 2023. [Online]. https://newsroom.aaa.com/2022/05/consumer-skepticism-toward-active-driving-features-justified/
10. European Commission and Directorate-General for Communications Networks, Content and Technology. Ethics guidelines for trustworthy AI, Publications Office, 2019. Accessed June 20, 2023. [Online]. https://data.europa.eu/doi/10.2759/346720
11. Long D, Magerko B, What is AI literacy? Competencies and design considerations. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu HI USA. 2020:1-16.
12. Henschke A. Trust and resilient autonomous driving systems. *Ethics Inf Technol*. 2020;22(1):81-92. doi:10.1007/s10676-019-09517-y
13. Hoff KA, Bashir M. Trust in automation: integrating empirical evidence on factors that influence trust. *Hum Factors*. 2015;57(3):407-434. doi:10.1177/0018720814547570
14. Thamrin MH, Bafadhal OM, Santoso AD. What promotes cognitive dissonance among anti-vaccine members in Indonesia? *IJPHS*. 2023;12(1):203. doi:10.11591/ijphs.v12i1.22125
15. Seefeld L, et al. Einstellungen, Wissen und Verhalten von Erwachsenen und Eltern gegenüber Impfungen—Ergebnisse der Repräsentativbefragung 2021 zum Infektionsschutz. 2022. doi:10.17623/BZgA:T2-IFSS-2021
16. Fieselmann J, Annac K, Erdsiek F, Yilmaz-Aslan Y, Brzoska P. What are the reasons for refusing a COVID-19 vaccine? A qualitative analysis of social media in Germany. *BMC Public Health*. 2022;22(1):846. doi:10.1186/s12889-022-13265-y

17. Haug S, Schnell R, Scharf A, Altenbuchner A, Weber K. Bereitschaft zur Impfung mit einem COVID-19-Vakzin—Risikoeinschätzung, Impferfahrungen und Einstellung zu Behandlungsverfahren. *Präv Gesundheitsf*. 2022;17(4):537-544. doi:10.1007/s11553-021-00908-y
18. Holthausen BE, Wintersberger P, Walker BN, Riener A, Situational Trust Scale for Automated Driving (STS-AD): development and initial validation. In: 12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Virtual Event DC USA. 2020:40-47.
19. Taddicken M, Reif A, Brandhorst J, Schuster J, Diestelhorst M, Hauk L. Wirtschaftlicher Nutzen statt gesellschaftlicher Debatte? Eine quantitative Framing-Analyse der Medienberichterstattung zum autonomen Fahren. *Medien Kommun*. 2020;68(4):406-427. doi:10.5771/1615-634X-2020-4-406
20. Maure Mr, Gerdes JC, Lenz B, Winner H, eds. *Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte*. Springer: 2015.
21. Koopman P, Wagner M. Autonomous vehicle safety: an interdisciplinary challenge. *IEEE Intell Transport Syst Mag*. 2017;9(1):90-96. doi:10.1109/MITS.2016.2583491
22. Chen H, Wen Y, Zhu M, et al. From automation system to autonomous system: an architecture perspective. *J Mar Sci Eng*. 2021;9(6):645. doi:10.3390/jmse9060645
23. Bundesministerium für Digitales und Verkehr. *BMVI—Gesetz zum autonomen Fahren tritt in Kraft*. [Online]. Accessed June 20, 2023. https://www.bmvi.de/SharedDocs/DE/Artikel/DG/gesetz-zum-autonomen-fahren.html
24. Birkemeyer L, Delventhal M, Schaefer I, Schmieder F. *Wann fahren wir autonom?. Eine Untersuchung aus technischer und rechtlicher Sicht*. Gesellschaft für Informatik e.V; 2022.
25. Nordhoff S, Kyriakidis M, van Arem B, Happee R. A multi-level model on automated vehicle acceptance (MAVA): a review-based study. *Theor Issues Ergon Sci*. 2019;20(6):682-710. doi:10.1080/1463922X.2019.1621406
26. Shneiderman B. Human-centered artificial intelligence: reliable, safe & trustworthy. *Int J Hum-Comput Int*. 2020;36(6):495-504. doi:10.1080/10447318.2020.1741118
27. Schulz-Schaeffer I. *Technik als Gegenstand der Soziologie. Berlin*. Technische Universität Berlin, Fak. VI Planen, Bauen, Umwelt, Institut für Soziologie Fachgebiet Techniksoziologie; 2008. Accessed June 20, 2023. [Online]. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-12318
28. Kohn SC, de Visser EJ, Wiese E, Lee Y-C, Shaw TH. Measurement of trust in automation: a narrative review and reference guide. *Front Psychol*. 2021;12:604977. doi:10.3389/fpsyg.2021.604977
29. Holthausen BE, Stuck RE, Walker BN. Trust in automated vehicles. In: Riener A Jeon M, Alvarez I, eds. *Springer eBook Collection*. 1st ed. Springer International Publishing; 2022:29-49.
30. Blöbaum B. Key factors in the process of trust. on the analysis of trust under digital conditions. In: Blöbaum B, ed. *Progress in IS, Trust and Communication in a Digitized World*. Springer International Publishing; 2016:3-25.
31. Wissenschaftsrat. *Perspektiven der Informatik in Deutschland. Positionspapier*. Wissenschaftsrat; 2020. Accessed June 20, 2023. [Online]. https://www.wissenschaftsrat.de/download/2020/8675-20.html
32. Wissenschaftsrat. *Wissenschaft im Spannungsfeld von Disziplinarität und Interdisziplinarität; "Positionspapier," Drs. 8694-20*. Wissenschaftsrat; 2020. Accessed September 13, 2021. [Online]. https://www.wissenschaftsrat.de/download/2020/8694-20.html
33. Damm W, Kalmar R. Autonome Systeme: Fähigkeiten und Anforderungen. *Inform-Spektrum*. 2017;40(5):400-408. doi:10.1007/s00287-017-1063-0
34. Liggesmeyer P. Autonome Systeme. *Inform-Spektrum*. 2017;40(5):399. doi:10.1007/s00287-017-1046-1
35. Schaefer KE, Chen JYC, Szalma JL, Hancock A. A meta-analysis of factors influencing the development of trust in automation: implications for understanding autonomy in future systems. *Hum Factors*. 2016;58(3):377-400. doi:10.1177/0018720816634228
36. Putzer H, Rueß H, Koch J. Trustworthy AI-based Systems with VDE-AR-E 2842–61: Structured development for trustworthy autonomous/cognitive systems. 2021. Accessed June 20, 2023. [Online]. http://www.cogitron.de/assets/files/ewC2021-ID10334docxTrustworthyAI-basedSystemsWithVDE-AR-E2842-612021-03-03.pdf
37. *VDE-Anwendungsregel VDE AR 2842-61-1 – Specification and design of autonomous systems*. 2021.
38. Avizienis A, Laprie J-C, Randell B, Landwehr C. Basic concepts and taxonomy of dependable and secure computing. *IEEE Trans. Dependable Secure Comput*. 2004;1(1):11-33. doi:10.1109/TDSC.2004.2
39. *ISO 26262:2018 Road Vehicles—Functional Safety*, International Organization for Standardization; 2011.
40. Anderson R. *Security Engineering: A Guide to Building Dependable Distributed Systems*. 3rd ed. John Wiley & Sons; 2020.
41. Pfleeger CP, Pfleeger SL. *Security in Computing*. 4th ed. Prentice Hall; 2007.
42. Dijkstra EW. Solution of a problem in concurrent programming control. *Commun ACM*. 1965;8(9):569.
43. Wensley JH, Lamport L, Goldberg J, et al. SIFT: design and analysis of a fault-tolerant computer for aircraft control. *Proc IEEE*. 1978;66(10):1240-1255. doi:10.1109/PROC.1978.11114
44. SafeTRANS e. V. Safety, Security, and Certifiability of Future Man-Machine Systems. 2021. Accessed June 20, 2023. [Online]. https://www.safetrans-de.org/de/Uploads/AK_2018_RLE_CPS/SafeTRANS_RM_SSC_FMMS_Roadmap_V2.pdf
45. IEEE Standard Glossary of Software Engineering Terminology. *IEEE Std 610.12-1990*. 1990;1-84. doi:10.1109/IEEESTD.1990.101064
46. Jackson S, Ferris TLJ. Resilience principles for engineered systems. *Syst Eng*. 2013;16(2):152-164. doi:10.1002/sys.21228
47. Shafique M, Naseer M, Theocharides T, et al. Robust machine learning systems: challenges, current trends, perspectives, and the road ahead. *IEEE Des Test*. 2020;37(2):30-57. doi:10.1109/MDAT.2020.2971217
48. Fey G, Fränzle M, Drechsler R. Self-explanation in systems of systems. In: 2022 IEEE 30th International Requirements Engineering Conference Workshops (REW). IEEE; 2022:85-91.
49. Miller T. Explanation in artificial intelligence: insights from the social sciences. *Artif Intell*. 2019;267:1-38. doi:10.1016/j.artint.2018.07.007
50. Koopman P. How safe is safe enough?. *Measuring and Predicting Autonomous Vehicle Safety*. Carnegie Mellon University; 2022.
51. Ethik-Kommission der Bundesregierung. Automatisiertes und Vernetztes Fahren. Bundesregierung; 2017. Accessed June 20, 2023. [Online]. https://www.bundesregierung.de/breg-de/service/publikationen/bericht-der-ethik-kommission-729110
52. Bundesministerium der Justiz. *Verordnung zur Genehmigung und zum Betrieb von Kraftfahrzeugen mit autonomer Fahrfunktion in festgelegten Betriebsbereichen*. 2022. Accessed June 20, 2023. [Online]. https://www.gesetze-im-internet.de/afgbv/index.html
53. Goodall NJ. Can you program ethics into a self-driving car? *IEEE Spectr*. 2016;53(6):28-58. doi:10.1109/MSPEC.2016.7473149
54. *LIT Robopsychology Lab, JKU—Johannes Kepler Universität Linz*. Accessed June 20, 2023. [Online]. https://www.jku.at/en/lit-robopsychology-lab/
55. Malsch T. Sozionik. *Informatiklexikon*. Gesellschaft für Informatik e.V; 2005. Accessed June 8, 2023. [Online]. https://gi.de/informatiklexikon/sozionik
56. Hepp A. *Deep Mediatization*. Routledge; 2020.
57. Rammert W. Wie die Soziologie zur ‚Künstlichen Intelligenz' kam: eine kurze Geschichte ihrer Beziehung. In: Muhle F, ed. Sozialwissenschaftliche Einführungen, *Band 4*, Soziale Robotik: Eine sozialwissenschaftliche Einführung. De Gruyter Oldenbourg; 2023:31-66.
58. Schraube E, Chimirri NA. Qualitative Technikpsychologie. In: Mey G, Mruck K, eds. Handbuch Qualitative Forschung in der Psychologie:

Band 1: Ansätze und Anwendungsfelder. Springer Fachmedien Wiesbaden; 2020:507-522.
59. Werthner H. *Perspectives on Digital Humanism*. Springer International Publishing; 2022.
60. Cools H, van Gorp B, Opgenhaffen M. Where exactly between utopia and dystopia? A framing analysis of AI and automation in US newspapers. *Journalism*. 2022;25:3-21. doi:10.1177/14648849221122647
61. Coeckelbergh M. *AI Ethics*. The MIT Press; 2020.
62. Misselhorn C. *Grundfragen der Maschinenethik*. 5th ed. Reclam Verlag; Reclam; 2022.
63. Baier A. Trust and antitrust. *Ethics*. 1986;96(2):231-260. http://www.jstor.org/stable/2381376 [Online]
64. McLeod C. Trust. In: Zalta EN, ed. *The Stanford Encyclopedia of Philosophy*. 2021st ed. Metaphysics Research Lab, Stanford University; 2021. Accessed June 20, 2023. [Online]. https://plato.stanford.edu/archives/fall2021/entries/trust/
65. Budnik C. Vertrauen als Gegenstand der Philosophie: Vertrauen ist ein Phänomen, das uns im Alltag auf Schritt und Tritt begegnet. 2016. Accessed June 20, 2023. [Online]. https://www.philosophie.ch/2016-12-20-budnik
66. Coeckelbergh M. Can we trust robots? *Ethics Inf Technol*. 2012;14(1):53-60. doi:10.1007/s10676-011-9279-1
67. Jones K. Trustworthiness. *Ethics*. 2012;123(1):61-85. doi:10.1086/667838
68. Häuslschmid R, von Bülow M, Pfleging B, Butz A. Supporting trust in autonomous driving. In: Proceedings of the 22nd International Conference on Intelligent User Interfaces, Limassol Cyprus. Association for Computing Machinery; 2017:319-329.
69. Mayer RC, Davis JH, Schoorman FD. An integrative model of organizational trust. *Acad Manage Rev*. 1995;20(3):709-734.
70. Wu Z, Zhou H, Xi H, Wu N. Analysing public acceptance of autonomous buses based on an extended TAM model. *IET Intelligent Trans Sys*. 2021;15(10):1318-1330. doi:10.1049/itr2.12100
71. *Functional Safety of Electrical, Electronic, programmable electronic safety related systems: International Standard IEC 61508-1*. International Electrotechnical Commission.
72. *ISO 17894:2005: Ships and marine technology — Computer applications — General principles for the development and use of programmable electronic systems in marine applications*. ISO/TC 8/SC 8 Ship design. 2005. [Online]. https://www.iso.org/standard/31619.html
73. Kramer B, Neurohr C, Büker M, Böde E, Fränzle M, Damm W. Identification and quantification of hazardous scenarios for automated driving. *Model-Based Safety and Assessment*. Springer; 2020:163-178.
74. Detjen H, Pfleging B, Schneegass S. A wizard of oz field study to understand non-driving-related activities, trust, and acceptance of automated vehicles. In: 12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications. 2020:19-29.
75. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*. 1st ed. 2019. Accessed June 20, 2023. [Online]. https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html
76. Huang X, Kwiatkowska M, Wang S, Wu M. Safety verification of deep neural networks. In: Majumdar R, Kunčak V, eds. *Lecture Notes in Computer Science, Computer Aided Verification*. Springer International Publishing; 2017:3-29.
77. Gunning D, Aha D. DARPA's Explainable Artificial Intelligence (XAI) program. *AIMag*. 2019;40(2):44-58. doi:10.1609/aimag.v40i2.2850
78. Lawless WF, Mittu R, Sofge DA, Shortell T, McDermott TA, eds. *Systems Engineering and Artificial Intelligence*. Springer International Publishing; 2021.
79. Woodward J, Ross L. Scientific explanation. In: Zalta EN, ed. *The Stanford Encyclopedia of Philosophy*. 2021st ed. Metaphysics Research Lab, Stanford University; 2021. Accessed June 20, 2023. [Online]. https://plato.stanford.edu/archives/sum2021/entries/scientific-explanation/
80. Hempel CG, Oppenheim P. Studies in the logic of explanation. *Philos Sci*. 1948;15(2):135-175. doi:10.1086/286983
81. Salmon WC. *Four Decades of Scientific Explanation*. University of Pittsburgh Press; 2006.
82. Schweder R. A defense of a unificationist theory of explanation. *Found Sci*. 2005;10(4):421-435. doi:10.1007/s10699-004-5250-5
83. Köhl MA, Baum K, Langer M, Oster D, Speith T, Bohlender D, Explainability as a Non-Functional Requirement. In: 2019 IEEE 27th International Requirements Engineering Conference (RE). IEEE; 2019:363-368.
84. Luhmann N. *Vertrauen: Ein Mechanismus der Reduktion sozialer Komplexität*. UVK Verlagsgesellschaft mbH; 2014.
85. Ramli MR, Törngren M. Towards an architectural framework and method for realizing trustworthy complex cyber-physical systems. In: 16th International Conference on Research Challenges in Information Science (RCIS 2022). CEUR Workshop Proceedings (CEUR-WS.org); 2022.

How to cite this article: Hoppe I, Hagemann W, Stierand I, Hahn A, Bolles A. Challenges for trustworthy autonomous vehicles: Let us learn from life. *Systems Engineering*. 2024;1-12. https://doi.org/10.1002/sys.21744