Trust your MUM: Trust as a Pillar of User Acceptance for the Autonomous Modifiable Underwater Mothership (MUM)

Trust your MUM: Trust as a Pillar of User Acceptance for the Autonomous Modifiable Underwater Mothership (MUM)

A Schmitz and H Braun

Resilience of maritime socio-technical Systems, Institute for the Protection of Maritime Infrastructures, German Aerospace Center (DLR), Bremerhaven, Germany

E-mail: alena.schmitz@dlr.de, hagen.braun@dlr.de

Abstract. Within the Large Modifiable Underwater Mothership (MUM) project, a consortium of research institutions and industry partners is developing a modular, fuel cell-powered underwater vehicle capable of extended autonomous operations. Its unique design aims to support missions in remote, high-risk, or high-latency environments—such as unexplored regions beneath Arctic ice—where human presence is limited or unfeasible. However, enabling such autonomy requires a transfer of control from human operators to the system itself, making the operator's trust in the system a critical prerequisite for acceptance.

This paper explores the complex relationship between trust and control in autonomous maritime systems through a mixed-method approach combining qualitative exploration and conceptual analysis. Findings highlight three interrelated dimensions: first, *trustworthiness* emerged as a key operator expectation, grounded in both technical performance and the credibility of human actors behind the system. Second, the *rationale for autonomy versus human control* revealed persistent ambivalence—participants acknowledged the benefits of autonomy but also expressed a strong desire to maintain human oversight, particularly in ethically or operationally uncertain scenarios. Third, the interplay between *human-machine coagency* and *perceived control* proved central to trust formation. Participants were more willing to delegate control when they retained a sense of personal agency, even without actual intervention capacity.

1. Introduction

Within the framework of the Large Modifiable Underwater Mothership (MUM) project, research institutions and industry partners are jointly developing a next-generation underwater vehicle. This vessel is distinguished by its modular design, fuel cell propulsion, and its capacity for largely autonomous operation. Its unique characteristics are intended to enable the vehicle to perform a wide range of tasks and missions over extended periods of time, even in hard-to-reach, high-risk, or high-latency environments. In particular, its autonomous capabilities could offer advantages that may revolutionise operations in various fields of application. For instance, researchers would gain access to regions beneath Arctic ice that have so far remained unexplored, enabling the conduct of measurements, the localisation of hydrothermal vents, and the documentation of biological communities within these ecosystems. Nevertheless, to realise these benefits, especially in environments where communication is limited and manual intervention is impossible, control must shift from human operators to the system itself. This transfer of control, however, presupposes a high degree of trust, particularly in light of the unpredictable and high-stakes nature of the operational context. Within this paper, a mixed-method approach has been used to explore the complex trust-control relation. The study takes an exploratory and conceptual perspective to identify the characteristics that potential operators expect from the MUM in order to consider it trustworthy. Furthermore, it investigates the rationale operators provide for wanting autonomous operation or human control, and examines the interplay between human-machine co-agency and perceived control. The findings aim to contribute to the broader question*:* How can trust be systematically established in autonomous maritime operations?

2. Methodology

To address the complexities of this subject, a mixed-method approach was adopted, combining an explorative qualitative study with conceptual analysis. The empirical foundation of this study consists of a set of semi-structured interviews exploring various aspects of operator acceptance. This qualitative approach uses a small, non-representative sample. The findings cannot be generalised to a broader population, but help to gain a deeper understanding of patterns, experiences and complex phenomena. Drawing from existing technology acceptance models, particularly those focusing on automation, key criteria such as trust and locus of control were identified. These informed the conceptual basis for the development of the interview guide.

The interviews were conducted between November 2022 and March 2023 with 21 potential operators of the MUM system. Unlike conventional vessels, MUM will neither be directly controlled by operators nor operated from onboard. Instead, operators may be involved in programming, monitoring, or collaborating with the system. Possible examples of operators could include marine biologists using the vehicle for sample collection, divers working alongside it on underwater infrastructure, remote observers in control centres, software developers programming mission plans, or maritime pilots guiding the vehicle through restricted areas. Given the diversity of potential operator roles, the interviews aimed to capture a wide range of perspectives across various use cases. Participants came from six occupational fields: Research and Development, Ship and Traffic Safety, Marine Research, Offshore Energy, and Teaching and Simulation.

The utilized interview guide is comprised of open-ended questions designed to gain insights into the experiences and perspectives of the interview participants. To ensure a structured approach, the guide was divided into thematic blocks aligned with potential acceptance criteria, including Perceived Usefulness, Perceived Ease of Use, Task-Technology Compatibility, Trust, Perceived Safety, Locus of Control, and Ethical Concerns. The interview guide itself is too expansive to reproduce in its entirety in this paper, but an example of a question from the guide is:

> *"If you had to name three aspects that are fundamental to your trust in a technology, what would they be? For a technology in your field of work? For MUM in particular?"*

The collected data was analysed using the systematic approach proposed by Gioia et al. (2013), which structures findings across three analytical levels: First-Order Concepts – direct quotes and terminology from interviewees, Second-Order Themes – researcher interpretations and categorisations, and Aggregated Dimensions - theoretical abstractions. Quotes from German-language interviews were translated into English. The Gioia approach was chosen for its strength in translating empirical material into theoretical insight while maintaining close adherence to the participants' original expressions. In a final step, the empirical findings presented in Chapter 3 were compared and discussed in light of current concepts from ethical discourse on autonomous systems presented in Chapter 4, highlighting similarities and differences between the two.

3. Findings

This paper presents a part of a user acceptance study for the MUM system, focusing specifically on the trust-control complex. Within the study, a preliminary adapted technology acceptance model for MUM (Figure 1) was developed, which—beyond the original acceptance criteria centred on perceived usefulness and ease of use—also highlights trust as a core pillar of operator acceptance. Where trust is maintained in a complex balance with control.

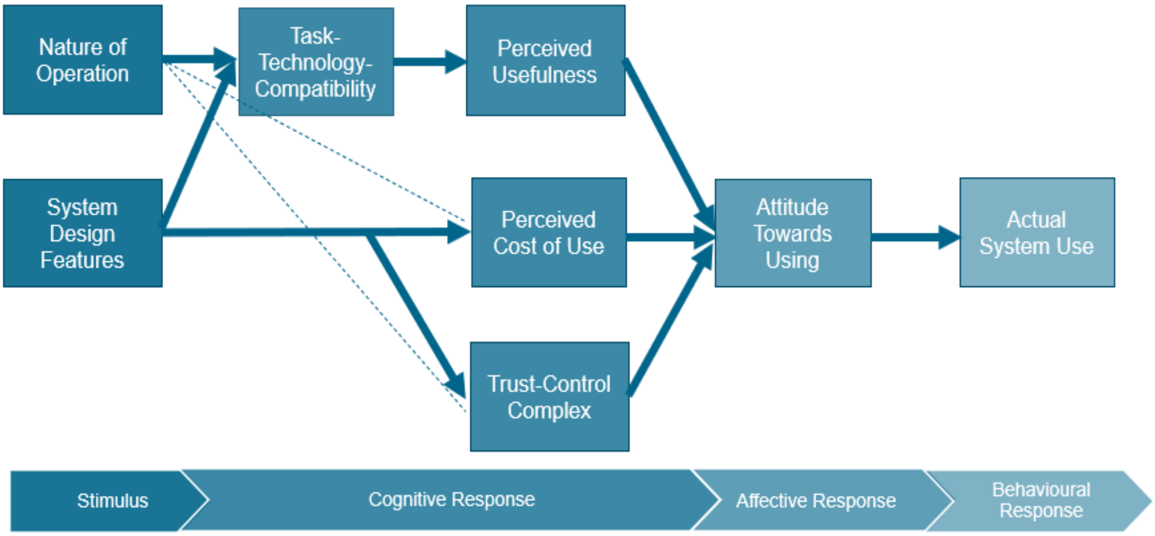

Figure 1. Preliminary Adaption of the Existing Technology Acceptance Model

Based on the interview data, three central dimensions have been identified within the trust–control complex (Figure 2). The first dimension**,** *trustworthiness*, addresses the expectations operators have regarding the system properties of MUM—such as *explainability*, *transparency*, and *redundancy*—which are considered essential for justifying laying trust in the system. The second aspect, *rationales for autonomy and human control*, captures the reasoning and conditions under which operators perceive autonomous functionalities as beneficial or problematic. Among others it reflects the operational and ethical considerations behind delegating control to the machine versus retaining it with humans. The third theme, *human–machine coagency*, explores how operators imagine the collaboration between humans and the autonomous system. This includes the individual perceptions of control in different coagency models. Together, these themes provide insight into the multidimensional relationship between trust, control, and acceptance in the context of autonomous maritime systems.

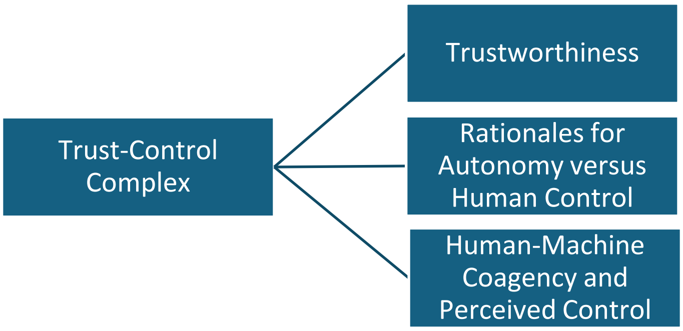
Trust-Control Complex
Trustworthiness
Rationales for Autonomy versus Human Control
Human-Machine Coagency and Perceived Control

Figure 2. Dimensions of trust within the MUM project

3.1 Trustworthiness

Within this dimension the interviewees described system characteristics, that need to be in place to rightfully trust the system. The participants do not limit their requirements solely to the MUM system as a technical artefact, but extend them to the human factor, such as the supplier, the service provider and the potential operator. The data suggests that the dimensions of trustworthiness are conceptually intertwined and not strictly separable. Two closely related concepts that were repeatedly emphasised by the participants are *transparency* and *explainability*. *Transparency* reflects the desire for a general comprehensibility of the system. Interviewee *V4679* derives from this a responsibility on part of the operator, who should know the system and its behaviour in certain situations entirely. *Transparency* was further frequently mentioned as imperative in the context of emergency situations. Interview participants emphasised the need for a clearly defined emergency protocol—not only to ensure that the system can respond appropriately to various scenarios, but also to guarantee that operators are trained in how the system behaves and what procedures to follow in the event of an incident (*P7781*). This includes the expectation that a wide range of potential situations has been thoroughly considered in advance. In addition, the interviewees stressed the importance of emergency-related *transparency* towards other traffic participants. These should be adequately warned (*M4571*): The MUM vessel must remain in a state that is predictable and assessable for others in the vicinity. *Explainability*, by contrast, refers to particular decisions. To be able to assume the responsibility as an operator, participant *L1539* argues, the operator must understand how the system arrives at the specific decision. „*Difficulties could arise if the vehicle makes decisions that cannot be tracked*" (W6658). The participants have discussed these concepts and their realisation with different levels of details. One intrviewee envisions an active *transparency* of the system, meaning that it provides an overview over its mission planning and the relevant data underlying its assessment of the situation:

> *"I am the MUM, this is my plan of today, these are the weather conditions. I think I can make it because of this and this and this. The first I want to this, after that we're going to do that. La la la la. I don't expect problems, but la la la la. We have a human look over my shoulder when I'm doing this and this and this, because we will look up this and this." (C7212)*

What the interview partners agree on is that the system must be easy to understand; if the operation becomes too complex and incomprehensible *transparency* and *explainability* are not given.

Redundancy was consistently referenced as a key requirement for the system. Participants described redundancy both as structural and functional. They referred to individual subsystems they considered essential—such as propulsion (*V4679*), sensor systems *(L1539)*, safety mechanisms *(Y3558)*, navigation (*V4679*), and tracking (*Y3558*), they believe to be necessary in duplicate or even triplicate. Since no personnel are on board to maintain or repair the vessel, any system failure would require external recovery (*S5968*). Therefore, several interviewees emphasised the importance of the system's overall ability to complete its mission and return to a safe state, even in the event of individual component failures (*G2405*). The system has to remain safe not only for itself but also for other traffic participants (V4679).

Table 1. Characteristics of trustworthiness

Aggregated Dimension	Second-Order Themes	First-Order-Concepts
Trustworthiness	Explainability	*"The traceability. What if I, as a human being, should still be in charge of this at all? So, if I'm supposed to give my opinion at all, then I should be able to understand how the system arrives at a certain decision, that the decision-making process." (L1539)*
	Transparency	*"The attentions of the machinery should be shown at all times to the human." (C7212)* *"The future operator should actually know this ship, let's call it by heart, [...] know all the modules [...] He really has to know it inside out [...]. He has to know the weak points or the limits of the system. And he must be able to assess the situation at all times." (V4679)*
	Redundancy	*"Even in the event of an error, a system must not immediately abort everything [...] it must be able to somehow maintain the system in a safe state." (G2405)* *"If it is to fail, you have a backup. If the backup fails, you can work around and still be able to do what you need to do." (C7212)* *"I think redundancies in a system are obligatory, redundancies in the sense of drive, navigation and not necessarily in the sense of function. I think it's important for these vehicles to always be secured in some way. In other words, safety for themselves, but also for other road users and other structures on land." (V4679)*

	Reliability	*“Of course it has to be reliable enough that in those moments when it's really important,[...]that when we're travelling straight out somewhere on the open sea, it's all feasible, but the moment another ship or other weather conditions suddenly occur, a change in current, it has to work 100%, because if it's too late, I'm saying that if there's a two-second delay or something like that, especially in our canal, then that's enough to make it impossible to avoid and possibly cause a collision.” (M4571)* *“Reliability is a very important aspect. So, how reliable. How well can I rely on the technology? In other words, how often does it make mistakes? If it makes a mistake once a week, then that shatters trust.” (W6658)* *“The manufacturer has, let's say, a good reputation.” (Y3558)* *“The team that designed the team MUM is also the team that we can call 24/7 for service. If you have a product and the group says, use your product, don't talk to us anymore. It's not as good. [...] You need constant support or you need to be able to have constant.” (C7212)*
	Competence	*“You will certainly start somewhere and then also test that you let a large number of future operands try it out and ask for feedback from them and then make improvements.” (W6658)* *“This is very clear proof that certain scenarios have been processed frequently and perhaps randomly and have been passed by the system.”* (T5656) *“I would assemble the MUM with technology that's already proven in another field.”* (C7212)

Within the dimension of *trustworthiness reliability* serves as a complementary concept to *redundancy*. Interview participants associate *reliability* to a variety of different aspects. From a functional perspective, they emphasise the technology’s *reliability* is fundamental for trust. Given the absence of onboard personnel, the system must operate with a high degree of consistency and faultlessness under extreme marine conditions (*M4571)*. Individual components need to fulfill their respective function consistently, as malfunctions can quickly erode the operators trust in the MUM (*W6658)*. The sensors, in particular, must operate flawlessly and with high precision (*L1539*). As there is a strong reliance on their input, the accuracy of sensor data is paramount (*B4910*). Participant *M5471* highlighted the importance of a stable signal, to ensure the transmission and reception of the sensor data. Moreover, interviewee *Y3558* accentuated the critical importance of dependable propulsion. Failure to achieve the required range, would disqualify the system. *T5656* concludes*: “If everything works, there can't be any negative consequences”.* Beyond technical performance, *reliability* also extends to adherence to applicable regulations. This is especially pertinent in interactions with other traffic participants. Both the operator and other parties must have confidence that the system consistently complies with international traffic rules and responds appropriately—such as yielding at the correct moment (*B4910*). Lastly, *reliability* as a characteristic of *trustworthiness* is also extended to the supplier or service provider of the MUM. Interviewee *C7212*, for example, recommends working with *"elephants in the business"*—established companies with global reach and components available worldwide, that have proven themselves reliable. A company's reputation is linked to the perception of its *reliability*. Furthermore, *reliability* refers to the ongoing relationship with the supplier or service provider. This includes the expectation that, even after system delivery, they will provide consistent support, maintain a cooperative relationship, and show interest in working collaboratively with customers to develop solutions (*A5007*).

The final requirement consistently emphasised by interview participants can be summarised under the term *competence*. To justifiably trust the MUM, participants expressed the need for some form of demonstration or rigorous testing of the vehicle and its individual components. These expectations ranged from live, firsthand demonstrations (*P7781*) to the use of system components with a positive track record through years of operation in other vehicles (*C7212*; *O7123*), and adoption of the MUM by competitors (*Y3358*). Evidence that the system can handle specific scenarios through randomised testing and even long-term studies (*S5968*) was also highlighted. These should be well documented (*A5007).* The system must have surpassed the status of a mere experimental prototype particularly regarding its decision-making logic (*O7123*). The goal is to ensure that the associated risks are assessable, manageable, and kept within acceptable limits (*O7123*). An iterative development process, including feedback from potential operators, was identified as helpful for the creation of trust (*W6658*).

3.2 Rationales for autonomy versus human control

Our interviews revealed a complex tension between the *rationale for autonomous operations* versus the *rationale for maintaining human control* over the vessel—a dilemma that many of the interviewed potential operators explicitly recognised within themselves. Five key arguments have been standing out debating this. On the one hand the participants could think of a broad range of advantages, that come with use of autonomy. One central point is the *operational advantage* that arises with it. The autonomous operation of the MUM has the potential to significantly improve working conditions in certain fields. If the MUM were to be operated, operators could perform their tasks e.g. on a offshore energy plant, from onshore workplaces, remaining close to their families (*R7321)*, and being exposed to far lower levels of safety risk than in the current high-risk environments that characterise their workplaces (*Y3558*). Especially, arduous, dangerous, and monotonous tasks could likely be handled by the system, which would perform them consistently and at a steady quality (*L1539; B49109*). As a result, a more attractive working environment and overall professional profile could emerge (*L1539)*—a development that was viewed positively, particularly in light of the current shortage of skilled labour (*R7321*). Further outstanding *operational advantages* associated with the use of an autonomous vessel include the ability to access regions that lie beyond the operational reach of humans or remotely operated vehicles (ROVs) (*H6309*). This aspect was particularly highlighted by researchers, for whom access to areas beneath Arctic ice could represent a significant breakthrough in their respective fields (*H6309*). In addition, the system's efficiency in surveying and processing extensive areas was repeatedly emphasised as a key benefit (*P7781*). The category *technical limitations* encompasses both constraints that autonomy can help to bypass and those that, conversely, limit the feasibility of deploying AI-based systems. The system's autonomous decision-making capability enables its deployment in environments where data transmission reaches its technical limits. In such contexts, where operators lack the necessary information to make informed decisions, the autonomy of the vehicle allows the mission to proceed (*W6658).* The system thus offers a solution to the inherent limitations of data transfer under water and ice. However, several scenarios were also discussed in which the technical implementation of the MUM reaches its boundaries. In particular, the representation of complex aspects through pre-programmed decision logic was described as problematic (*O7123*). Decision-making based on parameters that cannot be clearly defined or measured poses a significant challenge for autonomous systems:

> *"Deciding what a person does intuitively or what a person simply does based on a gut feeling. What they don't find in the law is, in my view, the most complex thing that needs to be mapped in an autonomous vehicle." (R7321)*

This refers to situations that humans typically handle using experience, intuition, and communication (*V4679*) — for example, interactions with other traffic participants in accordance with good seamanship, or navigating dynamic conditions such as operating at the edge of a current (07123, *M4571*). Generally, the interviewees expressed scepticism about whether the autonomous system can handle situations that have not been explicitly trained into the AI (*G2405*). The maritime context, at the same time, involves so many variables that it presents an endless range of potential scenarios (*S5968*). It seems an unattainable goal to prepare the artificial intelligence to adequately behave in all of those. This issue extends into that category *situational complexities*. The participants point out several complexities that hinder the use of autonomous maritime systems—particularly in contexts involving mixed traffic with both human-operated and autonomous vessels (*I3111*). Emergency situations were highlighted as particularly challenging for a system like the MUM, as multiple aspects must be prioritised simultaneously (*Q8244).* The question of how an autonomous maritime vehicle handles the duty to rescue at sea has also been raised (*I3111*). In connection with the deployment of an autonomous maritime system, interviewees raised *ethical and legal concerns*. A recurring theme was the responsibility gap—the challenge of assigning accountability in autonomous operations. Interviewee *R7321* highlights the precarious situation for operators of autonomous systems who might be expected to take responsibility without having full control over decision-making. That can leave the operator with a feeling of being subject to the system's decision. *Social and cultural aspects* were also part of the discourse, either in reference to seafaring culture or national societal values. Participant *T5656* pointed out that, unlike in aviation, maritime tradition has never envisioned a system overriding a human decision. Interviewee *O7123*, on the other hand, suggests that the socialisation of a tech-savvy society could influence trust in an autonomous system. Lastly, an interviewee questions the general social desirability of comprehensive autonomy:

> *"Completely autonomous driving means that these ships are not even observed by them, but really act on their own. But the question is do we really want it that way?" (V4679)*

Table 2. Rationale for Autonomous Operations versus the for Maintaining of Human Control

Aggregated Dimension	Second-Order Themes	First-Order-Concepts
Rationales for Autonomy versus Human Control	Operational Advantage	*“I would say that many seafarers would rather stay with their families. So that's a clear advantage of autonomy. (...) Less chance of people getting hurt. It's still a dangerous working environment. Fewer people on board or none at all, for me that also means risk of injury is significantly lower or loss of human life is significantly lower.” (R7321)* *“Now, of course, boring work, which tends to be so boring that the human will make a mistake. And complex work where a human could need help in decision.” (C7212)* *“Thanks to the autonomy, it would be relatively efficient (...) to cover several interesting locations or even large areas or long distances. (...) Without the need for large numbers of personnel. And also much faster than usual.” (P7781)* *“We want to send the systems into a region or into areas where we can't go ourselves, where we can't get to with the ROV.” (H6309)* *“You could carry out wonderful research tasks that you can't do at the moment.” (A5007)* *“Of course, the communication options are always a bit difficult.” (O7123)*
	Technical Limitations	*“If the transmission is disturbed, you can still make decisions under water that would no longer be possible above water because you don't have the information.” (W6658)* *“You make decisions with it, then it just does it and drives, drives away. So of course operator on the loop is a really nice story, especially as I think there is also this ROV module or something like that, then of course you have to interact somehow, but I don't see any chance of that, especially under ice. I mean, in open water you can always raise a buoy and somehow have satellite communication or something. But under ice, I have no idea how that would work.” (A5007)* *“Is that what I would do in that situation or do I have to catch a gust of wind or something? Or a stall, a flow edge, if I'm travelling in a current cut like that, I can't manage that. I can't do that with automation technology either, because I can only ever act on what I can measure and I can't measure*

		a current cut. And that means how is automation supposed to work?" (L1539) *"Deciding what a person does intuitively or what a person simply does based on a gut feeling. Sometimes. What they don't find in the law is, in my view, the most complex thing that needs to be mapped in an autonomous vehicle." (V4679)* *"Firstly, the purely rational parameters, i.e. price, speed, everything that is fixed and firm. And everything else are soft criteria as soft skills. So the law says safety and ease of traffic and good seamanship good seamanship is not a term that can be defined by hard values, but rather something like "How would I feel if I were in their situation?" (V4679)*
	Situational Complexities	*"But if situations arise that have not been trained, then you cannot ensure that the technical system will not even fail. In other words, in this case, humans are much more resilient to external disruptions than a technical system. And this hurdle has to be overcome somehow." (G2405)* *"In a dodging situation? Human versus computer? There are clear rules and the computer will follow clear rules. Humans follow rules according to their decision. This may not always be according to the rules." (W6658)* *"They then reach their limits, for example in the canal, where I have a lot of interaction between the ship or between the land and the ship. You can't manage these effects with this control system because it doesn't anticipate and is always based on the past and then derives actions from it, which of course gets better and better from the control system with the filters." (L1539)* *"I'll just say a fire alarm in the vehicle. From my point of view, this is very complex, because then I have several things that I have to deal with at the same time." (Q8244)*
	Legal and Ethical Concerns	*"I don't believe that a human being or anyone is prepared to take responsibility without having complete control over a machine. That. I don't know how legally that can be reconciled, but. You are a human being; a human being bears the responsibility. Whoever is in charge should also have full decision-making power." (R7321)*
	Social and Cultural Aspects	*"I don't know of any system in shipping where the machine intervenes [...] that doesn't exist in seafaring. Even an autopilot or something like that won't change course, it will simply hold it. And if it knows a hundred times that there's a rock ahead or that the ship is somehow travelling, that's not planned." (T5656)*

	„I found this a very interesting aspect. So how much technology does someone grow up with in a society? How tech-savvy is a society or how much faith in technology? I think that also influences the extent to which people trust certain things." (O7123)

3.3 Human-Machine Coagency and Perceived Control

Human-machine coagency refers to the interplay between a human operator and an autonomous system in executing actions and making decisions. It describes how authority, task allocation, and control are distributed—and how they are exercised—in a joint human-machine operation. Our interview data reveal that potential users have very different ideas about how this coagency should be structured in the context of MUM. Their visions range from goal-based mission planning with high levels of autonomy (*A5007*), to expectations of continuous monitoring (*W6658*) and the ability to intervene at any time (*N0386*). These differing expectations are also reflected in assumptions about personnel requirements: some anticipate multi-person teams with distributed responsibilities (*W6658*), while others imagine a single operator overseeing multiple vehicles, possibly embedded in a larger support team (*R7321*). At the same time, the operational environment of MUM imposes specific technical constraints on human-machine collaboration. In high-latency areas—where MUM is expected to operate—permanent monitoring and direct remote control by human operators is not technically feasible (*W6658*). As described in Chapter 2.2, precisely these remote and hard-to-access areas represent key operational advantages of MUM. The operational and technical advantages offered by autonomy cannot be fully realised if comprehensive control — in the sense of continuous 24/7 ability to observe and intervene — is maintained. The aim for constant control contradicts the original idea and purpose of autonomy (*H6309*).

> *"If we don't want to have this autonomy at all, but want people to be able to control and monitor and intervene all the time, then we can go straight on with ROVs. There's a cable attached to it, that's the umbilical cord to the human being." (H6309)*

Some interviewees – despite being aware of this conflict – still advocate for extensive control and intervention capabilities within the system – sometimes referring to the risk of the mission (*Y3558*). It seems plausible that this stance is related to the specific tasks, missions, and operational contexts in which they expect to use the system (*V4679*). However, based on this limited dataset, there is no clear indication that members of the same professional group share a consistent understanding of *human-machine coagency*. Some interviewees argue that, the necessary technological maturity for autonomous operations is not yet achieved. Nevertheless, it might be reached in the medium or long term (*V4679, W6658*).

Table 3. Dimensions of Human-Machine Coagency and Perceived Control

Aggregated Dimension	Second-Order Themes	First-Order-Concepts
Human-Machine Coagency	Decision-Making-Authority	*"If I think about control mechanisms in general, I would expect the operator to be able to overwrite all possible programming in the final instance. [The operator](...) can override or take over control himself, even if it means destroying the system. I compare that with the ships currently in operation. For example, even if I have an engine alarm and it tells me that you have no more oil pressure and if you carry on like this, your engine will break down, I can still say: No, I'm going to carry on like this, I want to avoid grounding or a collision."* *(B4910)* *"[The system interfering with the human-decision making?] It shouldn't do that at all. As a human being, I should always have top priority. I can't imagine a situation like that. […] In principle, I think that humans should always retain the ultimate decision-making power, even in autonomous systems." (Y3558)* *"I think that when people intervene, they do so consciously and (...) If they do so consciously, then they should also have control over the ship. As soon as people intervene, the automation should hold back." (W6658)*
	Intervention and Monitoring Capability	*"That's what I meant with an emergency stop button. But that you can always say now, no matter how much autonomy you have 'stop now'." (Y3558)* *"I would argue in favour of letting the operator or the supervisor do [the communication], at least in the beginning. (...) Or at least that he listens in and then has the opportunity to intervene again." (W6658)* *"That clearly depends on the situation. If the vehicle is operating in an open area and cannot endanger anyone, let me put it this way, then there is no need for human intervention. In my opinion, the vehicle can then operate freely." (V4679)* *"I would also assume the same for a MUM vehicle, which I can still prevent or cancel or change a certain maneuver or operation." (B4910)* *"Human control is mandatory or at least the ability to take over when you think a human would be better at intervening than the robot, which in most cases is… I think the human needs to needs to have the mandate to take over at any moment. That would be mandatory within the most cost." (C7212)* *"Because I, as an operator or pilot, would always like to have the option to intervene. As I said, I think that's very important." (Y3558)*

		"[The person should] be able to intervene at any time if possible." (P7781) *“But I wouldn't trust the automation to the extent that I would have it done completely without a human supervisor.” (W6658)*
	Operation	*“The MUM, if there's any obstacles, would be able to kind of avoid those obstacles. When you come to the place where you want to do something. Then it could be a kind of what we call a supervised autonomy. [...] if you're talking about having systems that are more or less autonomous, which I think should be the mission of your project, then it should be kind of just approval or disapproval. Or just sending a short message to the mum that now you can do this or that, and then it should continue in an autonomous way. Because you will not you will not be able to pilot it or operate manipulators the communication links.” (K9980)* *“Mission planning that is not fully autonomous or human supervision of mission planning.” (P7781)* *“My expectation would be, at least initially, that before a mission or deployment is launched, a lot is done, planned and considered. That would also be what I expect, that you have a lot of planning work beforehand. Then you get to a point where [...] you just press a button and the mission starts.” (O7123)* *“Autonomisation is more applicable to standard elements. (...) Drive somewhere and fulfil it. But the moment something happens, it needs a human.” (W6658)*
Perceived Control	Decision support as cognitive framing	*“[I would say that [the system should only be able to intervene in the human decision-making process] in the sense of a warning. So the system should be able to warn. Be careful, if you do this and that, then the following will most likely happen. But in the sense that the system says no, I'm not going to let you do that any more. No, I can't really imagine that.” (B4910)* *“[The system can intervene in the human decision-making process”, maybe [in form of] a warning of a wrong decision, of a future wrong decision. But we already have something like that on our ships. So as soon as there are any close calls. Then there's an alarm.” (M4571)*
	Delegated control as retained agency	*“Humans still program them like we do determine, how the machine should decide.” (H6309)* *“Do my conscience decisions or moral decisions change at some point? So as long as I still have a person behind me, the situation doesn't change. “ (B4910)* *“At the end of the day, of course, I give up on our vehicle. If we have the vehicle in the water and it dives somewhere alongside and yes, and the decision is that we now drive away from the ship, and into another research area, which is somehow a few miles away. And I know exactly that I'll be out of range of communication with my vehicle, so I'll give*

	up this decision. So, I can no longer intervene because I'm simply no longer within communication range. But that's my decision. I then say yes, okay, we can do it." (Y3558)

The data on *human-machine coagency* is particularly interesting when viewed through the lens of *perceived control* – that is, the subjective sense of the potential operators regarding the extent to which they can influence the vehicle and its actions. Some respondents already interpret alerts and support systems as forms of intervention in the human decision-making process (*M4571*). Through the presentation of certain information, a cognitive framing (*L1539*) occurs that can limit the operator's range of choices — even though no formal transfer of control takes place. At the same time, interviewee *Y3558* describes a scenario in which he decides to deploy the MUM on a mission where direct monitoring or intervention in the vehicle's decisions is no longer possible. In this case, control is delegated up to a certain point. Nevertheless, he still perceives himself as being in control, as he took the initial decision to deploy the system.

4. Discussion

The concept of trustworthy technology has been gaining more attention recently, most prominently in the context of artificial intelligence (AI). Starting with the publication of the European Commission's Assessment List for Trustworthy AI (EU High Level Expert Group on Artificial Intelligence 2019), the idea of trust as a basis for AI ethics has become a point of interest for many authors (Ryan 2020; Simion and Kelp 2023; Al 2023; Alvarado 2023; Kaur et al. 2023). Although nominally focused on AI ethics, in reality the debate covers all technological systems with a strong component of autonomous functioning. While the idea of trustworthy AI has an intuitive appeal, it also quickly becomes clear upon further reflection that a more in-depth approach to AI trust is needed to fully realise the advantages promised by it. Several questions remain open: Why should trust be considered as the governing concept of a technology ethics and acceptance framework? What is the definition of trust and trustworthiness? And finally, how do trust and trustworthiness lead to increased acceptance in the context of autonomous maritime systems?

4.1 What kind of trust?

The literature on trust knows a variety of different types of notion with distinct areas of application, such as systemic (Hawley 2017), emotional (Jones 1996) and epistemic (Alvarado 2023) trust. However, the most basic and paradigmatic kind of trust is agential trust (Horsburgh 1961; Baier 1986; Holton 1994; Hieronymi 2008; Faulkner 2007), the relation of trust between two (human) agents. This is also the type of trust we will take to be the operative concept in the context of AI and autonomous maritime systems. Trust is a three-place relation between two agents and a task:

$$T(A_1, A_2, \Phi)$$

Equation 1. *Trust (relation)*

The first agent A_1 (the *trustor*) trusts the second agent A_2 (the *trustee*) with the task Φ. With regard to its agential components, this relation is not symmetric, reflexive or transitive.

One feature of AI (and autonomous systems more generally) that justifies the idea of agential trust as the applicable governing ethics concept is the action-like nature of the workings of the aforementioned types of technology. Historically, the aim of AI research has been to achieve the capacities of action and deliberation in computer systems (Bringsjord and Govindarajulu 2022; Russell and Norvig 2022). We find that today, this vision is coming ever closer to fruition, with novel machine learning techniques allowing computerised systems to perform tasks that were long held to be exclusively the domain of human agents. While it remains an open question to what extent AI agents truly "act", for the purposes of this paper it is sufficient to note that the effects produced by modern autonomous and AI systems are close enough to those produced by human action that similar ethics standards can be applied to them.

Many authors see trust as a precondition for efficient cooperation (Dimock 2020; Gambetta 1988; Luhmann 1979). Given the theme of human-machine coagency that emerged from our interview base, that is another good justification to consider trust as one of the relevant ethical governing principles in the context of autonomous systems. The idea that human operators making use of systems with significant degrees of autonomy is best understood as an exercise of joint human-machine agency has some backing in the literature, as well (Nyholm 2018).

4.2 Trustworthiness as an ethics standard

How does trust lead to better ethical outcomes and higher technology acceptance by stakeholders? Intuitively, a high level of trust in a technology could be seen as constituting, rather than causing or grounding, a high degree of acceptance: Being highly trusted *equates to* a way for a technology to be accepted. If trust and trustworthiness are to lead be a means to the end of increased technology acceptance, it is necessary to provide a mechanism by which this increase can be explained.

Therefore, it is important to note that there is a difference between *unfounded trust*, and *well-founded trust.* Whereas unfounded trust can be given without any rational justification, well-founded trust requires the trustor to track certain properties in the trustee that give sufficient reason for their trust (Wanderer and Townsend 2013; Hollis 1998). We will jointly refer to these properties as trustworthiness (Hardin 2002; Jones 2012; Kelp and Simion 2023). Using well-founded inter-agential trust as the basis for trustworthy AI has the upside of this conception of trust requiring the trustee (the autonomous system) to have certain properties that can be empirically verified, which allows for the potential operationalisation of trustworthiness requirements.

It is clear that increasing the amount of unfounded trust that stakeholders have for a technology is not a way to achieve a higher degree of acceptance or better ethical outcomes. Increasing trust in a technological system that is not actually trustworthy should be expected to lead to betrayals of that trust over time, which would have a negative effect on the acceptance of the technology in question in the long term. One possible definition of trustworthiness based on the literature is the following: *Trustworthiness* is a complex property of agents. It reduces to the properties of *reliability*, *competence* with regard to a certain task, and the possession of an *appropriate normative backing* to perform that task.

We define *reliability* as the disposition of an agent to perform a task with which it is entrusted. Reliability is frequently conflated with trustworthiness as a whole in the debate surrounding AI ethics (Nickel 2013; Taddeo 2009; Taddeo 2010). It is definitely true that reliability is a major component of trustworthiness: If we are to trust an agent or system with important tasks, they need to fulfil certain standards of reliability. However, given that trustworthiness is supposed to act as the governing standard of an ethics framework, this cannot be sufficient. It is very possible, after all, for an automated system to reliably perform its functions in an ethically questionable way. *Redundancy* and the presence of *emergency protocols* can be viewed as features of technological systems that enhance reliability, especially if reliability standards are expressed as a probability of accidents or failure with regards to the performance of a given task.

Competence is distinct from reliability, as it refers to the ability of the agent to perform the task in question. While reliability can be viewed as a general characteristic the agent possesses with regards to any task, competence is task-specific. A reliable, but incompetent agent will perform the task it is entrusted with, but not necessarily do it well. An unreliable, but competent agent might not perform a given task, but will do it well when it does. It is worth noting at this point that the meaning of reliability as it is used by the interviewees and the technical definition derived from the academic literature we have supplied above are not completely identical. Stakeholders explicitly equate a low rate of errors made by the system with high reliability, which according to the definitions given above would be better captured as a high degree of competence. As the interviewees explicitly mention the term *reliability* in this context, we have accordingly categorized this concern as such in our analysis of the interviews, even though it does not conform to our proposed theoretical framework of technology trust. This semantic difference is not a significant problem, however, as we believe our analysis of trustworthiness is still able to capture all of the concerns expressed by the stakeholders. Even though the exact property of the autonomous system that results in a low rate of errors differs from the common-sense, intuitive categorization of the interviewees, a low rate of errors nevertheless follows as one of the required characteristics of a trustworthy system.

The definitions we propose for these terms have the advantage of conforming to the academic literature on trust and offering a more nuanced analysis by clearly distinguishing two separate requirements for trustworthiness. For example, defining reliability and competence in the above way makes it possible to more clearly analyse an especially pertinent issue in the context of AI ethics around large language models: The problem of *hallucinations*. Hallucinations occur when a large language model outputs statements of fact that do not actually have any basis in reality. Hallucinations are an example of high reliability but low competence: The user always receives an output, but there is no guarantee that the output meets the necessary level of quality to be useful. In the context of an automated naval vessel like MUM, competence as part of trustworthiness could take the shape of the system providing operators with a measure of confidence in its own capabilities, and advising them on which tasks it can reasonably be relied on to perform (for example as part of a decision support system).

Finally, a certain *normative backing* is needed to plausibly allow trust to help with ethical problems (Cohen and Dienhart 2013; Nickel 2007). It is necessary that the trustor not just perform the task with which they are entrusted, but that they do so in a manner consistent with the aims and desires of the trustor. For example, an autonomous maritime system that preforms its tasks according to the aims that primarily serve its parent company (harvesting its users' data without permission, trying to shield the company from legal liability, etc.) would not have and appropriate normative backing for its behaviour and would not be trustworthy even if it performs its tasks reliably and competently. Users of autonomous maritime systems have the reasonable expectation that the system will perform its tasks primarily with their own interest in mind. If that expectation is not fulfilled, this is a violation of trust. Of course, compliance with all relevant legal obligations is never the less non-optional (and should in any case increase trustworthiness for trustors without nefarious aims of their own).

When applying the trust relation to the context of autonomous maritime systems, the operator of the system takes the role of the trustor, while the system serves as the trustee (that is, trust is placed in the system by the operator). In order for this trust to be well-founded, the operator needs to be able to correctly track the trustworthiness of the system. In other words, they need to be able to verify that the system is reliable, competent and has good normative backing.
And since the dynamic of the trust relation relies on the trustor correctly tracking the trustee's trustworthiness, we can identify how trust connects to the properties of *explainability* and *transparency*: If the operator is expected to be able to track these properties, the relevant information needs to be made easily available to them. This is another way in which trustworthiness performs its function as a technology ethics standard. The developers of technological systems need to be transparent not just with regard to the inner workings of the technology in question (how and why it arrives at certain decision and produces certain behaviours), but also with regard to the normative backing of the system as defined above. This would likely require the disclosure of not just technical details about the system, but also information about the observance of ethical standards during the development process.

A final point of interest derived from the operator interviews is the concern about responsibility. Given the analysis of trust and trustworthiness above, trustworthy systems can help with problems of responsibility, but only in certain aspects. A major problem in the context of autonomous maritime systems is what is known as the responsibility gap (Nissenbaum 1996; Santoni de Sio and Mecacci 2021; Königs 2022). A responsibility gap occurs when a task that was previously performed by a human agent who was responsible in case of accidents or failure becomes automated. This can lead to situations in which it appears like there suddenly no longer exists a good justification to call anyone responsible anymore. Trustworthy systems can help in this regard by ensuring that appropriate safety standards are in place as part of the operative reliability and competence requirements. This could potentially help to clarify in which situations an operator, developer or manufacturer of a system is for an unforeseen accident. However, the problem of responsibility gaps is unlikely to be fully resolved through the trustworthy technology alone, as it has a significant legal dimension that is outside the scope of what ethics standards alone can affect.

5. Conclusion

The MUM is designed to operate autonomously over extended periods and in demanding, high-risk environments. Its autonomous capabilities are expected to significantly expand the vessel's operational range and enable missions where human presence is limited or not feasible. However, to fully realise these advantages, control must increasingly shift from the human operator to the system itself. For this shift to be accepted, human operators must place a high degree of trust in the system—especially given the uncertainty and criticality of its operational context. In this sense, trust is not merely desirable, but a necessary condition for realising the full potential of autonomous maritime systems such as the MUM. Being highly trusted, in fact, becomes a prerequisite for technological acceptance.

This raises a central question: How can trust be systematically established in autonomous maritime operations? If we are to trust a system with high-stakes tasks, it must meet clearly defined standards. In this study, trustworthiness emerged as a key and largely uncontested requirement among participants. The qualities associated with a trustworthy system were described as interdependent and mutually reinforcing. Importantly, these expectations extended beyond the technology itself to include human actors—such as operators and developers—underscoring the inherently relational character of trust.

At the same time, the study revealed a persistent tension between the rationale for autonomy and for human control. Participants acknowledged both the benefits of autonomous operation and the continued importance of maintaining human control—often voicing both perspectives at once. Autonomy promises efficiency, adaptability, and resilience in environments where human intervention is impractical or impossible. The concept of human control as a governing ethics standard in the context of maritime autonomous vehicles is not beyond reproach itself, as the demands of this standard might be unfeasibly high and its promised ethical upsides could not be as large as might be hoped (Albrecht, Braun, Kosack and Krüger 2025). Yet, especially in areas where technical feasibility is still uncertain or where ethical and legal questions remain unresolved, participants expressed a strong desire to retain human control. This ambivalence was particularly evident in discussions around *human-machine coagency*. Although participants recognised that ongoing human involvement may contradict the core concept and purpose of autonomy many still expressed a clear preference for continuous monitoring and the possibility of intervention during missions. Even if that means that some endeavours cannot be realised for the time being. This indicates that trust in the system, at least for now, is still limited.

A key insight from the study is the distinction between actual control and perceived control. This difference plays a central role in how trust is assigned and in shaping expectations for *human-machine coagency*. When control is experienced as delegated but still preserves a sense of personal agency, users are more willing to relinquish actual control to the system. This perception of retained agency facilitates trust and increases the likelihood of acceptance. Conversely, when cognitive framing or decision support is already perceived as a loss of control, the acceptance of fully autonomous missions is likely to remain low.

6. Outlook

This paper is part of the broader MUM project and contributes to its overarching goal of ensuring successful system integration through user-centered design. The acceptance criteria identified in this study (see Fig. 1) will serve as the foundation for a subsequent quantitative survey aimed at validating their relevance and applicability. Building on the survey results, an action plan will be collaboratively developed with subject-matter experts, detailing concrete measures to foster operator acceptance. In the following phase, a tailored method will be designed to assess the extent to which these acceptance criteria are fulfilled in practice. In this endeavor, a focus on the crucial properties of reliability and competence of the autonomous system as described in chapter 4, as well as the facilitation of the necessary degree of transparency to track these properties, will be employed to remedy deficits that are discovered and to positively affect the perceived trustworthiness of the system to the stakeholders. Continuous integration of expert input throughout all stages of the process ensures that operator acceptance supporting the successful adoption of the MUM system.

References

Al, Pepijn (2023). *(E)-Trust and Its Function: Why We Shouldn't Apply Trust and Trustworthiness to Human–AI Relations.* Journal of Applied Philosophy 40 (1), 95–108. https://doi.org/10.1111/japp.12613.

Albrecht, Lukas/Braun, Hagen/Kosack, Tim/Krüger, Thomas (2025). *The Limits of Meaningful Human Control of AI in the Maritime Domain. Transactions on Maritime Science*. doi: 10.7225/toms.v14.n03.w02

Alvarado, Ramón (2023). *What kind of trust does AI deserve, if any?* AI and Ethics 3 (4), 1169–1183. https://doi.org/10.1007/s43681-022-00224-x.

Baier, Annette (1986). *Trust and Antitrust*. Ethics 96 (2), 231–260. https://doi.org/10.1086/292745.

Bringsjord, Selmer/Govindarajulu, Naveen Sundar (2022). *Artificial Intelligence*. Available online at https://plato.stanford.edu/archives/fall2022/entries/artificial-intelligence/.

Cohen, Marc A./Dienhart, John (2013). *Moral and Amoral Conceptions of Trust, with an Application in Organizational Ethics*. Journal of Business Ethics 112 (1), 1–13. https://doi.org/10.1007/s10551-012-1218-5.

Dimock, Susan (2020). *Trust and Collaboration*. In: Judith Simon (Ed.). The Routledge handbook of trust and philosophy. New York/London, Routledge, Taylor & Francis Group.

EU High Level Expert Group on Artificial Intelligence (2019*). Ethics Guidelines for Trustworthy AI*. Available online at https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

Faulkner, Paul (2007). *A Genealogy of Trust.* Episteme 4 (3), 305–321. https://doi.org/10.3366/E174236000700010X.

Gioia, Dennis A./Corley, Kevin G./Hamilton, Aimee L. (2013). *Seeking Qualitative Rigor in Indictive Research: Notes on the Gioia Methodology*. Organizational Research Methods 16 (1), 15-31.

Gambetta, Diego (Ed.) (1988). *Trust. Making and breaking cooperative relations*. New York, NY/Oxford, Basil Blackwell.

Hardin, Russell (2002). *Trust and trustworthiness*. New York, Russell Sage Foundation.

Hawley, Katherine (2017). *Trustworthy Groups and Organizations*. In: Paul Faulkner/Thomas Simpson/Thomas W. Simpson (Eds.). The philosophy of trust. Oxford/New York, Oxford University Press, 230–250.

Hieronymi, Pamela (2008). *The reasons of trust*. Australasian Journal of Philosophy 86 (2), 213–236. https://doi.org/10.1080/00048400801886496.

Hollis, Martin (1998). *Trust within reason.* Cambridge, Cambridge Univ. Press.

Holton, Richard (1994). *Deciding to trust, coming to believe*. Australasian Journal of Philosophy 72 (1), 63–76. https://doi.org/10.1080/00048409412345881.

Horsburgh, H. J. N. (1961). *Trust and Social Objectives*. Ethics 72 (1), 28–40. https://doi.org/10.1086/291373.

Jones, Karen (1996). *Trust as an Affective Attitude*. Ethics 107 (1), 4–25. https://doi.org/10.1086/233694.

Jones, Karen (2012). *Trustworthiness*. Ethics 123 (1), 61–85. https://doi.org/10.1086/667838.

Kaur, Davinder/Uslu, Suleyman/Rittichier, Kaley J./Durresi, Arjan (2023). *Trustworthy Artificial Intelligence: A Review*. ACM Computing Surveys 55 (2), 1–38. https://doi.org/10.1145/3491209.

Kelp, Christoph/Simion, Mona (2023). *What Is Trustworthiness*? Noûs. https://doi.org/10.1111/nous.12448.

Königs, Peter (2022). *Artificial intelligence and responsibility gaps: what is the problem?* Ethics and Information Technology 24 (3). https://doi.org/10.1007/s10676-022-09643-0.

Luhmann, Niklas (Ed.) (1979). *Trust and power.* Two works. 1979th ed. Ann Arbor, Mich., UMI Books on Demand.

Nickel, Philip J. (2007). *Trust and Obligation-Ascription.* Ethical Theory and Moral Practice 10 (3), 309–319. https://doi.org/10.1007/s10677-007-9069-3.

Nickel, Philip J. (2013). *Trust in Technological Systems*. In: Marc J. Vries/Sven Ove Hansson/Anthonie W.M. Meijers (Eds.). Norms in Technology. Dordrecht, Springer.

Nissenbaum, Helen (1996). *Accountability in a computerized society.* Science and Engineering Ethics 2 (1), 25–42. https://doi.org/10.1007/BF02639315.

Nyholm, Sven (2018). *Attributing Agency to Automated Systems: Reflections on Human–Robot Collaborations and Responsibility-Loci.* Science and engineering ethics 24 (4), 1201–1219. https://doi.org/10.1007/s11948-017-9943-x.

Russell, Stuart J./Norvig, Peter (2022). *Artificial intelligence. A modern approach.* Harlow, Pearson.

Ryan, Mark (2020*). In AI We Trust: Ethics, Artificial Intelligence, and Reliability.* Science and engineering ethics 26 (5), 2749–2767. https://doi.org/10.1007/s11948-020-00228-y.

Santoni de Sio, Filippo/Mecacci, Giulio (2021). *Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them*. Philosophy & Technology 34 (4), 1057–1084. https://doi.org/10.1007/s13347-021-00450-x.

Simion, Mona/Kelp, Christoph (2023*). Trustworthy artificial intelligence*. Asian Journal of Philosophy 2 (1). https://doi.org/10.1007/s44204-023-00063-5.

Taddeo, Mariarosaria (2009). *Defining Trust and E-Trust*. International Journal of Technology and Human Interaction 5 (2), 23–35. https://doi.org/10.4018/jthi.2009040102.

Taddeo, Mariarosaria (2010). *Trust in Technology: A Distinctive and a Problematic Relation.* Knowledge, Technology & Policy 23 (3-4), 283–286. https://doi.org/10.1007/s12130-010-9113-9.

Wanderer, Jeremy/Townsend, Leo (2013*). Is it Rational to Trust?* Philosophy Compass 8 (1), 1–14. https://doi.org/10.1111/j.1747-9991.2012.00533.x.