The Limits of Meaningful Human Control of AI in the Maritime Domain

Lukas Albrecht[1], Hagen Braun[1], Tim Robin Kosack[2], Thomas Krüger[3]

This paper analyses the viability of Meaningful Human Control as a mechanism to ensure ethical and safe use of autonomous systems, focusing on the maritime context. With future maritime systems increasingly containing Artificial Intelligence components as a main driver for autonomous operation, vehicles like Maritime Autonomous Surface Ships promise substantial benefits in terms of efficiency and safety. Particularly in maritime settings, where hazardous environments and dangerous working conditions put humans at risk, the deployment of autonomous systems is appealing both from an efficiency and a safety point of view – removing humans both as source of and subject to risk. This is especially true for sophisticated AI-driven autonomous systems that can be deployed in unknown environments and are able to deal with unpredicted problems, as they can operate independently from human input in a wide variety of applications. However, truly autonomous AI also introduces characteristic risks like the occurrence of Responsibility Gaps, where the ascription of responsibility for the behavior of autonomous systems is obscured, as humans are prima facie not sufficiently in control of such systems. Simply put, sophisticated AI agents are considered too autonomous for holding human agents morally responsible. If due to special ethical concerns or safety engineering reasons the human operator needs to be involved in AI decision making, human oversight and human control in a meaningful way are indispensable. To address this need for human oversight and control, the concept of Meaningful Human Control (MHC) has been introduced, primarily to guarantee the ascription of responsibility in case of harmful events. Yet, reintroducing the human element to an autonomous AI-driven system not only limits its potential, but faces conceptual and material barriers. This paper starts by looking at autonomous systems in relation to risk, before exploring the call for Meaningful Human Control and the barriers to its implementation. It concludes that there are technical and conceptual barriers that make Meaningful Human Control non-viable in some maritime applications.

KEY WORDS
~ Artificial intelligence
~ Autonomous systems
~ Maritime vessels
~ Meaningful human control

[1] German Aerospace Center – Institute for the Protection of Maritime Infrastructures, Bremerhaven, Germany
[2] German Aerospace Center – Institute for AI Safety and Security, Sankt Augustin, Germany
[3] Aerodata AG, Braunschweig, Germany
e-mail: lukas.albrecht@dlr.de
doi: 10.7225/toms.v14.n03.w02

Received: 26 Apr 2025 / Revised: 4 Jun 2025 / Accepted: 6 Jun 2025 / Published: 25 Jul 2025

1. INTRODUCTION

Future autonomous systems, in maritime applications and otherwise, will contain Artificial Intelligence (AI) components as a main driver for highly automated and autonomous operation. This development not only promises increased convenience and efficiency, but also substantial safety advantages. In part, this is because AI may replace the human element in high-risk settings and allows for the delegation of tasks – thereby eliminating human error, deliberate malpractice, and removing humans from risks altogether. However, truly autonomous AI also introduces a variety of characteristic risks, including the occurrence of Responsibility Gaps that obscure responsibility ascription for the behavior of autonomous systems, especially when such systems perform tasks previously performed by humans. This is particularly important where human well-being is at stake. If due to special ethical concerns or safety engineering reasons the human operator needs to be involved in AI decision making, human oversight and human control in a meaningful way are indispensable. To address Responsibility Gaps and others safety concerns, the concept of Meaningful Human Control (MHC) has been introduced. Yet, reintroducing the human element to a highly autonomous AI system not only limits its potential, but faces conceptual and material barriers to establish the kind of control that serves to maintain ethical and safety requirements in the maritime domain.

2. AUTONOMOUS AI AND RISK

Developments in the field of Artificial Intelligence (AI) have brought significant benefits in terms of speed, accuracy, and reliability. This allows for a more efficient accomplishment of tasks and the delegation of more and more tasks from human operators to AI agents, promising to liberate humans from menial work and increase societal well-being (Danaher, 2019; European Commission, 2022). Moreover, AI also holds the potential to significantly increase safety for humans in a number of high-risk contexts. From a safety engineering perspective, AI provides extensive possibilities to mitigate or even eliminate risks. Algorithms that identify or assess (undiscovered) risks are already in place in several domains, e.g. healthcare (European Parliamentary Research Service, 2022), finance (Milana and Ashta, 2021), and predictive maintenance (Keleko et al., 2022). Furthermore, in unexpected accident scenarios, AI-based systems may be able to react faster and more accurately, thereby lowering situational risk.

Even more benefits are expected when human operators can be replaced entirely by advanced AI agents, and therefore, are removed from any contextual risk altogether. This kind of AI deployment is particularly useful because it removes the human element both in terms of being the source of, but also subject to harm. For some time, developments in automation have led to decreasingly human involvement in operation. Yet, it is sophisticated machine autonomy that holds the promise of complete delegation of tasks to AI agents that also strive to develop the skills required to handle tasks in unknown environments without significant human input (Ezenkwu and Starkey, 2019). At present, AI systems that pursue such levels of autonomy include self-driving cars (SAE International, 2012), manned and unmanned aircrafts (European Aviation Safety Agency, 2023), industrial robots (Tantawi et al., 2019), and Lethal Autonomous Weapons Systems (LAWS) (Scharre and Horowitz, 2015).

In the maritime domain, the trend towards autonomous operation is mostly known under the umbrella term of Marine Autonomous Surface Ships (MASS) – with maritime applications like offshore maintenance, hydrography, and subsea monitoring. Various sectors have started to explore unmanned vessels, remote control operation, and even partially autonomous systems as governmental stakeholders have commissioned economic research (London Economics, 2021) industry has provided roadmaps (Rolls-Royce, 2016), and international institutes have identified critical research areas (SINTEF Ocean & Technology Centre for Offshore and Marine Singapore, 2020), with some stakeholders particularly emphasizing the potential for enhanced safety (Lloyd's Register, 2024). Furthermore, armed forces all over the world are heavily invested in more and more autonomous technologies, to which maritime capabilities are no exception, and whose advent can already be seen in several ongoing conflicts zones that involve unmanned maritime vessels (Luck, 2024; Kirichenko, 2025).

Similar to land and air bound systems that are currently shaped by AI, autonomous maritime operation promises more efficient solutions in a variety of applications. But in many regards, the safety advantages are of elevated importance in the maritime domain because of its exceptional hazardous operational setting. In extreme environments, such as the high seas and under water, the removal of human operators is especially desirable. Furthermore, dangerous working conditions that include erratic work hours, a variety of chemical, electrical, and mechanical hazards (European Maritime Safety Agency, 2022), and more recently the increased risk of piracy in some areas (International Chamber of Commerce, 2025) make autonomous operation even more appealing to the maritime domain – not least because these factors contribute to low interest among young professionals.

The transportation sector serves as a good example of one of the main use-cases in which AI is headed towards replacing the human element. Self-driving (public) transport like the Zeam ferry (Zeam, 2025) or cargo ships like Yara Birkeland (Yara International, 2025) are prominent maritime examples with the goal of autonomous operation in mind. Much like self-driving cars or aerial drones, autonomous maritime operation provides a range of socio-economic benefits, convenience for users, and the potential for a substantial increase in safety, reducing the number of accidents by eliminating human error and deliberate malpractice in a field with a substantial number of accidents involving human shortcomings (Allianz Global Corporate & Specialty AG, 2012). With the fourth and final level of the International Maritime Organization's (IMO) degree of autonomy as "[t]he operating system of the ship [being] able to make decisions and determine actions by itself" (International Maritime Organization, 2025) in mind, a prospective AI takeover of sectors like personal and public transport could reduce the amount of risk passengers are exposed to. In the case of transportation of cargo, it might even remove human involvement, and thus human exposure to risk, completely.

Further developments in the field of AI and the design of more advanced AI systems not only hold the prospect of the delegation of more tasks in general, but also of more complex tasks in particular. And while current products and demonstrators still need to rely on the human operator as a safeguard, the operational performance of AI is likely to surpass human performance at some point in the future. With increased technological maturity, especially regarding explainability as well as safety engineering, the instrumental value of AI as a tool, able to replace human operators, can be found in virtually all settings where the use of autonomous systems provides enormous potential in terms of safety and efficiency.

While the use of autonomous AI provides significant potential to eliminate risk caused by humans, and may mitigate safety risks that arise in unexpected situations, AI agents, in turn, introduce their own risks. Apart from more general worries about the use of AI for nefarious purposes or looming threats of mass unemployment associated with disruptive technological change, AI driven autonomous systems come with more specific challenges. A considerable body of literature already addresses the issues that revolve around the technical means employed to harness AI's benefits and the deployment of autonomous agents – including (data) bias (Mehrabi et al., 2021), the occurrence of Black Boxes (von Eschenbach, 2021) and a host of issues concerning algorithmic decision making (Mittelstadt et al., 2016). Yet more importantly, the prospect of replacing humans with AI agents comes with a catch. Making use of autonomous systems that find their own creative solutions to unexpected problems and that are able to operate independently in unknown environments entails the inability to fully predict their behavior and accepting the possibility of undesired outputs. Simply put, valuing AI's autonomy necessitates facing the risk of that autonomy (Sparrow, 2007).

Subsequently, the issue of holding someone responsible in the face of autonomous systems operating independently from humans, and prima facie without control over such systems, has gained a lot of attraction. The difficulty to ascribe moral responsibility for the behavior of autonomous systems has become known as the problem of Responsibility Gaps (Matthias, 2004). In such cases it seems appropriate to blame someone for a harmful outcome of the operation of an autonomous system, but no individual can be justifiably identified to be blamed, because no individual is considered to be sufficiently in control of that operation. In other words, sophisticated AI agents are considered too autonomous for holding human agents morally responsible. The incident of an Uber vehicle in autonomous mode running over a pedestrian (Nyholm, 2023) is a much-quoted example when discussing gaps in responsibility, in which neither operator nor designer seem to be sufficiently in control to be held morally responsible for the outcome of the system's behavior. And with AI driven systems being deployed more often and with continuously more autonomous functions, similar cases are likely to arise more frequently.

These risks concerning the use of AI are not novel in the discussion on the governance of AI, but they are of particular relevance when discussing AI autonomy in this regard, as the introduction of new risks goes against the endeavor of minimizing them. More importantly, the types of risk mentioned above that the use of AI introduces are unlike those that are managed with the help of AI. Whereas safety related risk is expected to be managed more effectively with increased robustness and precision of AI systems, this does not apply to non-technical risk caused by their employment. For example, in the frequently discussed use case of LAWS, various stakeholders have endorsed the paramount importance of responsibility and dignity (Article 36, 2013; Human Rights Watch, 2012; International Committee for Robot Arms Control, 2009; International Committee of the Red Cross, 2014; United Nations Institute for Disarmament Research, 2014). In this case, the deployment of truly autonomous AI agents that are susceptible to gaps in responsibility is not feasible, regardless of how safe and effective they may be from a technical point of view.

The issues at hand, especially the risks regarding autonomy in AI agents, are not properly addressed by simply improving AI systems deployed in the context of risk management, if these improvements do not also manage the non-technical risk that they themselves cause. Hence, in contexts where aspects pertaining to these risks of AI play a key role, the use of autonomous AI might not be viable, and thus, prevent sophisticated AI systems from successful implementation. Unsurprisingly, addressing these challenges has become a pressing subject in the debate on AI governance, especially with regard to the above-mentioned aspects such as bias, explainability, and responsibility (Jobin et al., 2019; Kaur et al., 2023).

3. THE CALL FOR MEANINGFUL HUMAN CONTROL

One of the most prominent approaches to manage some of the characteristic risks of AI is the concept of Meaningful Human Control (MHC). Originating in the discussion on LAWS and the concern of delegating the use of force and decisions over life and death to machines (Article 36, 2013; Amoroso and Tamburrini, 2020; Ekelhof, 2019; Santoni de Sio and van den Hoven, 2018; Scharre and Horowitz, 2015; Schwarz, 2021), MHC has since become a popular instrument to manage various risks that the implementation of AI systems has introduced. The concept has spread to other domains like automated or autonomous driving systems (Mecacci and Santoni de Sio, 2020; Santoni de Sio et al., 2022) and automated decision-making systems (Cornelissen, 2022; Wagner, 2019). Yet, despite its apparently ubiquitous endorsement, there is no agreement on what exactly constitutes MHC (Ekelhof 2019; Cummings 2019; Davidovic, 2023). Furthermore, discussions on the subject frequently fail to clearly specify the purpose of implementing MHC – i.e. whether it shall increase safety, ensure responsibility or serve a completely different purpose (European Commission, 2019). Nonetheless, it should be stressed that MHC has one very specific advantage: It is the only answer to the problem of Responsibility Gaps that has been developed so far. MHC closes Responsibility Gaps by keeping a human operator close enough to the individual decisions made by an AI system that they can genuinely be held responsible for the consequences and possible harms caused by such a system.

MHC is not the only conceptual tool that has been explored in the context of AI ethics. For example, the most influential AI ethics and governance document of the recent past, the Ethics Guidelines for Trustworthy AI by the European Commission (2019) does not use the terminology of "meaningful human control" at all, and instead opts to engage with the matter in terms of "human agency and oversight" (pp. 15-16). In this context, three methods of oversight are specifically described: Human-in-the-loop (HITL), human-on-the-loop (HOTL) and human-in-command (HIC). These approaches to human oversight can be distinguished by the degree of direct control that a human exerts over individual decisions: whereas with HITL, a human has the capacity to intervene in every individual decision cycle of the AI, with HOTL, human input is limited to intervention during the design cycle and monitoring roles. HIC can be viewed as the minimum that is potentially compatible with the ethical AI use, as the degree of human input in this governance method is limited to decisions of when and how to use AI systems.

What differentiates MHC from mere human oversight is the degree of immediacy with which a human is involved in individual decisions made by the AI system. As perhaps the most important upside of MHC is supposed to be the avoidance of Responsibility Gaps, it is necessary that a human operator is involved in all ethically weighty decisions made by the AI system. Therefore, with these methods of oversight, as direct human control decreases, so does Meaningful Human Control: HITL exhibits the highest compatibility with the concept, since, by definition, a human is able to intervene in every decision cycle when a HITL approach is employed. The HOTL approach is still compatible with MHC, as human monitoring of AI performance may be sufficient to ensure ethical adequacy and responsible use in less ethically sensitive contexts. The HIC method should be viewed as incompatible with MHC, since following this approach individual decisions made by AI agents after deployment do not fall under the control of humans at all.

4. BARRIERS TO MEANINGFUL HUMAN CONTROL

The main advantage of MHC is its promise to ensure that responsibility is preserved in high automation and autonomy contexts. Keeping human operators meaningfully in control of AI-driven systems such that these operators can be held responsible for the behavior of the system takes two requirements to be met: First, the human operator needs to be *significantly involved* in the decision-making process and second, the human operator needs the *necessary expertise* to evaluate the decisions of the AI system in an informed way. However, there are conceptual and material barrier to the implementation of this kind of control.

Direct involvement in every decision cycle, such as when HITL is applied as the method of human oversight, does not guarantee by itself Meaningful Human Control. The problems of rubber stamping and automation bias illustrate why that might be the case. Rubber stamping refers to a human operator accepting an AI decision without the ability to properly assess it. In the case of rubber stamping, a human operator is nominally in control of the AI system, authorizing or validating AI decisions before they are executed. However, due to a lack of expertise on part of the operator or other contravening factors such as time pressure, the presence of the human operator does not actually result in improvements in the problem areas where MHC is supposed to be a solution. For example, while a human operator can in principle act as a nexus of responsibility when a fully autonomous AI system cannot, this is not the case for obviously unqualified operators or those that feel pressured to quickly authorize AI decision in order to not undermine the performance of the system. The process can be thought of as a person pressing a button every time they see a light turn on. If the light signifies an ethically weighty decision made by an AI, the human operator authorizing it by pressing the button obviously does nothing to improve the

adequacy of an AI-enhanced system in terms of the relevant ethical dimensions and does not make the operator meaningfully responsible for any resultant harm.

However, non-meaningful control can also occur in contexts where the human operator is, in theory, able to properly validate the decision of an AI agent. In these cases, the problem of automation bias can potentially undermine meaningful human control. Automation bias occurs when the judgement of an expert operator is undermined by giving undue weight to the output of an automated (in this case, AI) system. It is a documented phenomenon that humans tend to put increased weight on data provided by automated systems even when trusting them goes against their own, best judgement (Goddard et al. 2012; Skitka et al. 1999). The ALTAI addresses this problem with the requirement that AI systems should not undermine human agency. Rather, it must be ensured that AI systems enhance human decision-making abilities. Automation bias as a general phenomenon is a barrier to this aim, since it occurs on a subconscious level. However, the existence of the phenomenon shows that implementing MHC is not trivial even in cases where expert human operators who are not subject to contravening situational pressure are confronted with highly pre-refined judgements of AI agents.

Ensuring significant involvement in the decision-making process is not trivial in many contexts in which high automation or autonomy is desirable, either. In the maritime domain, communication with autonomous ships is significantly limited by the low bandwidth available at sea. Due to the high rate of absorption of electromagnetic waves in water, radio-based communications are generally not feasible in the underwater environment. Alternative methods of communications come with downsides that make the implementation of MHC difficult, such as the limited range of wired communications or the significant latency inherent to acoustic communications. This means that a significant barrier to fast, reliable, long-range communication exists in one of the prime use-cases of AI in the maritime field, (partially-)autonomous underwater vehicles (AUVs) (Aziz El-Banna and Wu 2021). The more constrained communication between the operator and AI system becomes, the larger the material barrier to the implementation of an MHC paradigm grows: A human operator cannot be expected to be able to intervene in or supervise every decision cycle of the AI system when there is no reliable means of communication between the two parties.

Moreover, issues regarding maritime communication become more pronounced when considering the regulatory framework relevant to MASS. Governing most aspects of international law relating to the sea, the United Nations Convention on the Law of the Sea (UNCLOS) implicitly assumes that ships are crewed, with Art. 94(3b) referring to “the manning of ships” and (4b) stating “that each ship is in the charge of a master and officers” (United Nations Convention on the Law of the Sea 1982). To reconcile the advent of autonomous maritime operation and unmanned vessels with the regulatory framework, a working group by the IMO is addressing legal issues related to MASS. It agrees that “there should be a human master responsible for a MASS, regardless of mode of operation or degree or level of autonomy [and] have the means to intervene when necessary.” Yet, the working group also states that “such master may not need to be on board, depending on the technology used on the MASS and human presence on board, if any” (Maritime Safety Committee 2023). While enabling unmanned operation and potentially high degrees of autonomy, these considerations also stress the importance of responsibility. This is unsurprising, as removing the crew from vessels also removes an *immediate* human nexus for responsibility in cases of harmful events. On these grounds, insisting on a responsible master to have the means to intervene retains this nexus of responsibility, as long as the kind of significant involvement in the decision-making process relevant to the MHC paradigm is ensured. But again, control and supervision while not being on board requires reliable means of communication.

Considering all the (raw) data necessary for safe and responsible operation shows how difficult this task is. This includes information like velocity, swell, inclination, and any status from mechanical systems that is provided by sensors available on board. Yet, skilled personnel on site can also take any immediate feedback into account that they receive during operation and that is not fully quantifiable with current sensor technology, like the ship’s dynamic response, irregular audio or vibration patterns, or other environmental information. To transfer this kind of data would not only require additional bandwidth, but also supplementary tech solutions that provide an equivalent to the senses of traditional on-site operators. The difficulty to substitute such a comprehensive perception and the experience in operating and controlling a vessel is also reflected in the challenge of realizing Good Seamanship in the context of the Convention on the International Regulations for Preventing Collisions at Sea (COLREGs) (Margat & Stadermann, 2024). It is worth noting that remote-control centers without access to this information may still be able to meaningfully intervene, and thus stay in control of an autonomous vessel, but have to accept the possibility of decreased safety.

All these issues become even more problematic when a mismatch between the method of oversight and the desired degree of autonomy of the AI system occurs, such as when a human-in-the-loop approach is applied to an otherwise fully autonomous system. At first glance, a more extensive degree of human involvement appears to be desirable in this context, since the loss of human involvement in the decision-making process is the source of the types of issues that MHC is supposed to address (Responsibility Gaps, ethical inadequacy, etc.). However, we can observe that this kind of approach will either

limit the performance of the system as a whole to ensure meaningful control, or undermine the meaningfulness of the control exerted by the human operator in order to maintain performance.

The idea of MHC comes full circle with the reintroduction of expert operators as controllers of AI systems. Whereas humans were originally excluded from specific tasks to reap the benefits of highly automated or autonomous AI systems, the reintroduction of the human operator leads the whole process back to its beginning. The starting point is marked by a task that is being executed without assistance of AI. In order to achieve the benefits of autonomous AI – better performance in terms of speed, accuracy or safety and the delegation of more complex tasks – the involvement of AI reaches a level where the human being is increasingly excluded from the task. Various problems follow from this exclusion of the human operator, both ethical and practical. One of the answers prospective to this development is MHC, where the reintroduction of the human operator is the essential aspect. But with the key aspect being human control, MHC is conceptually incompatible with the highest levels of AI autonomy. The reasons to deploy highly autonomous AI systems are simply in too much tension with the idea of a human operator being meaningfully involved in all relevant decisions made by that AI. Thus, in any scenario in which the removal of the human operators is an advantage in itself (for reasons of safety, efficiency or otherwise), MHC loses applicability.

This is not the case at lower levels of autonomy. Human operators are already a necessary element in decision support systems and human-AI teaming set ups. These two approaches differ from the case described above, in which a human operator acts as a vetoing agent to an otherwise autonomous decision-making AI. Instead, the human operator is assumed to be in charge of decision-making ab initio, with the purpose of the AI agent being to either enhance their capabilities (support systems) or to take over ethically non-critical parts of the overall operation (human-AI teaming).

5. THE LIMITS OF MEANINGFUL HUMAN CONTROL

If we accept that MHC is not a viable solution to problems relating to responsibility and ethical concerns that arise in the context of highly autonomous AI systems, that raises the question: What is the use of MHC? The answer might be that MHC gives us insight regarding the appropriate method of human oversight in use contexts with different levels of ethical relevance. If, for example, guaranteeing ethical standards and avoiding Responsibility Gaps is non-negotiable in the context of autonomous weapons systems, and MHC is the only way to meet this goal, then we can conclude that the use of fully autonomous AI weapons is unwarranted, and that at most we should be considering decision support systems or human-AI teaming approaches in this context.

Of course, not all domains are characterized by such high degrees of ethical sensitivity. It is not a conceptual necessity that MHC is a requirement for all use cases of AI. For example, if a sufficient level of safety can be clearly demonstrated for highly autonomous (maritime) vehicles in contexts such as IMO level 4 – similar to self-driving cars' level 5 "full driving automation" (SAE International, 2021) or aviation's level 3 "advanced automation" (European Aviation Safety Agency, 2023) – it becomes difficult to articulate what further ethical barrier to the use of such systems would still remain. The Ethics Guidelines for Trustworthy AI note: "Oversight mechanisms can be required in varying degrees to support other safety and control measures, depending on the AI system's application area and potential risk (European Commission, 2019). All other things being equal, the less oversight a human can exercise over an AI system, the more extensive testing and stricter governance is required" (p. 16). If the area of application is not subject to special ethical concerns and risk is demonstrably low, governing mechanisms other than MHC can take over. This illustrates an upside to the Trustworthy AI approach of the ALTAI guidelines, since it is compatible with MHC in contexts where direct control is necessary, but provides mechanisms to ensure ethical AI development and use in contexts incompatible with MHC, as well.

6. CONCLUSION

MHC is a valuable concept in AI ethics that helps us understand which type of human oversight mechanism is necessary for different use cases for AI systems. It is the only oversight mechanism developed so far that can truly be said to deal with the problem of Responsibility Gaps. However, the approach is conceptually incompatible with highly autonomous AI systems, as it is impossible to both realize the strict control required for a human operator to truly be responsible for the decisions of the AI system and maintain the benefits that highly autonomous AI provides in terms of speed, accuracy, and safety. Domains in which communication between an AI system and human operator are harder to realize, such as the underwater domain in the maritime context, suffer from additional material barriers that could further complicate the implementation of MHC. The existence of certain domains that naturally resist an MHC approach should be recognized as a limiting factor for the feasibility of MHC as a general approach to AI ethics. In some domains, where the problem of Responsibility Gaps is of elevated importance, this implies that the use of highly autonomous AI systems is unjustifiable on ethical grounds. In these contexts, alternative approaches such as AI decision support and human-AI teaming should be pursued instead. However, in less ethically demanding contexts, one may find other human oversight paradigms that do not

preclude the use of highly autonomous systems, if a sufficient level of safety can be demonstrated. The authors declare that they have no known financial or non-financial conflicting interests in any material discussed in this paper.

CONFLICT OF INTEREST

Authors declare that they have no known financial or non-financial conflicting interests in any material discussed in this paper.

NOTICE

The present paper has been selected to be published as an extended version of the authors' previous work on this subject. It is an extension of the conference paper 'The Limits of Meaningful Human Control of AI in the Maritime Domain' presented at the 4th European Workshop on Maritime Systems, Resilience and Security 2024 (MARESEC 2024), Bremerhaven, Germany (doi: https://doi.org/10.5281/zenodo.14214794). The overall structure of the original paper has been modified to meet the requirements of this journal. The content of this paper is based on the previously published version, but it has been revised and supplemented with several new aspects.

REFERENCES

Allianz Global Corporate & Specialty AG (2012) Safety and Shipping 1912–2012. Available at: https://www.allianz.com/content/dam/onemarketing/azcom/Allianz_com/migration/media/press/document/other/agcs_safety_shipping_1912-2012.pdf (Accessed: 19 March 2025).

Amoroso, D. and Tamburrini, G. (2020) 'Autonomous weapons systems and meaningful human control: Ethical and legal issues', Current Robotics Reports, 1(4). Available at: https://doi.org/10.1007/s43154-020-00024-3.

Article 36 (2013) Killer Robots: UK Government Policy on Fully Autonomous Weapons. Available at: https://article36.org/wp-content/uploads/2013/04/Policy_Paper1.pdf (Accessed: 19 March 2025).

Aziz El-Banna, A. A. and Wu, K. (2021) 'Introduction to underwater communication and IoUT networks', in Machine Learning Modeling for IoUT Networks. SpringerBriefs in Computer Science. Cham: Springer. Available at: https://doi.org/10.1007/978-3-030-68567-6_1 (Accessed: 19 March 2025).

Cornelissen, N. A. J., van der Arend, S., Maas, L., and Dignum, F. (2022) 'Reflection machines: Increasing meaningful human control over decision support systems', Ethics and Information Technology, 24(2). Available at: https://doi.org/10.1007/s10676-022-09645-y.

Cummings, M. L. (2019) 'Lethal autonomous weapons: Meaningful human control or meaningful human certification?', IEEE Technology and Society Magazine, 38(4). Available at: https://doi.org/10.1109/mts.2019.2948438.

Danaher, J. (2019) Automation and Utopia: Human Flourishing in a World Without Work. Cambridge, MA: Harvard University Press. Available at: https://doi.org/10.2307/j.ctvn5txpc.

Davidovic, J. (2023) 'On the purpose of meaningful human control of AI', Frontiers in Big Data, 5. Available at: https://doi.org/10.3389/fdata.2022.1017677.

Ekelhof, M. A. C. (2019) 'Moving beyond semantics on autonomous weapons: Meaningful human control in operation', Global Policy, 10(3). Available at: https://doi.org/10.1111/1758-5899.12665.

European Aviation Safety Agency (2023) EASA Concept Paper: First Usable Guidance for Level 1 & 2 Machine Learning Applications – A Deliverable of the EASA AI Roadmap. Available at: https://www.easa.europa.eu/en/document-library/generalpublications/easa-artificial-intelligence-concept-paper-proposedissue-2 (Accessed: 19 March 2025).

European Commission (2019) Ethics Guidelines for Trustworthy AI. Available at: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419 (Accessed: 19 March 2025).

European Commission (2022) Annual Statistical Report. Available at: https://road-safety.transport.ec.europa.eu/europeanroad-safety-observatory/data-and-analysis/annual-statistical-report_en (Accessed: 19 March 2025).

European Maritime Safety Agency (2022) European Maritime Safety Report 2022. Available at: https://doi.org/10.2808/914730.

European Parliamentary Research Service (2022) Artificial Intelligence in Healthcare: Applications, Risks, and Ethical and Societal Impacts. European Parliament. Available at: https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2022)729512 (Accessed: 19 March 2025).

Ezenkwu, C. P. and Starkey, A. (2019) 'Machine autonomy: Definition, approaches, challenges and research gaps', in Proceedings of the Computing Conference, London, United Kingdom, 16–17 July, pp. 335–358. Available at: https://doi.org/10.1007/978-3-030-22871-2_24.

Ferencz, C. and Zöldy, M. (2024) 'Exhaustive investigation of the promises and perils of autonomous mobility technology', Periodica Polytechnica Transportation Engineering, 52(1). Available at: https://doi.org/10.3311/PPtr.22573.

Goddard, K., Roudsari, A., and Wyatt, J. C. (2012) 'Automation bias: A systematic review of frequency, effect mediators, and mitigators', Journal of the American Medical Informatics Association, 19(1). Available at: https://doi.org/10.1136/amiajnl-2011-000089.

Human Rights Watch (2012) Losing Humanity: The Case Against Killer Robots. Available at: https://www.hrw.org/report/2012/11/19/losing-humanity/case-againstkiller-robots (Accessed: 19 March 2025).

International Chamber of Commerce (2025) 'Maritime piracy dropped in 2024 but crew safety remains at risk'. Available at: https://iccwbo.org/news-publications/news/maritime-piracy-dropped-in-2024-but-crew-safety-remains-at-risk/ (Accessed: 19 March 2025).

International Committee for Robot Arms Control (2009) Mission Statement. Available at: http://www.icrac.net/statements/ (Accessed: 19 March 2025).

International Committee of the Red Cross (2014) Autonomous Weapon Systems: Technical, Military, Legal and Humanitarian Aspects. Available at: https://www.icrc.org/en/document/report-icrc-meeting-autonomousweapon-systems-26-28-march-2014 (Accessed: 19 March 2025).

International Maritime Organization (2025) Autonomous Shipping. Available at: https://www.imo.org/en/MediaCentre/HotTopics/Pages/Autonomous-shipping.aspx (Accessed: 19 March 2025).

Jobin, A., Ienca, M., and Vayena, E. (2019) 'The global landscape of AI ethics guidelines', Nature Machine Intelligence, 1(9). Available at: https://doi.org/10.1038/s42256-019-0088-2.

Kaur, D., Gupta, D., Khanna, A., and Arora, A. (2023) 'Trustworthy artificial intelligence: A review', ACM Computing Surveys, 55(2). Available at: https://doi.org/10.1145/3491209.

Keleko, A. T., Adepoju, S. A., and David, O. M. (2022) 'Artificial intelligence and real-time predictive maintenance in industry 4.0: A bibliometric analysis', AI and Ethics, 2(4). Available at: https://doi.org/10.1007/s43681-021-00132-6.

Kirichenko, D. (2025) 'Ukraine's marauding sea drones bewilder Russia'. Available at: https://cepa.org/article/ukraines-marauding-sea-drones-bewilder-russia/ (Accessed: 19 March 2025).

Lloyd's Register (2024) Maritime Autonomous Surface Ships (MASS): Creating a Framework for Efficiency, Safety and Compliance. Available at: https://www.lr.org/en/knowledge/research-reports/2024/maritime-autonomous-surface-ships/ (Accessed: 19 March 2025).

London Economics (2021) Consultancy Research into the UK Maritime Technology Sector. Available at: https://londoneconomics.co.uk/wp-content/uploads/2021/11/LE-DfT_SmartShipping-Final-Report.pdf (Accessed: 19 March 2025).

Luck, A. (2024) 'Chinese experimental aviation platform and combat USV emerge in detailed new imagery'. Available at: https://www.navalnews.com/naval-news/2024/11/chinese-experimental-aviation-platform-and-combat-usv-emerge-in-detailed-new-imagery/ (Accessed: 19 March 2025).

Margat, P. and Stadermann, M. (2024) 'The application of COLREGs by autonomous and unmanned vessels: Issues raised by situational awareness, night-time navigation and good seamanship', in Proceedings of the MARESEC 2024, Bremerhaven, Germany, 6–7 June. Available at: https://doi.org/10.5281/zenodo.14214447.

Matthias, A. (2004) 'The responsibility gap: Ascribing responsibility for the actions of learning automata', Ethics and Information Technology, 6(3). Available at: https://doi.org/10.1007/s10676-004-3422-1.

Maritime Safety Committee (2023) Development of a Goal-Based Instrument for Maritime Autonomous Surface Ships (MASS). Available at: https://wwwcdn.imo.org/localresources/en/MediaCentre/HotTopics/Documents/MSC%20107-5-1-Report%20of%20the%20MSC-LEG-FAL%20Joint%20Working%20Group.pdf (Accessed: 19 March 2025).

Mecacci, G. and Santoni de Sio, F. (2020) 'Meaningful human control as reason-responsiveness: The case of dual-mode vehicles', Ethics and Information Technology, 22(2). Available at: https://doi.org/10.1007/s10676-019-09519-w.

Mehrabi, N. (2021) 'A survey on bias and fairness in machine learning', ACM Computing Surveys, 54(6). Available at: https://doi.org/10.1145/3457607.

Milana, C. and Ashta, A. (2021) 'Artificial intelligence techniques in finance and financial markets: A survey of the literature', Strategic Change, 30(3). Available at: https://doi.org/10.1002/jsc.2403.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S. and Floridi, L. (2016) 'The ethics of algorithms: Mapping the debate', Big Data & Society, 3(2). Available at: https://doi.org/10.1177/2053951716679679.

Nyholm, S. (2023) 'Responsibility gaps, value alignment, and meaningful human control over artificial intelligence', in Placani, A. and Broadhead, S. (eds.) Risk and responsibility in context. New York: Routledge, pp. 191–213. Available at: https://doi.org/10.4324/9781003276029.

Robbins, S. (2023) 'The many meanings of meaningful human control', AI and Ethics, pp. 1–12. Available at: https://doi.org/10.1007/s43681-023-00320-6.

Rolls-Royce (2016) Autonomous ships: The next step. Available at: https://www.rolls-royce.com/~/media/Files/R/Rolls-Royce/documents/%20customers/marine/ship-intel/rr-ship-intel-aawa-8pg.pdf (Accessed: 19 March 2025).

SAE International (2021) Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles: J3016_202104. Available at: https://www.sae.org/standards/content/j3016_202104/ (Accessed: 19 March 2025).

Santoni de Sio, F., Mecacci, G., Rebera, A. and Yampolskiy, M. (2022) 'Realising meaningful human control over automated driving systems: A multidisciplinary approach', Minds and Machines, 33. Available at: https://doi.org/10.1007/s11023-022-09608-8.

Santoni de Sio, F. and van den Hoven, J. (2018) 'Meaningful human control over autonomous systems: A philosophical account', Frontiers in Robotics and AI, 5. Available at: https://doi.org/10.3389/frobt.2018.00015.

Scharre, P. and Horowitz, M. C. (2015) An introduction to autonomy in weapon systems. Center for a New American Security. Available at: https://s3.us-east-1.amazonaws.com/files.cnas.org/hero/documents/EthicalAutonomy-Working-Paper_021015_v02.pdf (Accessed: 19 March 2025).

Schwarz, E. (2021) 'Autonomous weapons systems, artificial intelligence, and the problem of meaningful human control', Philosophical Journal of Conflict and Violence, 5(1). Available at: https://doi.org/10.22618/TP.PJCV.20215.1.139004.

SINTEF Ocean and Technology Centre for Offshore and Marine Singapore (2020) R&D roadmap for smart and autonomous sea transport systems. Available at: https://www.smashroadmap.com/files/2020-10-SINTEF-TCOMS---RD-Roadmap-for-Smart--Autonomous-Sea-Transport-Systems.pdf (Accessed: 19 March 2025).

Skitka, L. J., Mosier, K. L. and Burdick, M. (1999) 'Does automation bias decision-making?', International Journal of Human-Computer Studies, 51(5). Available at: https://doi.org/10.1006/ijhc.1999.0252.

Sparrow, R. (2007) 'Killer robots', Journal of Applied Philosophy, 24(1). Available at: https://doi.org/10.1111/j.1468-5930.2007.00346.x.

Tantawi, K. H., Ali, K. M., Murad, M. A. and Rahman, A. A. (2019) 'Advances in industrial robotics: From industry 3.0 automation to industry 4.0 collaboration', in Proceedings of the 4th Technology Innovation Management and Engineering Science International Conference (TIMES-ICON), Bangkok, Thailand, 11–13 December, pp. 1–4. Available at: https://doi.org/10.1109/TIMES-iCON47539.2019.9024658.

United Nations Convention on the Law of the Sea (1982) Available at: https://www.un.org/Depts/los/convention_agreements/convention_overview_convention.htm (Accessed: 19 March 2025).

United Nations Institute for Disarmament Research (2014) The weaponization of increasingly autonomous technologies: Considering how meaningful human control might move the discussion forward. Available at: https://unidir.org/files/publication/pdfs/considering-how-meaningfulhuman-control-might-move-the-discussion-forward-en-615.pdf (Accessed: 19 March 2025).

von Eschenbach, W. J. (2021) 'Transparency and the black box problem: Why we do not trust AI', Philosophy & Technology, 34(4). Available at: https://doi.org/10.1007/s13347-021-00477-0.

Wagner, B. (2019) 'Liable, but not in control? Ensuring meaningful human agency in automated decision-making systems', Policy & Internet, 11(1). Available at: https://doi.org/10.1002/poi3.198.

Yara International (2025) Yara Birkeland press kit. Available at: https://www.yara.com/news-and-media/media-library/press-kits/yara-birkeland-press-kit/ (Accessed: 19 March 2025).

Zeam (2025) Zeam autonomous and electric ferry. Available at: https://www.zeam.se/en (Accessed: 19 March 2025).