Journal of
Marine Science and Engineering

MDPI

Article

Human-Centered Explainable Artificial Intelligence for Marine Autonomous Surface Vehicles

Erik Veitch * and Ole Andreas Alsos

Department of Design, Norwegian University of Science and Technology (NTNU), Kolbjørn Hejes Vei 2b, 7491 Trondheim, Norway; ole.alsos@ntnu.no
* Correspondence: erik.a.veitch@ntnu.no

Abstract: Explainable Artificial Intelligence (XAI) for Autonomous Surface Vehicles (ASVs) addresses developers' needs for model interpretation, understandability, and trust. As ASVs approach wide-scale deployment, these needs are expanded to include end user interactions in real-world contexts. Despite recent successes of technology-centered XAI for enhancing the explainability of AI techniques to expert users, these approaches do not necessarily carry over to non-expert end users. Passengers, other vessels, and remote operators will have XAI needs distinct from those of expert users targeted in a traditional technology-centered approach. We formulate a concept called 'human-centered XAI' to address emerging end user interaction needs for ASVs. To structure the concept, we adopt a model-based reasoning method for concept formation consisting of three processes: analogy, visualization, and mental simulation, drawing from examples of recent ASV research at the Norwegian University of Science and Technology (NTNU). The examples show how current research activities point to novel ways of addressing XAI needs for distinct end user interactions and underpin the human-centered XAI approach. Findings show how representations of (1) usability, (2) trust, and (3) safety make up the main processes in human-centered XAI. The contribution is the formation of human-centered XAI to help advance the research community's efforts to expand the agenda of interpretability, understandability, and trust to include end user ASV interactions.

Keywords: human-AI interaction; human-centered design; autonomous surface vehicles; shore control center; explainable AI; automation transparency; collaborative systems; trust

check for updates

Citation: Veitch, E.; Alsos, O.A. Human-Centered Explainable Artificial Intelligence for Marine Autonomous Surface Vehicles. *J. Mar. Sci. Eng.* **2021**, *9*, 1227. https://doi.org/10.3390/jmse9111227

Academic Editor: Rosemary Norman

Received: 1 October 2021
Accepted: 4 November 2021
Published: 6 November 2021

1. Introduction

Artificial Intelligence (AI) is being increasingly used in maritime applications. This is perhaps most clearly seen in Autonomous Surface Vehicles (ASVs), a category of maritime vessels that emerged in oceanographic and marine biological data collection [1–3] and has recently branched out into urban mobility [4–6] (Figure 1). A related category is the Maritime Autonomous Surface Ship (MASS), now formally acknowledged by the International Maritime Organization (IMO) [7], the world regulatory agency for safety and environmental protection at sea. MASS are distinct from ASVs primarily in scale and by virtue of their reliance upon a hybrid format of AI-human collaborative seafaring coordinated from a 'Shore Control Center' (known alternatively as 'Remote Control Station/Center') [8–10]. However, trends toward the development of ASVs with passengers and automated shipboard navigation systems, such as auto-crossing and auto-docking [11,12], are blurring the line between the two categories. Work towards scaling MASS and ASVs into widespread use raises new challenges related to ensuring that AI system goals are aligned with the values of those who will be interacting with them. This is broadly the motivation behind the growing field of Explainable AI (XAI), characterized, as expressed by [13], by its mission to 'create AI systems whose learned models and decisions can be understood and appropriately trusted by end users' (p. 44). This mission is necessarily multi-disciplinary, meeting at the crossroads of fields as diverse as cognitive science, human-computer interaction, cybernetics, safety engineering, computer science, human factors, sociology, and others.

The work in this article was motivated by the growing need for such a multi-disciplinary XAI focus in ASV development. Our objective was to investigate the extent to which a human-centered design approach to XAI can contribute to aligning ASV technology towards real-world stakeholders.

The relationship between humans and technology is at the core of the fields of interaction design and human-computer interaction. Within these fields, practical approaches for designing AI for human use have existed for at least two decades. Early work envisioned 'mixed-initiative user interfaces' [14] focused on effective collaboration between AI systems and human users and on design approaches incorporating safeguards against unintended outcomes of autonomous agents [15]. More recently, scholars have proposed human-centered AI design frameworks [16] that aim to reconcile advancements in machine autonomy with humans' fundamental need for their own autonomy. In the past decade, these efforts have taken on new urgency. Hardware advancements, for example, have unlocked possibilities in Machine Learning (ML) previously considered unfeasible (e.g., Graphical Processor Units, [17]), as have the growth of open-source training datasets and prediction competitions (e.g., ImageNet, [18]). The adoption of AI systems into autonomous cars and passenger ASVs has raised the stakes of unintended consequences, with the possibility for real harm for people involved.

The recent growth of the XAI field is a testament to the broad range of disciplines contributing. New forums for scientific discussion have emerged, like the ACM conference on fairness, accountability, and transparency (https://facctconference.org/, accessed on 22 October 2021). Here, implications surrounding lack of predictability of AI systems are weighed against their benefits. 'Responsible AI' [19] expands the XAI audience from its core of computer scientists addressing 'black box' networks [20] towards 'large-scale implementation of AI methods in real organizations with fairness, model explainability, and accountability at its core' [19] (p. 82). Despite the field's widespread growth, however, it remains unclear how the values at the core of XAI—interpretability, understandability, explainability, and trust—will be practically addressed in system design.

Some critics of XAI have shown that increases in model interpretability generally lead to reduced performance [21]. Moreover, experiments conducted by [22] showed that increasing transparency of an ML model may not influence user trust at all and may even detract from users' ability to notice mistakes. Reporting on the latter, the authors emphasized 'the importance of testing over intuition when developing interpretable models' [22] (p. 1). Still, the sentiment among AI researchers suggests that model interpretability stands to benefit developers, users, and downstream societal stakeholders by virtue of a better understanding of AI system mechanisms [23]. Researchers envision AI systems offering societal benefits by supporting and enhancing human decision-making [24], including 'hybrid systems' composed of autonomous agents and humans working together [25].

ASV use cases present unique challenges to XAI. While safety records have steadily improved, the maritime environment is still considered a dangerous one, with a high rate of fatal injuries and high consequences for accidents [26,27]. The IMO calls shipping 'one of the most dangerous' of the world's industries [28]. The barrier for trust in passenger ASVs seems especially high, with one recent survey by [29] suggesting that public perception of autonomous ferries is conditional upon onboard operator presence. IMO's recent 'Regulatory Scoping Exercise for the Use of Maritime Autonomous Surface Ships,' initiated in 2017 with the aim of building a new regulatory framework, found current regulatory shortcomings so 'complex and extensive' that they suggested a new 'MASS Code' is needed [7] (pp. 8–9). Among the top high-priority regulatory gaps listed by the scoping exercise were issues related to interactions between AI navigation systems and human backup control and oversight. Efforts to address conventions, such as Safety of Life at Sea (SOLAS), Standards of Training, Certification and Watchkeeping (STCW), and Collision Regulations (COLREGS) extend the aim of AI system interpretability and accountability to operational and regulatory oversight domains.

This work investigates the issues surrounding the explainability and trust of AI systems from the lens of a practice-based, human-centered design approach tailored specifically to the needs of ASVs. From this perspective, efforts towards automation transparency aiming to improve model understandability are just one important factor among several others, including the affordances necessary for understanding its use and establishing trust among a broader stakeholder base emerging from more widespread deployment. Our research question is: can a human-centered approach to XAI contribute to building trust among real-world ASV users?

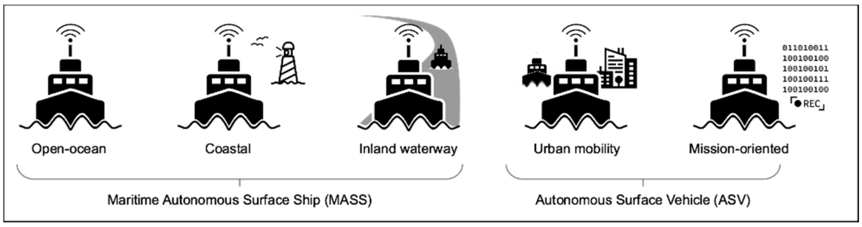


Figure 1. Categories of MASS and ASVs.

2. Method

Our aim in this paper was to introduce the concept of human-centered XAI for ASV applications. In this aim, methods describing the process of creating scientific concepts were particularly helpful. We drew inspiration from the cognitive and social sciences, and used the principles behind model-based reasoning [30] to structure our findings. Model-based reasoning posits that concept formation in science occurs through analogy, visualization, and mental simulation. We organized the findings accordingly, with analogy, visualization, and mental simulation each allotted their own sub-section. These sub-sections contain examples from our own research and from recent research of our peers at our university. The examples illustrate an important point: they are not the findings of the article in themselves; rather, they show how, through the lens of model-based reasoning, they can be used to construct the concept of human-centered XAI for ASVs.

The social sciences have placed great importance on the process of concept formation, possibly even more so than the natural sciences. We thus drew inspiration from this field, too. Weber, in unequivocal terms, described the concept as 'one of the great tools of all scientific knowledge' ([31], p. 151), highlighting their role in allowing research to proceed through their capacity 'to establish knowledge of what is essential' ([32], p. 213). He also pointed out that existing concepts can be used as building material for new concepts, which is the case for human-centered XAI in our work: an amalgamation of human-centered design and Explainable AI. Swedberg [33] described the process of creating new concepts as an essential part of building theory in science. Observation, according to Swedberg, leads to naming of a phenomenon, which then 'often needs to be turned into a concept . . . to a get a firm grip on the phenomenon' ([33], pp. 58–59).

The empirical data we present were drawn from practice-based research activities over the period 2018–2021 at the Norwegian University of Science and Technology (NTNU). Lacking the space to include everything, we selected examples that we think helped to formulate the concept of human-centered XAI for ASVs in the framework of model-based reasoning. While the research we presented stemmed from a range of projects, one is featured predominantly: the design and construction of an autonomous passenger ferry called the milliAmpere2, which is described in Section 2.1.

2.1. The milliAmpere2 Autonomous Passenger Ferry

Up until the mid-1960s, there was a century-old service available in central Trondheim for transporting passengers across a 100-m-long urban canal (Figure 2). Locally called the Fløttmann, this canal-crossing service was administered by a person in a rowboat (Figure 3a). Today the service is available in the summer months as a tourist attraction. In 2018 the local municipality proposed to construct a pedestrian bridge at the location where the Fløttmann crosses, meeting resistance among stakeholders. Out of these discussions emerged the idea of an automatic passenger ferry, put forward by an associate professor at NTNU. Under the theme of 'digital transformation', the Autoferry Project was kicked off in the same year [34]. Before long, the first prototype was ready. It was called milliAmpere, named after the first electric ferry in Norway, the Ampere. This prototype continues to be used by students, PhDs, and postdocs as a research platform for testing and development of sensor fusion, powering, maneuvering [35], safety, cyber-security, and automated collision avoidance [36]. Meanwhile, work started in 2019 to design an operational version of the automated ferry, capable of carrying up to twelve passengers for public use at the same location as the Fløttmann. As of late-2021, the milliAmpere2 (Figure 3b) has been commissioned and is undergoing field testing.

From a design perspective, the milliAmpere2 offers a unique opportunity to investigate human-centered XAI because it poses the hypothetical question: which would you choose given a choice of the human-operated Fløttmann and the autonomous milliAmpere2? As we approach the operational stage of milliAmpere2, this question will soon represent a real choice. The question motivates our research question: can a human-centered approach to XAI contribute towards building trust among real-world ASV users?

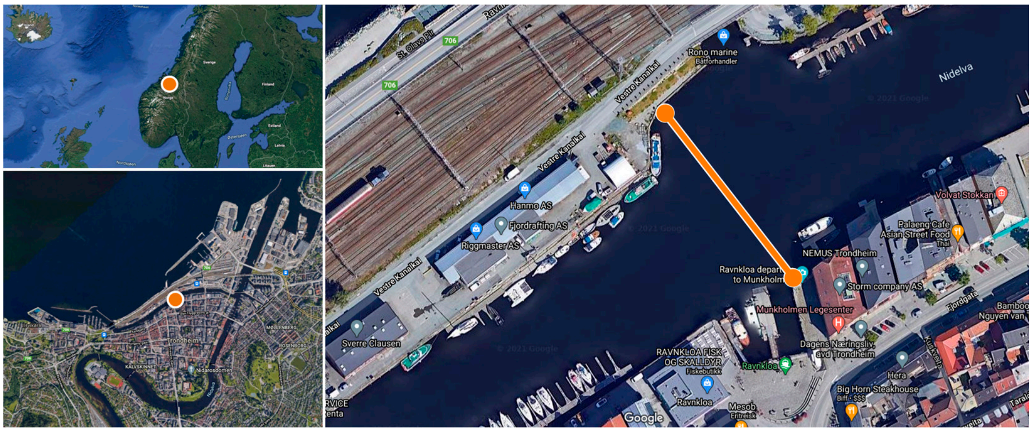
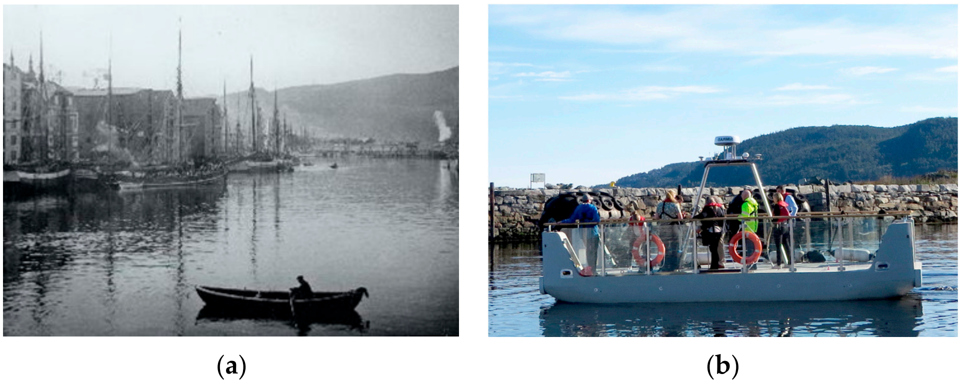


Figure 2. The milliAmpere2 will operate in Trondheim as a canal-crossing service over a 100-m-long urban canal.

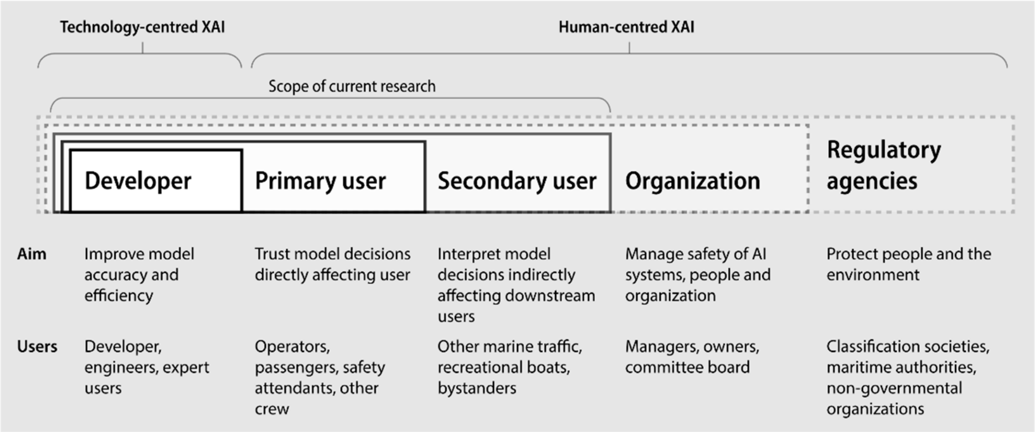


(**a**) (**b**)

Figure 3. The service of crossing an urban canal in Trondheim, Norway is undergoing a digital transformation in the Autoferry project [34] (**a**) The Fløttman in 1906 (photo credit A. Holbæk Eriksens Publishers and The Municipal Archives of Trondheim; licensed under a Creative Commons Attribution 2.0 Generic License); (**b**) The autonomous milliAmpere2 in October 2021 (photo Erik A. Veitch).

2.2. XAI Audience and Scope

Figure 4 depicts the XAI audience we considered in this work, tailored specifically to the ASV application case. Figure 4 also illustrates the scope of this work in terms of what segments of the XAI audience we did not consider. Continued efforts are needed to expand the XAI audience towards organizational stakeholders, such as managers and owners, as well as regulatory agencies and non-governmental organizations.



	Developer	Primary user	Secondary user	Organization	Regulatory agencies
Aim	Improve model accuracy and efficiency	Trust model decisions directly affecting user	Interpret model decisions indirectly affecting downstream users	Manage safety of AI systems, people and organization	Protect people and the environment
Users	Developer, engineers, expert users	Operators, passengers, safety attendants, other crew	Other marine traffic, recreational boats, bystanders	Managers, owners, committee board	Classification societies, maritime authorities, non-governmental organizations

Figure 4. Human-centered XAI expands the field's audience towards end user stakeholders.

To illustrate the distinct XAI needs for different user groups, consider a visualization of an object classification algorithm (Figure 5). While a developer needs to know the probability that a given classification of an identified object is correct, a ship operator needs to know practical navigation details (e.g., the ship's name, current speed and heading, and destination). In contrast, a passenger may only need to know whether the ASV has discovered it.

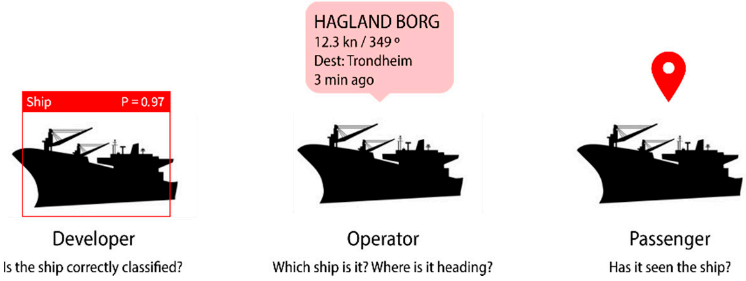


Figure 5. Different user groups have different XAI needs.

2.3. Methodological Considerations

At the outset of our research, we did not intend to create the concept of human-centered XAI for ASVs, nor, once we stumbled upon the idea, did we immediately begin using model-based reasoning to systematically give the idea structure. Model-based reasoning, as a method, was used a posteriori, as a way of lending our reasoning, which occurred 'abductively' [37], a kind of organization that is only possible in hindsight. Thus model-based reasoning has a decidedly historical perspective, adding up to an account of conceptual change grounded in cognitive phenomena. It represents what Nersessian calls the cognitive-historical method: a kind of 'bootstrapping procedure commonly used in science' where the 'range of historical records, notebooks, diaries, correspondence, drafts, publications, and artifacts, such as instruments and physical models, serves as the source of empirical data on the scientific practices' ([30], p. 7). Such a historical perspective may even be a necessary part of concept formation because, as pointed out by Swedberg, 'it is often not possible to create a concept until the research is well underway' ([33], p. 60).

The scientific literature on concept formation is extensive, covering not just the cognitive and social sciences, as we have mentioned, but also the philosophy of science, linguistics, mathematics, and other fields. Defining a concept also raises the question of defining other concepts, as pointed out by Wittgenstein: 'What should we gain from a definition, as it can only lead us to more undefined terms?' ([38], p. 26). Similarly, it is widely recognized that even the word 'concept' has no formal definition. This circular problem is exemplified by the concept of 'mental model,' a useful concept we invoke throughout the paper, but that nonetheless has no agreed-upon definition across the many fields of science using it. This has led to, as expressed by [39], a 'muddling' of the term. Here, we use the definition offered by Nersessian, whose model-based reasoning framework for concept formation we also adopted, and whose definition is agnostic to the myriad fields applying it: 'a mental model is a structural, behavioral, or functional analog representation of a real-world or imagined situation, event, or process' ([30] p. 95).

3. Results

Here we present observations of research activities as they relate to three elements of concept formation for human-centered XAI for ASVs. These three elements are listed below and originate from the model-based reasoning framework for concept formation (Section 2):

1. Analogy (representation of an unfamiliar concept in terms of a familiar one);
2. Visualization (representation of internal states through external imagistic processes);
3. Mental simulation (representation of behavior through 'mental models' and thought experiments, including numerical simulation).

3.1. Analogy

During the early-stage design of the human-machine interface for the milliAmpere2 prototype ferry, it emerged that trust was important for establishing an interaction relationship among passengers. The ASV technology represented a new concept, after all—one that without a human operator broke with convention. The interaction relationship, we reasoned, could be designed into the process of introducing the ferry to passengers, including an explanation of what the ferry was and how it worked. The prototype information post, depicted in Figure 6, contains such an introductory message:

> *'World's first driverless passenger ferry. The service is free and open for everyone. The ferry works like an elevator. You press the Call button, and it calls the ferry. You can take aboard everything from your pets to your bike and stroller. The ferry goes every day from 07:00 to 22:00. The ferry crosses between Ravnkloa and Venstre Kanalkai.'*

The information post in Figure 6 contained the phrase: 'The ferry works like an elevator.' The analogy with an elevator promoted the desired change in passenger trust by enabling understanding one new representation in terms of another, more familiar one. The elevator analogy also served to encode a 'mental model' of how the service works (see Section 2.3 for our definition of 'mental model'). To illustrate this, consider a new passenger mentally simulating pressing a button to call the ferry, understanding that one must wait for its arrival after being called. Then, stepping inside once the doors open, this passenger can press another button inside to close the doors and initiate the crossing. Finally, upon arrival, the doors open on the other side, and they can disembark. An interaction relationship is thus established based on encoding a representation of its use in a mental model and reinforcing that mental model with the expected user interfaces. Buttons for calling the ferry and for initiating it upon entry also reinforce to the user that they are using it correctly and have control.

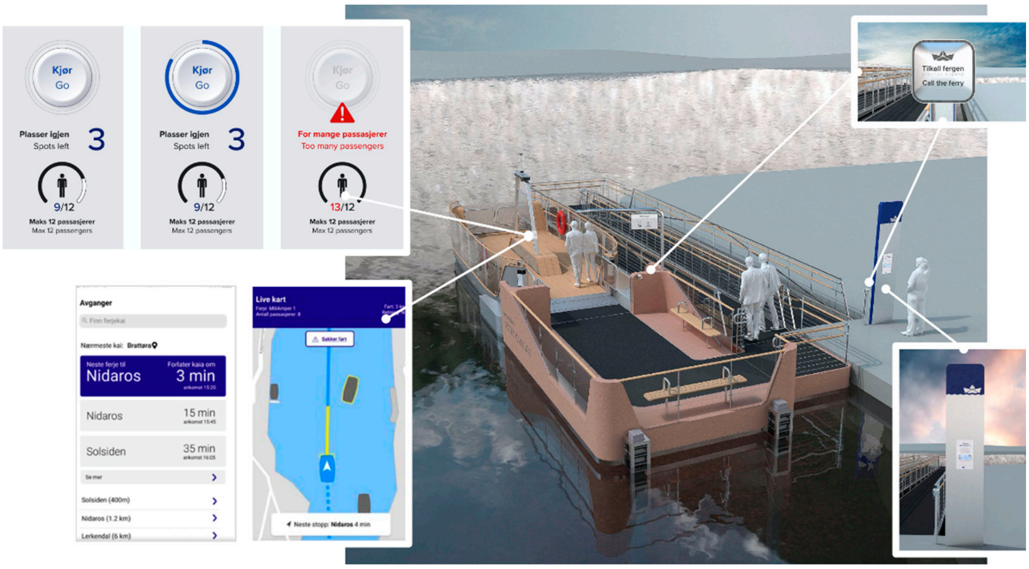


Figure 6. Establishing an interaction relationship with passengers of milliAmpere2.

This example shows that analogy plays an important role in explaining ASV functionality to end users. In the milliAmpere2 example, the analogy worked by transferring the representation of a familiar concept (an elevator) to an unfamiliar concept (ASV technology). In this example, the analogy enhanced the usability of the ASV. Analogy has much in common with mental simulation, a method of concept formation we return to in Section 3.3.

3.2. Visualization

3.2.1. User Displays

As in the previous example where functional representation was transferred to the ASV through an analogy (Section 3.1), so could useability be transferred by affirming the resulting 'mental model' with expected design affordances. In this example, those expected affordances consisted of elevator-like inputs and user displays explaining the internal representation of the ASV. For example, referring to the example in Figure 6, upon pressing the 'Go' button, the 'Closing doors' screen is displayed to the passenger, and the 'Go' button begins indicating the extent to which the ferry has reached its destination using an animated radial dial. If passenger count exceeds twelve then the expectation is met that the ferry cannot depart, and the 'Go' button is greyed out with a warning message that passenger count has been exceeded. The screen also displays an avatar of the ferry along with safety information and the expected time of arrival for different stops. Other objects, including land, coastal infrastructure, and other ships and leisure craft, are depicted on the screen, displaying to passengers what the autonomous system has detected in the environment. Such visualizations serve two purposes: they support usability by affirming the users' mental model of how the ferry works, and they transform internal representations of the ferry ('what is it thinking?') to external imagistic representations.

This example builds the case for human-centered XAI because it shows how visualization on user displays can influence user understanding, interpretability, and trust. In this example, the visual representation of 'what the ferry is thinking' enhanced the ASV's usability. Rather than trying to accomplish this in traditional XAI terms, a distinctly human-centered approach involved the representation of internal states to end users through visual means.

3.2.2. Design, Form and Aesthetic

During the early-stage design of the milliAmpere2, several data collection efforts were launched to gain insights into how the ferry design affected users' perception of the new technology. For example, in [40], surveys, interviews, and workshops were used to understand how design, form, and aesthetics conveyed human-centered values, such as 'safety, stability, user-friendliness.' This culminated in the design of a physical model (Figure 7) and eventually the full-scale ferry (Figure 3b). As described in [40], the design portrayed the familiar curvature and materials of a sailboat—its broad, open deck, the elements of a bridge. These visual representations transferred meaning in a similar way to direct analogy (Section 3.1). In [41], researchers interviewed pedestrians and gained insights into how people interacted with the design as a service. The latter found, for example, that 'older users enjoyed teaching one another how to use new technology,' suggesting that actions involved in discovery and learning were highly valued in the technology interaction. This suggested that enhancing reflexive experiences like discovery and learning promoted overall interpretability, explainability, and trust in the ASV.

This example showed that design, form, and aesthetics played an important role in a human-centered approach to XAI because they helped to build trusting interaction relationships between end users and the ASV. The mechanism worked in a similar way to the analogy (Section 3.1) by transferring representations of familiar concepts (sailboats, bridges, interaction with 'new technology') to the ASV representation through visual means.

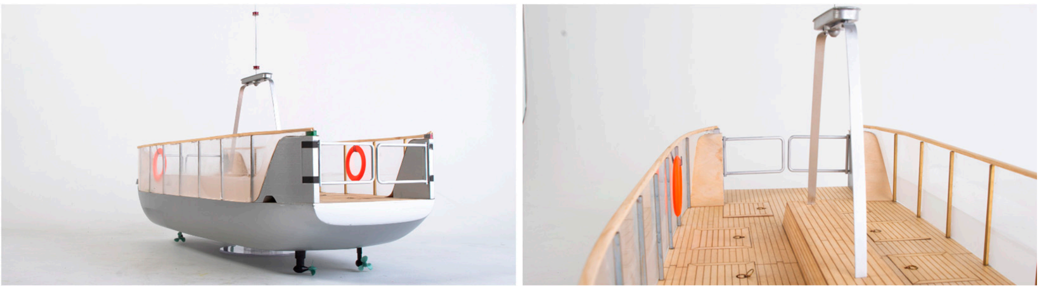


Figure 7. Design, form, and aesthetics played an important role in explaining 'safety, stability, and friendliness' to ASV passengers, according to user research done by [40] (Images depict a 1:10 scale model of milliAmpere2; Reproduced with permission from Petter Mustvedt, published by NTNU Open under Creative Commons CC BY 4.0 License, 2019).

3.2.3. Sensor Data

We observed developers using visualizations of the autonomous navigation systems aboard the milliAmpere prototype ferry with the aim of understanding and improving how the system made decisions. The milliAmpere has been used extensively for testing the sensor fusion algorithms driving the navigation and collision avoidance systems. For example, Figure 8 shows a visualization of sensor data aboard the milliAmpere in the Robotic Operating System (ROS). For an ASV like the milliAmpere to apprehend its environment and minimize the risk of detecting false positives and false negatives affecting motion planning, a vast amount of real-time data is processed 'under the hood.' Synthesizing this data into a single input for motion planning underpins efforts in the field of sensor fusion. Explainability in these efforts represented a major challenge for the milliAmpere given the large volume of data. In Figure 8 we see no fewer than three raw data sources overlayed onto a navigation map: Lidar, radar, and Infrared (IR) video. The visualization also includes real-time data processing, including radar edge-tracing and object tracking and classification.

Such visualizations point to a human-centered XAI process in that they represent the internal state of the ASV ('what it is sensing') in an external imagistic representation. This external representation is used by developers in mental modeling processes to evaluate trust in ASV decision-making in the context of end user interactions.

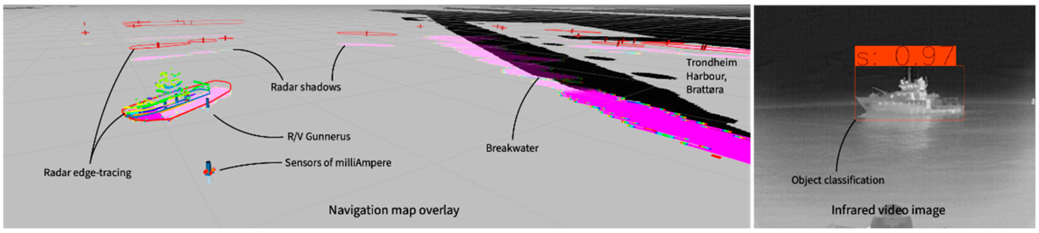


Figure 8. Visualization in Robotic Operating System (ROS) helping a developer of milliAmpere understand sensor fusion (image credit Øystein Helgesen, used with permission).

Another example of XAI visualizations stemmed from ML-based autonomous docking and undocking maneuvers. In [42], several XAI visualizations were created for the developer that display the changing values of azimuth forces and angles (f1, f2, a1, a2) and tunnel thruster forces (f3) of a simulated ship docking. One such visualization, in the form of an action plot (Figure 9), helps the developer to monitor the training status of the algorithm, identify bugs, and improve the autonomous system.

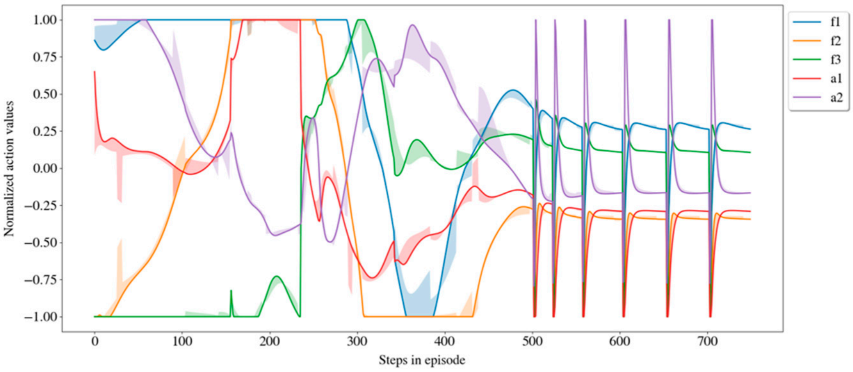


Figure 9. Visualization showing changing values of azimuth thruster forces and angles of a simulated ASV docking. This is meant for the developer (adapted from [42], image used with permission).

Visualizations like the plot in Figure 9 have a limited XAI audience because their meaning is decipherable only to experts of ML-based techniques for the motion control of ASVs. In [42], the same scientists who generated this developer-centric plot recast the information to a user-centric interface (Figure 10), visually conveying which features in the model are weighted most heavily, as well as an avatar showing corresponding thruster power and direction. The display, intended to explain an automated docking process to a navigator, also depicts green lines representing target destination and red lines representing the distance to the nearest object.

This example shows how visualization of AI models can enhance trust in AI decision making among end users in a real-world context. This example, while starting with specialized visualizations intended for developers to improve model interpretability, evolved towards broader user-centered interpretability when confronted with user interaction. This evolution supported the case that human-centered XAI processes were present during the design of ASVs that involved end user interactions.

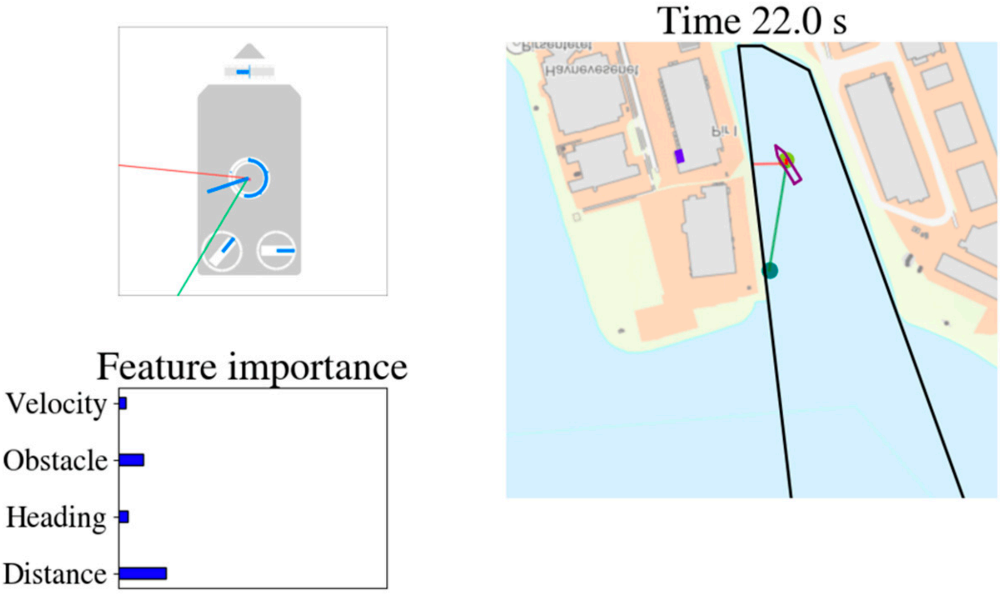


Figure 10. Visualization showing the same information presented in Figure 9 but intended for a ship operator (adapted from [42], image used with permission).

3.2.4. Data Visualization for Shore Control Center Operators

Central for the milliAmpere2, with its goal of operational functionality, was remote monitoring and control supervision. These aims were at the core of the Shore Control Lab [43,44] (Figure 11), a research platform developed in parallel with the milliAmpere2 project. The Shore Control Lab is a Shore Control Centre equipped for research activities and designed with active engagement from various stakeholders and expert groups. Designed to be able to support a fleet of ASVs from one location, it will start with the milliAmpere2, supporting operation of just the one vessel. Having freed up the attention that would normally have been allocated to mundane control tasks on the ferry, the capacity to apprehend problematic out-of-the-ordinary events is enhanced at the Shore Control Center (e.g., handling emergencies, rescues, and special weather events). Operators are also the first line of support to the milliAmpere2 via an onboard video and audio link activated by a call button. Of central importance to Shore Control Center work is the presence of a 'control threshold' that demarcates where AI control ends and where human control begins (for takeovers), and vice versa (for handover). Understanding where the control threshold lies relative to the AI system limitations defines the operators' primary role because it defines when they need to intervene. This control threshold also depends on contextual and local factors (it may, for example, be lowered in adverse weather, resulting in more frequent interventions, and raised after AI system improvements, resulting in fewer).

This example points to an underlying human-centered XAI process whereby interpretability of the AI system is necessitated by virtue of the operators' being 'in the loop' of ASV operations. Because there is a level of collaboration between the ASV system and Shore Control Centre operator, the success of this teamwork hinges on the explainability of the system's constraints and of relevant safety-critical data. Human-centered XAI, in this aim, continues to play a role in designing the Shore Control Centre to enhance human decision-making in contextually nuanced, out-of-the-ordinary events that challenge the constraints of the ASV system.

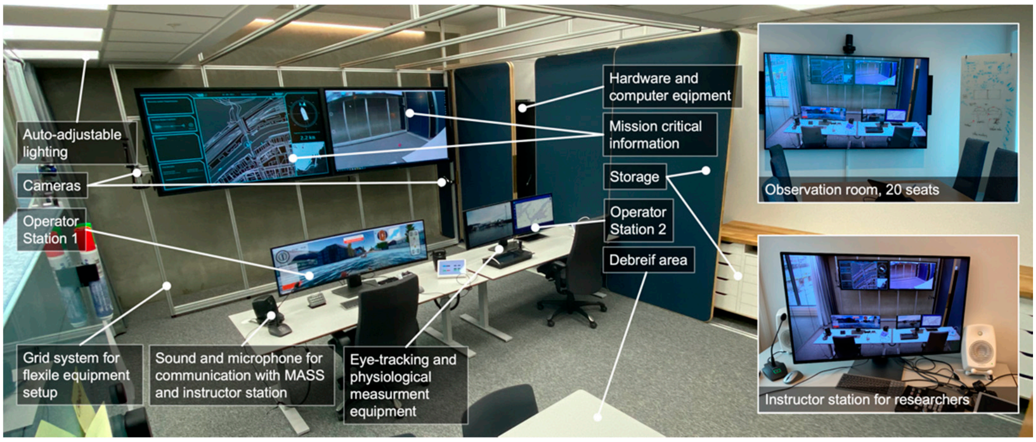

Auto-adjustable lighting
Cameras
Operator Station 1
Grid system for flexile equipment setup
Sound and microphone for communication with MASS and instructor station
Eye-tracking and physiological measurment equipment
Hardware and computer eqipment
Mission critical information
Storage
Operator Station 2
Debreif area
Observation room, 20 seats
Instructor station for researchers

Figure 11. The NTNU Shore Control Lab is designed to enhance human perception and decision-making for ASV operation in out-of-the-ordinary situation handling (see [43,44] for details).

3.2.5. Visual Signals to Predict Future States

Visual signals can be useful in explaining the behavior of ASVs and enhancing traffic coordination and usability. Figure 12 depicts a 1:10 scaled model of milliAmpere2 mounted with light-emitting diodes (LEDs) programmed to light up in different colors, intensities, and light patterns. Early-stage designs of the milliAmpere2 involved testing how light signals explained the ASV's future state to observers. Preliminary experimental results showed that light signals improved understanding and predictability of a diverse range of states, including docking, crossing, vessel detection, autonomous or manual mode, direction changes, speed reductions, emergency stops, and distress signals.

In the broader aim of enabling understandability, interpretability, and trust among ASV end users, the use of visual signals showed how the prediction of future states could be represented as 'mental models' of behavior. Aside from enhancing general usability, this is especially important for other vessels seeking to coordinate traffic in confined waterways.

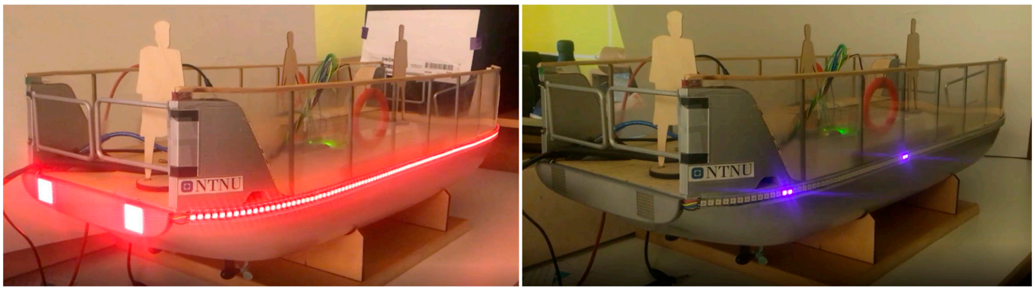


Figure 12. Visual signals used to communicate the milliAmpere2′s state and intention to secondary users, such as other vessels and bystanders (**Left**: distress signal; **right**: vessel detection; image credit Jonas Selvikvåg, used with permission).

3.3. Mental Simulation

3.3.1. Path Planning

Humans make predictions about certain entities using simulated enactments in a type of 'mental model' representing that entity's behavior. To explore how this relates to processes in human-centered XAI, we presented an illustrated example of a collision avoidance maneuver in Figure 13. In this example, different motion paths of a target vessel represent different behaviors for an observing vessel, affecting the observer's predictions of the target vessel's planned path. In Figure 13, the milliAmpere2 (target) is the stand-on vessel, and the R/V Gunnerus (observer) is the give-way vessel, and according to Rule 8 in the COLREGS must milliAmpere2 give 'ample time' to initiate a collision avoidance maneuver. It is assumed that the crew onboard Gunnerus is carefully watching for the signs of such a maneuver. The observing vessel is, in other words, mentally simulating the future path of the milliAmpere2 based on interpretation of its current motion characteristics. In Figure 13 one could imagine that different parameters programmed in the onboard milliAmpere2 navigation system would result in different turning radii around the stand-on vessel. Furthermore, one parameter might appropriately explain that 'ample time' is being accounted for in collision avoidance initiation, whereas another may not.

This is an example of human-centered XAI because it links qualitative properties (in this case, 'ample time' as expressed in COLREGS) to computational decision-making (control parameters to execute a collision avoidance maneuver) that characterize the AVS's behavior. In doing so, the AI system embodies interpretable motion behavior whose actions are simulated in the minds of observers as a type of 'mental model.' Making motion interpretable in this way is important in shared waterways where normal traffic is expected to interact with ASV traffic, especially for collision avoidance in constrained waterways.

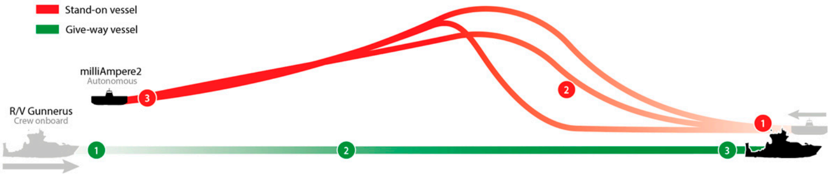


Figure 13. Simulations of various turning radii of milliAmpere2 around a stand-on vessel to evaluate the 'ample time' requirement in COLREGS Rule 8.

3.3.2. Trustworthiness

Apprehending subtle cues in how the milliAmpere2 moves, passengers will form subjective notions of trustworthiness of the automated system's decision-making capacity. Here we assume that the passengers are on board the ASV and therefore feel its motions in a kinesthetic sense. To illustrate this, Figure 14 compares two motion paths of the ASV Telemetron developed in the AutoSea Project [45]. While the first motion path accomplishes the task of collision avoidance, it is characterized as jagged and disjointed (Figure 14, left) when compared to that of a human operator or a better algorithm (Figure 14, right). Were the milliAmpere2 to navigate in this way, it is likely that passengers on board would conflate its motion with unpredictability, thus undermining trust despite what was objectively a successful collision avoidance.

This example shows that human-centered XAI for ASVs involves interpretating system trustworthiness through its motion. This is a type of 'mental model' that represents the inner workings of the navigation system as a kinesthetic experience. The process of aligning ASV motion to kinesthetic experience suggests a decidedly human-centered approach to XAI in the broader aim of aligning ASV actions with human-centered values like predictability, comfort, and trustworthiness.

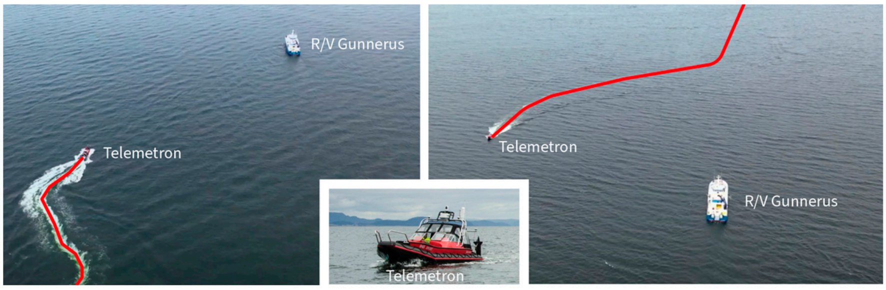
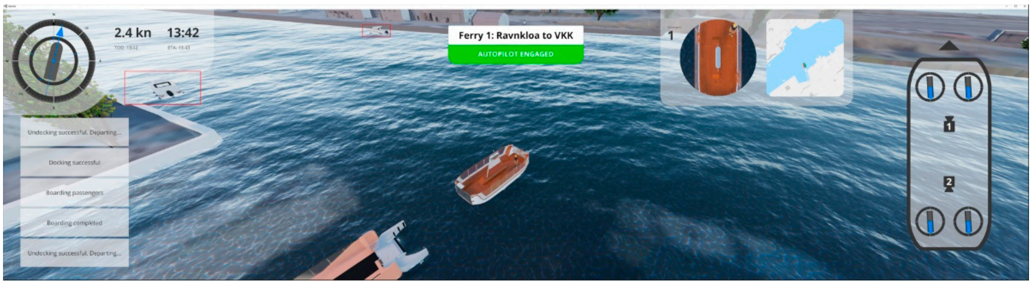


Figure 14. Collision avoidance of the ASV Telemetron: (**Left**) The vessel avoids collision, but the path is jagged, disjointed, and unpredictable. This undermines the trust of the crew, operators, passengers, and other vessels in the environment. (**Right**) Telemetron avoids collision with a smooth and predictable path. Crew, operators, passengers, and other vessels trust the autonomous vessel.

3.3.3. Human-AI Collaboration

Shared control between the ASV and remote operator can be simulated in a virtual environment. In Section 3.2.4, we introduced the Shore Control Centre as a safety-critical element in the overall ASV system because it ensured that humans were 'in the loop' and able to take over control when needed. To design the collaborative system and train operators in its use, waiting for such an intervention situation to arise was not practical. Instead, virtual simulations at the Shore Control Lab were used to reenact them (Figure 15). The virtual simulator is custom-built on top of the Gemini open-source platform [46] and, in effect, lets the researcher carry out thought experiments about ASVs in an immersive environment. For example, by having operators run simulated operations of the milliAmpere2 (as depicted in Figure 15), researchers can investigate what levels of operator involvement in control induce sufficient readiness for safety-critical interventions.

This example showed that virtual simulation can be used to design for human-AI collaboration for ASVs. Virtual simulation is, in this sense, a type of immersive thought experiment whose aim is to represent real-world operations where collaborative control is required to ensure safety. This type of simulation points to a human-centered XAI approach aiming to align the ASV system with real-world operational conditions and expectations of safety.



Undocking successful. Departing...

Docking successful

Boarding passengers

Boarding completed

Undocking successful. Departing...

Figure 15. The virtual simulator based on Gemini open-source platform [46] allowed for immersive thought experimentation of operational design of milliAmpere2.

3.4. Summary

In Table 1 we summarize the Findings, tracing the various processes of human-centered XAI in the examples above.

Table 1. Summary of human-centered XAI processes present in the examples presented.

	XAI Audience			
Model-Based Reasoning Elements	***Developers***	***Passengers***	***Operators***	***Other Vessels***
Analogy	Represents ASV as a familiar concept to explain AI interaction (enhances usability)			
Visualization	Represents ASV sensory perception and numerical models visually to explain AI decision-making (enhances trust)	Represents ASV functionality and affordances through user displays, design, form, and aesthetics to explain AI interaction (enhances usability and trust)	Represents ASV system constraints visually to explain human-AI collaboration (enhances usability and safety)	Represents ASV behavior through visual signals to explain AI decision-making (enhances safety)
Mental simulation	Represents ASV behavior in terms of 'mental models' (often synthesized as numerical simulations) to explain AI decision-making (enhances trust)	Represents ASV behavior using 'mental model' of kinesthetic experience to explain AI decision-making (enhances trust)	Represents ASV behavior in terms of 'mental models' (often synthesized as immersive virtual simulations) to explain human-AI collaboration (enhances safety)	Represents ASV behavior in terms of 'mental models' of motion characteristics to explain AI decision-making (enhances safety)

4. Discussion

In this section we expand upon the findings, examining how the concept of human-centered XAI can be applied to address end user trust in ASVs and contribute to the broader multi-disciplinary discussions around XAI. To help frame this discussion, we start with a definition of human-centered XAI as it applies to ASVs (see Box 1). We also present a table listing human-centered XAI processes compared to corresponding technology-centered XAI processes (Table 2).

Box 1. Human-centered Explainable AI defined.

> Human-centered Explainable AI is the process of making AI technology understandable, interpretable, explainable, and trustworthy for end user interactions. For ASVs, this involves representing capabilities and constraints of the AI system in line with developers', primary users', and secondary users' interaction needs.

Table 2. Technology-centered XAI compared to human-centered XAI.

	Technology-Centered XAI	Human-Centered XAI
1.	Mission is to ensure and improve model accuracy and efficiency	Mission is to establish and maintain an interaction relationship
2.	Provides data-driven visualizations of models based in mathematics	Uses mental models based on analogy-making to explain technology use
3.	Frames increases in machine autonomy as subsequent reductions in human autonomy	Frames increases in machine autonomy as independent of human autonomy
4.	Considers the 'black box' interpretability problem as a barrier to AI development	Accepts that user does not need to know inner working of AI to use it successfully
5.	Considers humans as a source of error in overall system safety	Considers humans as a source of resilience in overall system safety
6.	Often assigns AI human-like qualities, leading to over-selling and not meeting user expectations	Explains limitations to users and actively manages expectations
7.	Seeks to improve the performance of human task by prioritizing computation	Seeks to enhance the performance of human tasks by prioritizing collaboration

4.1. Explainability Needs for Different End User Interactions

One of the features of human-centered XAI is its ability to orient the explaining to the specific needs of the end user. To illustrate this point, consider that AI developers work primarily in a simulated environment without any direct risks for crew or equipment. They also have plenty of time and several tools at their disposal with which to present and analyze data visualizations of the AI model (e.g., Figures 8 and 9). If the algorithm fails, they will not intervene but use it to learn and to improve the model. Operators, on the other hand, either onboard the vessel or in a Shore Control Center, are faced with direct risks and responsibility for crew, vessel, and equipment. They do not have time to analyze complex data visualizations but need simpler visualizations mapped to the ship's layout to be able to quickly decide whether to intervene (Figure 10).

In Section 2, three end user groups were identified for investigation in this research:

1. Developers (researchers, engineers);
2. Primary users (passengers, operators);
3. Secondary users (other vessels, bystanders).

In Table 3 we characterize the various XAI needs of these different groups. A similar effort was presented in [42], where the authors analyzed the differences between developers and vessel operators in terms of characteristics, the context of use, and XAI needs. Here, we expand the analysis to a broader range of end users, such as passengers and other vessels. Note that in Table 3 we refer to Automatic Information Service (AIS), which is a globally standardized information-sharing platform for marine traffic, broadcasting information like ship name, flag, location, and speed (https://www.marinetraffic.com/, accessed on 22 October 2021). Vessel Traffic Services (VTS) refer to land-based coordination centers for aiding mariners in safe and efficient use of navigable waterways [47].

Table 3. Characteristics of developers and primary/secondary users of ASVs in the context of XAI needs.

Characteristics	Developers	Primary Users		Secondary Users	
Users	Developers, engineers, expert users	Operators, safety attendants, crew	Passengers	Sailboats, leisure boats, kayaks (not AIS-equipped); bystanders	Fishing boats, ferries, cruise ships, other marine traffic (AIS-equipped)
Background knowledge	Trained developers; strong technical and analytical skills	Trained mariners, strong safety culture	No or little understanding of navigation nor of onboard safety systems	Varying navigation skills and safety culture	Trained mariners; strong safety culture
Context of use	Primarily office work; methodical testing of simulation-based scenarios; indirect consequences for crew and environment	Primarily field work; time-critical decision-making; direct consequences for crew and environment	Present onboard the vessel with the purpose of safe, comfortable, enjoyable, and timely transportation	Shared traffic especially in confined waterways (especially during holidays); unreliable radio communication and VTS detection; bystanders may become future passengers	Shared traffic especially in marine traffic lanes; reliable radio communication and VTS detection
Interaction with AI	Improve ASV model accuracy and efficiency; train models with data	Directly affected by ASV model decisions	Directly affected by ASV model decisions	Indirectly affected by the ASV state and intention	Indirectly affected by the ASV state and intention
XAI needs	Visualizations of ASV models; real-time visualization and processing of sensor data; description of training data	Current state and intention of ASV models; definition of AI-human control boundary; understanding of when to intervene	Confirmation that ASV 'sees' and avoids collisions with other objects	Confirmation that ASV 'sees' them to avoid collisions, dangerous situations, and 'deadlocks'	Clear information about ASV intentions for avoiding collisions and 'deadlocks;' traffic flow maintained

4.2. XAI to Establish Interaction Relationships and Build User Trust

Interacting with an ASV involves interpretability, understandability, explainability, and trust—phenomena that are also at the core of the XAI field. While developers focus on explaining AI techniques using mathematics and data visualization, a generalized approach based on leveraging more universal interaction elements, such as analogy-making, design, aesthetics, form, and motion characteristics, was used to reach a wider audience of end users in the case studies we examined. These elements were particularly important in establishing an interactive relationship with new end users during their first encounter with an ASV.

Explaining AI can be done by comparing it to familiar, 'human-friendly,' concepts. Recently, the use of 'human-friendly' concepts has seen increasing use among developers testing the interpretability of deep learning models [48]. Similarly, we observed that explaining ASV functionality to potential passengers played a central role in trust (Section 3.1), suggesting that explaining using familiar concepts plays an important role in human-centered XAI. For the milliAmpere2, for example, we observed the use of an information post containing a useful analogy that helped relate the abstract concept of ASV technology into something more concrete: an elevator (we return to this theme in Section 4.3 in the context of using metaphors). Using familiar concepts may also be useful for explaining safety. Typically, in the risk sciences, safety measures are represented using probabilities, expressed, for instance, as a percent likelihood for false positives and false negatives. However, as shown by [49], interpretation of what such error rates mean in practice, especially when related in terms of base rate information (such as total prevalence of all errors) runs counter to intuition. Explaining the ASV's safety to users may be more effectively accomplished by comparing it to an existing baseline; for example, whether it can be shown to be significantly safer than a corresponding manned surface vessel.

Explaining AI can also be done by rending activities visible. In computer vision, saliency-based XAI techniques that visualize the pixels that a network is fixated upon for its predictions have made important strides towards model interpretability [50]. In [51], the authors showed how similar visualization techniques can lead to insights into how image classification models function in intermediate layers, leading to insights about model improvements. A similar discussion about 'where to look' arises for the shore control center operators whose tasks are to apprehend and respond to out-of-the-ordinary events (Section 3.2.4). In the context of XAI for collaborative systems, this raises the need for explaining actions taken during coordinated work. In [52], the authors demonstrated the importance of 'rendering activities visible' in control room work, observing, for example, the central role of out-loud explanations, verbal exclamations, pointing, and other cues and gestures in coordinated action among operators handling a situation. XAI for collaborative AI-human work, such as in the Shore Control Centre, should, in the light of such 'mutual monitoring' practices, avoid the pitfall of rendering activities in AI systems invisible.

The 'rendering actions visible' corollary applies in equal measure to secondary users. These users, as defined by [53], are those who are not interacting with a system directly and yet are affected by the system's and primary users' actions. In the context of autonomous ships, these users are indirectly influenced by the decisions of the AI system and include marine traffic, recreational boats, and bystanders. Motion planning characteristics (Section 3.3.1) and visual signals (Section 3.3.3) were shown to be especially useful for conveying an understanding of ASV state and intention.

4.3. XAI by Encoding Abstract Concepts as Mental Models to Support User Interaction

Humans have the remarkable ability of understanding concepts through abstraction, even when faced with something entirely new and with very little information. In [54], the authors studied the use of metaphors and framed them as a central mechanism by which humans make sense of the world. Metaphors go beyond just natural language, guiding the most fundamental interactions in our daily lives, usually without our even being aware. For example, the authors describe 'orientational metaphors' that lend abstract concepts of spatial orientation relative to what we know best: our bodies. Consciousness, for instance, is up; unconsciousness is down: 'Get *up*. Wake *up*. He *rises* early in the morning. He *fell* asleep. He's *under* hypnosis. He *sank* into a coma' (p. 15, authors' original emphasis). The human-centered approach to XAI leverages humans' ability to understand abstract concepts via metaphors using 'mental models.' There are many definitions for mental models, and we presented a working definition in Section 2.3. A 'mental model' is, in this context, a type of abstraction that encodes in the mind of an end user an explanation of how to use a technology. For the milliAmpere2, we observed that the 'elevator mental model' was especially useful (Section 3.1). In this application, the simplicity of the mental model belied its power in establishing and maintaining a trusting interaction relationship. Simply providing a mental model, though, was not enough to maintain a trusting interaction relationship. Interface inputs were thus designed to affirm the mental model, and outputs were mapped to actions that reinforced it.

4.4. XAI That Frames Human Autonomy Independent of Machine Autonomy

Adopting Level of Automation (LOA) taxonomies, such as of the Norwegian Forum for Autonomous Shipping (NFAS) framework [55], was useful for explaining the extent to which the milliAmpere2 system was automated. For example, the LOA called 'Constrained Autonomy' fits the description of the milliAmpere2:

> 'The ship [or ASV] can operate fully automatic in most situations and has a predefined selection of options for solving commonly encountered problems... It will call on human operators to intervene if the problems cannot be solved within these constraints. The SCC or bridge personnel continuously supervises the operations and will take immediate control when requested to by the system. Otherwise, the system will be expected to operate safely by itself.' [55] (pp. 11–12)

However, [16] showed that conventional taxonomies for LOA imply that introducing machine autonomy comes at the cost of human control: a 'one-dimensional' zero-sum relationship. In a comprehensive review, [56] compared twelve taxonomies for automation, all of which presumed such a one-dimensional LOA, composed of an incremental scale from full human control to full autonomous control. Since their review, several new taxonomies have been proposed specifically for ASVs that adopt a similar one-dimensional LOA format [7,55,57]. Scholars have warned about the loss of human autonomy such taxonomies might unintentionally introduce, proposing instead LOA frameworks that consider joint human and machine autonomy as two-dimensional 'stages of automation' [16]. As explained by [58], examples of highly reliable, automated systems already exist in cars, including Anti-lock Braking Systems (ABS), Traction Control Systems (TCS), and Electronic Stability Control (ESC). The success of these technologies lies in their ability to enhance, rather than supplant human decision-making when braking and maneuvering in safety-critical situations.

Inspired by the 'stages of automation' framework proposed by [16], which considers joint human and machine autonomy in AI systems, we mapped several prominent maritime surface vessels onto four quadrants of AI-human control (Figure 16). The basis for positioning these vessels in Figure 16 is described in Table 4, which characterizes their respective levels of human and machine control. The results show a trend moving from the upper left quadrant (high human control; low machine control) toward the bottom right quadrant (high machine control; low human control). A necessary adjustment for achieving trust in ASVs is to shift toward the upper right quadrant (high machine control; high human control).

4.5. XAI That Accepts That Users Do Not Need to Know Inner Workings of AI to Use It Successfully

To illustrate the motivation behind this guiding principle, we can represent 'black box' AI techniques with a technology we are familiar with: a dishwasher. The dishwasher user has control over several inputs in the form of loading dishes and pressing buttons to control the setting and start the machine. When the machine is activated, it runs automatically and, after a time, produces clean dishes. While it is possible to implement a transparent door on the dishwasher so that users could see what is happening, it is unlikely that this level of transparency will contribute to an enhanced user experience. We are, in this case, simply projecting the value of transparency onto the user, possibly at the risk of a less efficient and more expensive product. This may be what [22] observed when finding that an ML algorithm's end user trust levels were not affected by model transparency. Similarly, we observed in the milliAmpere2 user research that transparency of analogous 'black box' AI techniques may be less of an influencing factor than design, aesthetics, form, and motion characteristics for building trust (Section 3.2.2).

Product developers have known for a long time that predicting user experiences is notoriously difficult [63]. Part of the reason is that user experience is entangled with higher-order values and ambitions of which the user themselves may be only dimly aware. In the dishwasher example, these ambitions could be related to the higher-order value of preserving harmony in the household rather than the more obvious benefit of skipping kitchen chores. In the milliAmpere2 case, high-order ambitions are more likely linked to discovery and learning (Section 3.2.2) than to simply crossing a canal. Since the early days of AI development, scholars have discussed the dissonance that can result from automated machine decision-making in the context of societal values. More than half a century ago, Norbert Wiener wrote that 'should we create a mechanical agency with whose operation we cannot efficiently interfere once we have started it . . . then we had better be quite sure that the purpose put into the machine is the purpose which we really desire and not merely a colorful imitation of it' [64] (p. 1358). The issue remains elusive today and has been recast into a discussion about 'value alignment in AI' [23]. A new field of AI ethics has also emerged, where scholars cast questions about morals in decision-making, previously considered theoretical, into a new practical light [65,66]. A human-centered approach to XAI does not assume that making AI decisions transparent will necessarily benefit an interaction relationship. Instead, it eschews explanations that risk projecting the values of the developers onto users, steering the user instead towards reflexive and subjective experiences.

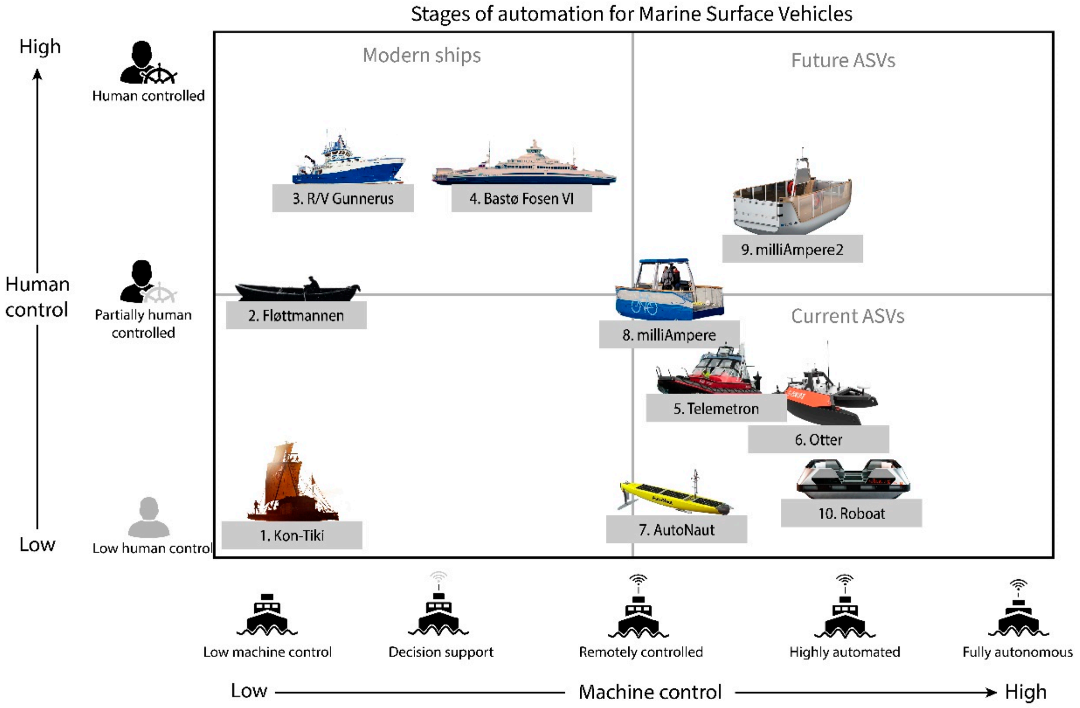


Figure 16. Two-dimensional 'stages of automation' plot adapted from [16] showing human-AI system control for various vessels. These are described further in Table 4.

Table 4. Stages of automation described for the example shown in Figure 16.

ID	Vessel/Project	Description	Autonomous Control	Human Control
1.	Kon-Tiki [59]	Balsa wood raft used in Thor Heyerdahl's 1947 expedition in the South Pacific Ocean	None	Very low; poorly maneuverable raft relying upon wind and current
2.	Fløttmannen	Passenger rowboat crossing a 100 m canal in Trondheim (same location where the milliAmpere ferries are planned for operation)	None	Medium; maneuverable in urban canals; visible actions for passengers and other vessels
3.	R/V Gunnerus [60]	Research vessel owned and operated by NTNU	Low; Dynamic Positioning (DP) system for automatic station-keeping	High; navigation wheelhouse with crew and complement of four

Table 4. *Cont.*

ID	Vessel/Project	Description	Autonomous Control	Human Control
4.	Bastø Fosen VI [12]	Roll-on-roll-off car ferry (length overall 140 m) operating on Oslofjord's Horten-Moss crossing, the busiest crossing in Norway	Medium; Auto-crossing and auto-docking; DP system	High; navigation wheelhouse with two deck officers, total crew of five
5.	Telemetron [45]	Converted sport vessel with autonomous functionality; used for field testing of collision avoidance	Automated navigation including collision avoidance	Original steering and controls onboard; vehicle control station onboard for control of autonomous system
6.	Otter [61]	Portable research and data-acquisition ASV power by battery packs	Medium; autonomously follows motion trajectory defined by user	Medium; remote control with joystick; set trajectories via user-friendly software interface and mobile app
7.	AutoNaut [62]	Long-range research and data-acquisition ASV propelled by ocean waves	Medium; when there are no waves it drifts with current	Low; remote control of passive maneuvering system; waypoint navigation in custom user interface
8.	milliAmpere [34]	Electric autonomous passenger ferry (length overall 5 m) crossing a 100 m canal in Trondheim; designed for research purposes only	Medium; Automated crossing, docking, undocking, and collision avoidance	Medium; remote control console, emergency stop, oars
9.	milliAmpere2 [34]	Electric autonomous passenger ferry (length overall 8 m) crossing a 100 m canal in Trondheim; designed for real-world operation and field research	Automated crossing, docking, undocking, and collision avoidance; Dynamic Positioning (DP)	Medium; collaborative control with Shore Control Centre; direct remote motion control via DP system; passenger user interface and communication link
10.	Roboat [4]	On-demand multipurpose autonomous platforms (floating bridges and stages, waste collection, goods deliver, urban transportation, data collection)	High; automated motion planning, obstacle avoidance, predictive trajectory tracking	Low; indirect control only, no direct remote motion control or collaborative control infrastructure

For operators in the Shore Control Centre who had specific roles and safety responsibilities, this approach led to designing interactions that enhanced experts' ability to apprehend out-of-the-ordinary events the AI system is unable to handle (Section 3.2.4). As [67] observed from experts' prediction making when compared to simple regression models, the experts, despite not producing as accurate predictions as the models, demonstrated an ability to 'know what to look for' in making those predictions [67] (p. 573). In other words, experience has taught experts to identify which factors are linked to future events if not able to accurately synthesize that information into an accurate prediction. Well-aligned collaboration of AI-human systems, such as Shore Control Centers, should, in this light, combine the synthesizing ability of computation with experts' semantic knowledge of which factors are meaningful to decision-making. For such collaborative systems, understanding the respective roles of the AI system and its human expert backup underpinned this emerging 'hybrid' format of collaborative work. More work is needed in this area, and the field of XAI may contribute to techniques and strategies for rendering collaborative work more visible to operators.

5. Conclusions

As more ASVs are deployed, and AI-based navigation systems become more widespread, trust is emerging as a barrier to acceptance. This is especially the case as we move towards passenger ASVs, where unintended consequences of model errors can result in real harm to people. The tenants of the field of Explainable AI (XAI)—with its emphasis on interpretability, understandability, explainability, and trust—represent common values for multiple disciplines interested in addressing the challenge of building trust in ASVs. Traditionally, XAI has been composed of computer scientists with the aim of improving model efficiency and accuracy by better interpretation of opaque 'black box' Machine Learning (ML) models. However, as downstream consequences of AI deployment in real-world maritime applications become clearer, more disciplines are contributing to the field. This includes human-computer interaction and interaction design, two fields positioned at the crossroads of humans and technology. In this article, we constructed the concept of 'human-centered XAI' to address the distinct needs of end user interactions for ASVs. We did this by drawing from examples of ongoing research that pointed to processes of analogy, visualization, and mental simulations that, according to model-based reasoning [30], are signatures of formation for new concepts. Our investigation was guided by the research question: can a human-centered approach to XAI contribute towards building trust among real-world ASV users?

The findings showed that the main aims of human-centered XAI processes were to enhance (1) usability, (2) trust, and (3) safety. Usability was enhanced through the representation of AI functionality; trust was enhanced through the representation of AI decision-making, sensory perception, and behavior; and safety was enhanced through the representation of AI constraints and human-AI collaboration. All representations were enabled by interrelated processes of analogy making, visualization, and mental simulation that sought to align AI goals and actions with the values held by end users during interaction with ASVs.

Several important limitations exist. Firstly, given that the nature of concepts is to lack permanence, the processes that we define as 'human-centered XAI' may change over time, across cultures, and with shifting societal values. Secondly, all the examples presented were drawn from either our own research or from the research of our collaborators at our university. This entailed a high degree of reflexivity in the research process that risked entangling the processes observed with the formation of the concept we constructed.

For the future direction of research, we hope to see lively discussion about how the concept can be taken further in practice-based research. Could a 'Handbook of human-centered XAI' or similar synthesize the processes we underlined for developers, engineers, designers, and other practitioners? Furthermore, can we expand the XAI audience to include organizational stakeholders and even regulatory agencies?

This work has demonstrated that as ASVs scale for widespread deployment, design practices are orienting towards end user interaction design. However, this is occurring without the guidance of a fully formulated conceptualization of what this design practice entails. The main contribution of this work is the formation and definition of human-centered XAI for this purpose, helping to advance ASV design toward wide-scale deployment.

Author Contributions: E.V.: Conceptualization, methodology, investigation, visualization, and writing—original draft preparation. O.A.A.: Supervision, project administration, conceptualization, methodology, visualization, and writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Norwegian Research Council under the MAROFF-2 project 'Land-based Operation of Autonomous Ships' (LOAS) directed by project leader Kongsberg Maritime AS and partnered with the Institute for Energy Technology (IFE), grant number 296527.

Acknowledgments: The authors acknowledge the support of the Department of Design, the Department of Electronic Systems, and the Department of Engineering Cybernetics at NTNU. The authors gratefully acknowledge the support of several former Master's students from the Department of Design who contributed to this work: Petter Mustvedt, Henrik Ellingsen, and Peter Glesaaen. We would also like to thank Erik Scott Hognestad, Maia Sløveren Andressen, and Jonas Klev Selvikvåg (Department of Design) for their contributions to the interaction design visualizations and to doctoral candidates Øystein Helgesen, Emil Thyri, and Vilde Gjærum for their contributions to XAI visualizations and to Thomas Kaland, Kjetil Vasstein, Mikael Hanssen, and Felix Petermann for their contributions to visualizations of the Shore Control Lab. Finally, we would like to thank the whole team behind milliAmpere and milliAmpere2 for supporting this work.

Conflicts of Interest: The authors declare no conflict of interest. The sponsors had no role in the design, execution, interpretation, or writing of the study.

References

1. Dunbabin, M.; Grinham, A.; Udy, J. *An Autonomous Surface Vehicle for Water Quality Monitoring*; Australian Robotics and Automation Association: Sydney, Australia, 2 December 2009; pp. 1–6.
2. Kimball, P.; Bailey, J.; Das, S.; Geyer, R.; Harrison, T.; Kunz, C.; Manganini, K.; Mankoff, K.; Samuelson, K.; Sayre-McCord, T.; et al. The WHOI Jetyak: An Autonomous Surface Vehicle for Oceanographic Research in Shallow or Dangerous Waters. In Proceedings of the 2014 IEEE/OES Autonomous Underwater Vehicles (AUV), Oxford, MS, USA, 6–9 October 2014; pp. 1–7.
3. Williams, G.; Maksym, T.; Wilkinson, J.; Kunz, C.; Murphy, C.; Kimball, P.; Singh, H. Thick and Deformed Antarctic Sea Ice Mapped with Autonomous Underwater Vehicles. *Nat. Geosci.* **2015**, *8*, 61–67. [CrossRef]
4. MiT Roboat Project. Available online: http://www.roboat.org (accessed on 19 November 2020).
5. Reddy, N.P.; Zadeh, M.K.; Thieme, C.A.; Skjetne, R.; Sorensen, A.J.; Aanondsen, S.A.; Breivik, M.; Eide, E. Zero-Emission Autonomous Ferries for Urban Water Transport: Cheaper, Cleaner Alternative to Bridges and Manned Vessels. *IEEE Electrif. Mag.* **2019**, *7*, 32–45. [CrossRef]
6. Wang, J.; Xiao, Y.; Li, T.; Chen, C.P. A Survey of Technologies for Unmanned Merchant Ships. *IEEE Access* **2020**, *8*, 224461–224486. [CrossRef]
7. IMO. *Outcome of the Regulatory Scoping Exercise for the Use of Maritime Autonomous Surface Ships (MASS)*; IMO: London, UK, 2021.
8. Burmeister, H.-C.; Bruhn, W.; Rødseth, Ø.J.; Porathe, T. Autonomous Unmanned Merchant Vessel and Its Contribution towards the E-Navigation Implementation: The MUNIN Perspective. *Int. J. e-Navig. Marit. Econ.* **2014**, *1*, 1–13. [CrossRef]
9. Peeters, G.; Yayla, G.; Catoor, T.; Van Baelen, S.; Afzal, M.R.; Christofakis, C.; Storms, S.; Boonen, R.; Slaets, P. An Inland Shore Control Centre for Monitoring or Controlling Unmanned Inland Cargo Vessels. *J. Mar. Sci. Eng.* **2020**, *8*, 758. [CrossRef]
10. Kongsberg. Kongsberg Maritime and Massterly to Equip and Operate Two Zero-Emission Autonomous Vessels for ASKO. Available online: https://www.kongsberg.com/maritime/about-us/news-and-media/news-archive/2020/zero-emission-autonomous-vessels/ (accessed on 29 September 2021).
11. Rolls-Royce Press Releases. Available online: https://www.rolls-royce.com/media/press-releases.aspx (accessed on 18 April 2021).
12. Kongsberg. First Adaptive Transit on Bastøfosen VI. Available online: https://www.kongsberg.com/maritime/about-us/news-and-media/news-archive/2020/first-adaptive-transit-on-bastofosen-vi/ (accessed on 29 September 2021).
13. Gunning, D.; Aha, D. DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Mag.* **2019**, *40*, 44–58. [CrossRef]
14. Horvitz, E. Principles of Mixed-Initiative User Interfaces. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Pittsburgh, PA, USA, 15–20 May 1999; Association for Computing Machinery: New York, NY, USA; pp. 159–166.
15. Höök, K. Steps to Take before Intelligent User Interfaces Become Real. *Interact. Comput.* **2000**, *12*, 409–426. [CrossRef]
16. Shneiderman, B. Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *Int. J. Hum. Comput. Interact.* **2020**, *36*, 495–504. [CrossRef]
17. Cui, H.; Zhang, H.; Ganger, G.R.; Gibbons, P.B.; Xing, E.P. GeePS: Scalable Deep Learning on Distributed GPUs with a GPU-Specialized Parameter Server. In Proceedings of the Eleventh European Conference on Computer Systems, London, UK, 18–21 April 2016; Association for Computing Machinery: New York, NY, USA.
18. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
19. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
20. Voosen, P. The AI Detectives. *Science* **2017**, *357*, 22–27. [CrossRef]
21. Došilović, F.K.; Brčić, M.; Hlupić, N. Explainable Artificial Intelligence: A Survey. In Proceedings of the 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 21–25 May 2018; pp. 210–215.

22. Poursabzi-Sangdeh, F.; Goldstein, D.G.; Hofman, J.M.; Wortman Vaughan, J.W.; Wallach, H. Manipulating and Measuring Model Interpretability. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 8–13 May 2021; Association for Computing Machinery: New York, NY, USA, 2021.
23. Christian, B. *The Alignment Problem: Machine Learning and Human Values*; WW Norton & Company: New York, NY, USA, 2020.
24. Kleinberg, J.; Lakkaraju, H.; Leskovec, J.; Ludwig, J.; Mullainathan, S. Human Decisions and Machine Predictions. *Q. J. Econ.* **2018**, *133*, 237–293. [CrossRef]
25. Shirado, H.; Christakis, N.A. Locally Noisy Autonomous Agents Improve Global Human Coordination in Network Experiments. *Nature* **2017**, *545*, 370–374. [CrossRef] [PubMed]
26. Hansen, H.L.; Nielsen, D.; Frydenberg, M. Occupational Accidents Aboard Merchant Ships. *Occup. Environ. Med.* **2002**, *59*, 85–91. [CrossRef] [PubMed]
27. Hetherington, C.; Flin, R.; Mearns, K. Safety in Shipping: The Human Element. *J. Saf. Res.* **2006**, *37*, 401–411. [CrossRef]
28. IMO. Maritime Safety. Available online: https://www.imo.org/en/OurWork/Safety/Pages/default.aspx (accessed on 27 April 2021).
29. Goerlandt, F.; Pulsifer, K. An Exploratory Investigation of Public Perceptions towards Autonomous Urban Ferries. *Saf. Sci.* **2022**, *145*, 105496. [CrossRef]
30. Nersessian, N.J. *Creating Scientific Concepts*; MIT Press: Cambridge, MA, USA, 2010.
31. Weber, M. Science as a Vocation. In *From Max Weber*; Gerth, H.H., Mills, C.W., Eds. and Translators; Oxford University Press: New York, NY, USA, 1946; pp. 129–156.
32. Weber, M. *Roscher and Knies: The Logical Problems of Historical Economics*; Oakes, G., Translator; Free Press: New York, NY, USA, 1975.
33. Swedberg, R. *The Art of Social Theory*; Princeton University Press: Princeton, NJ, USA, 2014.
34. NTNU. Autoferry—NTNU. Available online: https://www.ntnu.edu/autoferry (accessed on 1 October 2020).
35. Bitar, G.; Martinsen, A.B.; Lekkas, A.M.; Breivik, M. Trajectory Planning and Control for Automatic Docking of ASVs with Full-Scale Experiments. *IFAC-PapersOnLine* **2020**, *53*, 14488–14494. [CrossRef]
36. Thyri, E.H.; Breivik, M.; Lekkas, A.M. A Path-Velocity Decomposition Approach to Collision Avoidance for Autonomous Passenger Ferries in Confined Waters. *IFAC-PapersOnLine* **2020**, *53*, 14628–14635. [CrossRef]
37. Paavola, S. On the Origin of Ideas: An Abductivist Approach to Discovery. Ph.D. Thesis, University of Helsinki, Helsinki, Finland, 2006.
38. Wittgenstein, L. *The Blue and the Brown Book*; Harper: New York, NY, USA, 1958.
39. Rips, L.J. Mental muddles. In *The Representation of Knowledge and Belief*; Arizona Colloquium in Cognition; The University of Arizona Press: Tucson, AZ, USA, 1986; pp. 258–286.
40. Mustvedt, P. Autonom Ferge Designet for å Frakte 12 Passasjerer Trygt over Nidelven. Master's Thesis, Norwegian University of Science and Technology (NTNU), Trondheim, Norway, 2019.
41. Glesaaen, P.K.; Ellingsen, H.M. Design av Brukerreise og Brygger til Autonom Passasjerferge. Master's Thesis, Norwegian University of Science and Technology (NTNU), Trondheim, Norway, 2020.
42. Gjærum, V.B.; Strümke, I.; Alsos, O.A.; Lekkas, A.M. Explaining a Deep Reinforcement Learning Docking Agent Using Linear Model Trees with User Adapted Visualization. *J. Mar. Sci. Eng.* **2021**, *9*, 1178. [CrossRef]
43. NTNU. NTNU Shore Control Lab. Available online: https://www.ntnu.edu/shorecontrol (accessed on 30 September 2021).
44. Veitch, E.A.; Kaland, T.; Alsos, O.A. Design for Resilient Human-System Interaction in Autonomy: The Case of a Shore Control Centre for Unmanned Ships. *Proc. Des. Soc.* **2021**, *1*, 1023–1032. [CrossRef]
45. Brekke, E.F.; Wilthil, E.F.; Eriksen, B.-O.H.; Kufoalor, D.K.M.; Helgesen, Ø.K.; Hagen, I.B.; Breivik, M.; Johansen, T.A. The Autosea Project: Developing Closed-Loop Target Tracking and Collision Avoidance Systems. *J. Phys. Conf. Ser.* **2019**, *1357*, 012020. [CrossRef]
46. Vasstein, K.; Brekke, E.F.; Mester, R.; Eide, E. Autoferry Gemini: A Real-Time Simulation Platform for Electromagnetic Radiation Sensors on Autonomous Ships. *IOP Conf. Ser. Mater. Sci. Eng.* **2020**, *929*, 012032. [CrossRef]
47. VTS Manual 2021-Edition 8. IALA: Zeebrugge, Belgium. Available online: https://www.iala-aism.org/product/iala-vts-manual-2021/ (accessed on 5 November 2021).
48. Kim, B.; Wattenberg, M.; Gilmer, J.; Cai, C.; Wexler, J.; Viegas, F.; Sayres, R. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018*; ICML: San Diego, CA, USA.
49. Kahneman, D.; Tversky, A. On the Psychology of Prediction. *Psychol. Rev.* **1973**, *80*, 237–251. [CrossRef]
50. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks. *Nature* **2017**, *542*, 115–118. [CrossRef] [PubMed]
51. Zeiler, M.D.; Fergus, R. Visualizing and Understanding Convolutional Networks. In Proceedings of the Computer Vision—ECCV 2014; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 818–833. Available online: https://arxiv.org/abs/1311.2901 (accessed on 3 November 2021).
52. Heath, C.; Luff, P. Collaboration and Control: Crisis Management and Multimedia Technology in London Underground Line Control Rooms. *Comput. Supported Coop. Work.* **1992**, *1*, 69–94. [CrossRef]

53. Alsos, O.A.; Das, A.; Svanæs, D. Mobile Health IT: The Effect of User Interface and Form Factor on Doctor–Patient Communication. *Int. J. Med. Inform.* **2012**, *81*, 12–28. [CrossRef]
54. Lakoff, G.; Johnson, M. *Metaphors We Live by*; University of Chicago Press: London, UK, 2003.
55. Rødseth, Ø.J. *Definitions for Autonomous Merchant Ships*; NFAS: Trondheim, Norway, 2017.
56. Vagia, M.; Rødseth, Ø.J. A Taxonomy for Autonomous Vehicles for Different Transportation Modes. *J. Phys. Conf. Ser.* **2019**, *1357*, 012022. [CrossRef]
57. Utne, I.B.; Sørensen, A.J.; Schjølberg, I. Risk Management of Autonomous Marine Systems and Operations. In Proceedings of the ASME 2017 36th International Conference on Ocean, Offshore and Arctic Engineering, Trondheim, Norway, 25–30 June 2017; Volume 3B: Structures, Safety and Reliability. ASME: New York, NY, USA, 2017.
58. Stone, P.; Brooks, R.; Brynjolfsson, E.; Calo, R.; Etzioni, O.; Hager, G.; Hirschberg, J.; Kalyanakrishnan, S.; Kamar, E.; Kraus, S. *Artificial Intelligence and Life in 2030: The One Hundred Year Study on Artificial Intelligence*; Stanford University: Stanford, CA, USA, 2016.
59. Heyerdahl, T. The Voyage of the Raft Kon-Tiki. *Geogr. J.* **1950**, *115*, 20–41. [CrossRef]
60. Skjetne, R.; Sørensen, M.E.N.; Breivik, M.; Værnø, S.A.T.; Brodtkorb, A.H.; Sørensen, A.J.; Kjerstad, Ø.K.; Calabrò, V.; Vinje, B.O. *AMOS DP Research Cruise 2016: Academic Full-Scale Testing of Experimental Dynamic Positioning Control Algorithms Onboard R/V Gunnerus*; Volume 1: Offshore Technology; ASME: Trondheim, Norway, 2017.
61. Maritime Robotics Otter. Available online: https://www.maritimerobotics.com/otter (accessed on 1 October 2021).
62. Dallolio, A.; Agdal, B.; Zolich, A.; Alfredsen, J.A.; Johansen, T.A. Long-Endurance Green Energy Autonomous Surface Vehicle Control Architecture. In Proceedings of the OCEANS 2019 MTS/IEEE SEATTLE, Seattle, WA, USA, 27–31 October 2019; pp. 1–10.
63. Norman, D. *The Design of Everyday Things: Revised and Expanded Edition*; Basic Books: New York, NY, USA, 2013.
64. Wiener, N. Some Moral and Technical Consequences of Automation. *Science* **1960**, *131*, 1355–1358. [CrossRef] [PubMed]
65. Allen, C.; Smit, I.; Wallach, W. Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches. *Ethics Inf. Technol.* **2005**, *7*, 149–155. [CrossRef]
66. Gabriel, I. Artificial Intelligence, Values, and Alignment. *Minds Mach.* **2020**, *30*, 411–437. [CrossRef]
67. Dawes, R.M. The Robust Beauty of Improper Linear Models in Decision Making. *Am. Psychol.* **1979**, *34*, 571–582. [CrossRef]