EMERGE

WP2 Ethics: Mapping risks and potential

D2.3 Map of risks and potentials for humans

Version: 1.0

Date: 30/09/2024

Funded by the European Union

European Innovation Council

Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

Document control

Project title	Emergent awareness from minimal collectives
Project acronym	EMERGE
Call identifier	HORIZON-EIC-2021-PATHFINDERCHALLENGES-01-01
Grant agreement	101070918
Starting date	01/10/2022
Duration	48 months
Project URL	http://eic-emerge.eu
Work Package	WP2 Ethics: Mapping risks and potential
Deliverable	D2.3 Map of risks and potentials for humans
Contractual Delivery Date	30/09/2024
Actual Delivery Date	30/09/2024
Nature[1]	R
Dissemination level[2]	PU
Lead Beneficiary	LMU
Editor(s)	Bahador Bahrami (LMU), Jurgis Karpus (LMU)
Contributor(s)	Nadine Meertens (LMU)
Reviewer(s)	Riccardo Guidotti (UNIPI)
Document description	*This document discusses potential ethical risks and benefits that could arise with the introduction of aware AI systems and reports findings from two experiments with which we investigated several of these concerns.*

[1]R: Document, report (excluding the periodic and final reports); DEM: Demonstrator, pilot, prototype, plan designs; DEC: Websites, patents filing, press & media actions, videos, etc.; DATA: Data sets, microdata, etc.; DMP: Data management plan; ETHICS: Deliverables related to ethics issues.; SECURITY: Deliverables related to security issues; OTHER: Software, technical diagram, algorithms, models, etc.

[2]PU – Public, fully open, e.g. web (Deliverables flagged as public will be automatically published in CORDIS project's page); SEN – Sensitive, limited under the conditions of the Grant Agreement; Classified R-UE/EU-R – EU RESTRICTED under the Commission Decision No2015/444; Classified C-UE/EU-C – EU CONFIDENTIAL under the Commission Decision No2015/444; Classified S-UE/EU-S – EU SECRET under the Commission Decision No2015/444

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

Version control

Version	Editor(s) Contributor(s) Reviewer(s)	Date	Description
0.1	Bahador Bahrami (LMU), Jurgis Karpus (LMU)	03.09.2024	TOC and section outlines
0.2	Bahador Bahrami (LMU), Jurgis Karpus (LMU)	09.09.2024	Section 1 draft completed
0.3	Bahador Bahrami (LMU), Jurgis Karpus (LMU)	13.09.2024	Section 1 draft updated, Section 2 draft in progress
0.4	Bahador Bahrami (LMU), Jurgis Karpus (LMU)	17.09.2024	Section 2 draft updated and completed
0.5	Bahador Bahrami (LMU), Jurgis Karpus (LMU), Nadine Meertens (LMU)	19.09.2024	Sections 1 and 2 drafts updated, Section 3 draft in progress
0.6	Bahador Bahrami (LMU), Jurgis Karpus (LMU), Riccardo Guidotti (UNIPI)	24.09.2024	Sections 3 and 4 drafts completed and ready for review
0.7	Bahador Bahrami (LMU), Jurgis Karpus (LMU)	26.09.2024	Section 5 draft completed, report updated based on reviewer's suggestions and passed on final review and sign-off
1.0	Davide Bacciu (UNIPI)	30.09.2024	Document submitted

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

Abstract

We review a number of ethical risks and potential benefits that may arise from the development and introduction of the concept of machine awareness into human dealings with artificial intelligence (AI) systems. We also report findings from two experiments, with which we addressed several of our raised concerns. Specifically, we investigated people's understanding of the concept of machine awareness that is developed by the project EMERGE and assessed people's demands for higher standards from explainable AI that ought to be anticipated from possible deployment of aware machines.

Funded by the European Union
European Innovation Council

Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

Disclaimer

This document does not represent the opinion of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA), and neither the European Union nor the granting authority can be held responsible for any use that might be made of its content.

Neither the EMERGE consortium as a whole, nor a certain party of the EMERGE consortium warrant that the information contained in this document is capable of use, nor that use of the information is free from risk and does not accept any liability for loss or damage suffered by any person using this information.

Acknowledgement

This document is a deliverable of the EMERGE project. This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement Nº 101070918.

Table of contents

Contents

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

List of figures

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

1. Introduction

Artificial intelligence (AI) research is progressing fast and ethicists, particularly those who focus on developing policy to regulate this new technology, are working hard to keep pace (Coeckelbergh 2020a; Nyholm 2022). Here we focus on the need to further expand the field of AI ethics to cover two new developments in robotics and AI: a shift from singular to collective and another from non-aware to aware AI systems. Specifically, novel risks and opportunities will have to be considered before people begin to interact with collective (as opposed to singular) and/or aware (including partially and gradually aware, as opposed to outright non-aware) systems powered by AI (Fig. 1).

Most work in AI ethics to date has focused on human dealings with non-aware, or thus far considered to be non-aware, systems. That is because *awareness* has been traditionally construed as closely related to *consciousness* (e.g., Black 2017; Seth and Bayne 2022) and, some notable exceptions aside (e.g., Chalmers 1995), most researchers agree that present-day technology is still a far way off from being conscious (McDermott 2007; Li et al. 2021; Chalmers 2023). That said, several large-scale interdisciplinary research projects, including EMERGE, are currently exploring ways to reconceptualize what it means for a human or a machine to be aware. Importantly, they make a deliberate point of separating machine awareness from consciousness and focusing specifically on the former. These developments call for a preemptive and fresh look at the ethics of possibly aware AI systems.

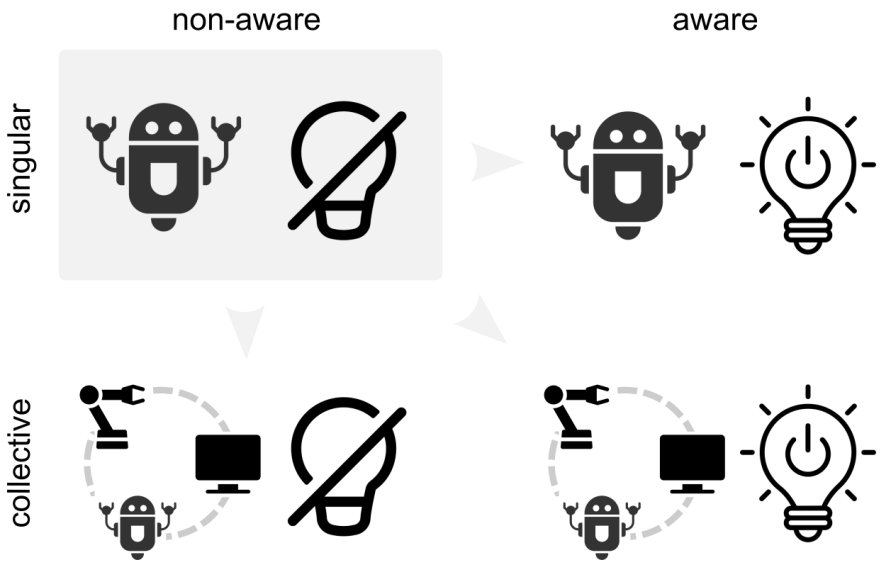

Figure 1. Two novel developments proposed by the project EMERGE. *We consider a shift from singular (top) to collective (bottom) and from non-aware (left) to aware (right) systems in human dealings with AI. Aware versus non-aware is not necessarily a binary distinction. Awareness may well be graded and multidimensional. Similarly, a shift from singular to collective systems can be further nuanced. For example, a collective system may consist of several singular, but connected individually aware components. Alternatively, collective awareness may be an emergent property of the system as a whole that supervenes on properties of its individual components.*

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

Also, while most work in AI ethics has been addressing challenges that arise in human dealings with singular AI systems, futuristic visions about the impending age of the "internet of things" and rapid developments in modular and swarm robotics urge us to explore the ethics of human interactions with collective (as opposed to singular) systems powered by AI.

While the two novelties—machine awareness and collectives—are conceptualised and developed simultaneously, it is prudent to assess and scrutinise them from the point of view of potential risks, opportunities, and ethical concerns that may arise in human dealings with aware and/or collective machines independently from one another. More generally, putting the spotlight on the ethical assessment of novel developments in AI research from an early stage is the natural step in the human-centred AI approach recommended by the EU AI Act (European Commission 2021).

Taking this approach, in this report we primarily focus on machine *awareness* and will return to machine *collectives* in future works (see also our earlier published D2.2 report for an initial assessment and empirical investigations concerning the latter).

Our report consists of four parts. In the first part we explain the concept of machine awareness and identify potential ethically relevant risks and benefits associated with the introduction of aware machines (Section 2). In the second part we report findings from an experiment, in which we investigated people's general perception of human and machine awareness. In particular, we sought to find out whether people's views of what awareness means to them coincides with the concept of machine awareness developed by the project EMERGE (Section 3). In the third part we discuss results of another experiment, in which we investigated people's evaluations of explainable aware AI (Section 4). In the fourth part we return to our identified ethically relevant risks in light of our initial empirical findings and offer practical recommendations for developers, deployers, supervisors, and regulators of AI systems as well as future directions in research.

1.1 Terminology

In this report, we use the terms **artificial agent**, **AI agent**, **AI system**, **machine**, and **intelligent machine** interchangeably even if there can be meaningful and important differences between them in other contexts. As we will explain shortly in more detail below, we treat **awareness** as an agent's state that is distinct from **consciousness** and take the view that awareness is best construed as something that comes in degrees (more on this in Section 2.1).

1.2 Connections to the EU AI Act

The ethically relevant aspects of human dealings with AI systems that we cover here are closely connected to several themes in the proposal for the European Union's Artificial Intelligence Act—the "EU AI Act" (European Commission 2021). Particularly important are demands for appropriately attributed and shared responsibility, clear and effective communication and transparency, and explainability of AI systems in human dealings with machines. We will review what novel benefits and risks may arise with the introduction of aware machines in connection to these topics in Section 2.2 and Section 2.3.

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

2. Machine awareness

2.1 Defining and measuring awareness

Philosophical debate and everyday usage of the terms often treat awareness and consciousness interchangeably, or see awareness as an aspect of consciousness (Zeman 2006; Seth and Byne 2022). Chalmers (1995) challenged this tendency, proposing awareness to be a somewhat simpler notion than consciousness. Awareness, according to him, allows one to explain what he called “easy problems”, for example, an agent’s ability to discriminate, categorise and react to environmental stimuli. The difficulty in disentangling awareness from consciousness, however, remains due to their ambiguous usage to date (Black 2017). Theories of consciousness vary widely and range from providing fairly minimal to highly complex and demanding concepts of that state (the state of being conscious). The science about consciousness, therefore, addresses different challenges depending on one's chosen theory and definition. Taking consciousness in the phenomenal sense, for example, refers to subjective experiences or ‘what it is like’ to experience something (Nagel 1974). In essence, this concerns the mental states or the ‘inner life’ of a conscious agent. When it comes to artificial agents, however, it is worthwhile to look at and assess their abilities and capacities whilst remaining neutral on the possibility of them having ‘inner lives’, adopting the distinction between awareness and consciousness.

The extent to which different systems can engage with their environment, themselves, and other systems varies. The concept of awareness is useful to explore this variation, thereby allowing us to look at, for instance, the detection, registration, and responsiveness of an artificial agent to external environmental events or changes in its internal states. This could reveal the type of strategies a system can employ in the attainment of its goals and the space of action-perception capacities that are available to it. Take, for instance, a vacuuming robot like Roomba. We can ask how aware it is of the space it is navigating through and the type of objects and other agents that it may encounter in that space by assessing how successfully it can adapt to changes in its environment (for example, the introduction of a new couch or the presence of another moving agent) without being interested in what, if anything, it might feel like for the robot to be able to do just that. Crucially, by disentangling awareness from consciousness we can sidestep the debate concerning artificial agents’ consciousness—a property that is often reserved for living beings—but nevertheless have a useful and generalizable concept to discuss and assess their capacities to successfully perform their tasks when environments and circumstances change (for more on this idea, see McDermott 2007; Li et al. 2021; Chalmers 2023).

Based on these considerations, the project EMERGE is developing a functional account of (machine) awareness that is best construed as distinct from consciousness. It is a functional account because, since awareness is associated with an agent’s capacities to perform certain tasks, the agent’s awareness itself can be assessed and measured in terms of that agent’s relevant experimentally testable and observable performance in varying environments and circumstances. This is roughly in line with the behaviourist approach in AI ethics research to assess and determine what status could or should be attributed to machines based on their performance in relevant domains, the general idea of which was originally proposed by Turing, but has been developed further since and most recently advocated in AI ethics circles by Danaher (2019, 2021; see also Nyholm 2022, chapters 8 and 9). This also presents the opportunity to conceptualise awareness as multi-dimensional and graded, which can be of great use in practice. For example, a robot that has a high degree of spatial awareness of the environment that it navigates could be expected and relied on to perform certain tasks more

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

efficiently and more robustly compared to a robot that has a low degree of awareness in the relevant (spatial) dimension.

2.2 Benefits of machine awareness

There are numerous potential benefits from developing and introducing the concept of (machine) awareness to people's interactions with AI. One of them we already mentioned. Knowing an AI system's awareness dimensions and its degrees of awareness in those dimensions should make it easier for human users of that system to rely on and delegate tasks to it. For example, knowing that one's fully automated (self-driving) vehicle has a high degree of awareness of the variety of traffic participants and other agents that it may encounter—pedestrians, cyclists, other vehicles, careless children playing ball close to schools, and so on—would make it easier for one to delegate the task of driving to it in busy urban traffic. Conversely, knowing that one's conversational AI assistant has a low degree of awareness of the truthfulness of its boldly stated claims on a particular topic would make one rightly more sceptical about the assistant's communicated "facts".

Relatedly, accurately and clearly communicated awareness dimensions of a machine and its degrees of awareness in those dimensions will allow its human users and supervisors to have a better understanding of where human oversight is the most important and required in collaborative human-machine teams. For example, suppose that one's job requires one to speedily disseminate important information to a wide audience in multiple languages and that one does that with the help of an automated translation tool. Knowing gaps in the translation tool's awareness of important nuances in specific languages will make it easier for one to know translations of which topics and into which languages will require a more careful human oversight and checking (and concerning which topics and languages the tool can be safely relied upon in this collaborative work).

Machine awareness can also safeguard people from unintended and unwelcome side effects of automation. An automated system's awareness of potential problems associated with factors on which it bases its decisions and that therefore contribute to its performance success, whether in the domain of recommending fitness training plans to human users or in filtering job applicants' CVs in a human resources department, will allow it to flag those problematic cases to its human supervisors and thereby help avoid potentially disastrous outcomes. This is particularly relevant in the context of automated algorithmic nudging of human behaviours (Sætra 2019; Schmauder et al. 2023). Simply put, the problem with unaware AI systems is that, without being aware of how and why exactly they achieve their tasked objectives, they may produce unintended side effects without anyone noticing. Crucially, in order to be able to flag a potential problem with what one does, one needs to be aware of that potential problem in the first place (for a more detailed discussion of this particular point, see also our previously published D2.2 report).

Machine awareness also brings about several important and useful implications for learning. As discussed in a separate report D1.3, one important dimension of awareness is awareness of one's agency, or *agentive awareness.* This enables an agent to assign credit, for example, to distinguish between what was caused by the agent itself and what was caused by other parties or environmental events (e.g., other agents, natural phenomena such as wind, and so on). Credit assignment is particularly useful for any agent that needs to learn things about its environment and possible courses of action through trial and error.

Another important dimension of awareness is confidence—a form of cognition about cognition, i.e., metacognition. Confidence refers to an agent's sense of uncertainty about the present

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

and future states of the agent's environment and its probabilistic estimates for future outcomes associated with different possible courses of action. Confidence is important for learning because it allows the agent to evaluate its observations and decide whether they are critical enough for the agent to adapt to. An agent endowed with such metacognitive awareness can choose what to adapt to and what can be ignored based on its level of confidence.

2.3 Risks of machine awareness

While there certainly are benefits to using the developed concept and the term *machine awareness* (and *degrees* thereof) in human dealings with AI systems, this also carries risks. Firstly, there is the danger of arising confusions and misunderstandings between different stakeholders—developers, deployers, supervisors, end users, and people who will be affected by the deployment of advanced technologies—about what machine awareness actually means when AI systems and/or robots are described or marketed as aware of certain things. For example, will all parties agree to the idea that awareness is graded as opposed to being categorically "on" or "off"? If people think of awareness as an all-or-nothing state, they may over-rely on machines that are described to them as aware of certain things albeit to a small degree and fail to differentiate between capabilities of machines that differ in the degree of their awareness. Relatedly, will everyone view awareness as distinct from consciousness? If not, should they be asked to change their view and how should then the term *machine awareness* be best explained?

In a recent e-mail newsletter, Ben Schneiderman, a prominent computer scientist and an astute promoter of the human-centred AI approach, discussed potential problems with using metaphors and human-like terms to describe computer systems. The point that is relevant to us here was well captured by Google's Gemini chatbot, which Schneiderman queried about why it uses the pronoun "I" in conversations with its human users. Here is Gemini's response (as quoted by Schniederman): "Using 'I' can make computers feel more human-like and relatable to users… It can help people understand and interact with computers as if they were sentient beings." However, Gemini also recognized (or rather, reported) that "It can create a false impression of sentience or consciousness… It can lead to unrealistic expectations and emotional responses… It raises questions about the potential for computers to develop self-awareness or consciousness… It's important to consider the potential implications and avoid creating confusion or misunderstandings." This response aptly captures some of the problems that may arise with the introduction of the term *machine awareness* to human dealings with AI. It is important that everyone has the same and clear understanding of what machine awareness means so as not to misattribute certain capacities to computer systems that are described as aware of certain things and/or aware of them to some degree.

Another point to scrutinise is that machine awareness may come hand-in-hand with heightened demands for ascriptions of responsibility to machines or their deployers. Whether intelligent machines themselves (or their deployers) can be held responsible for anything that they produce or otherwise contribute to bring about has been extensively debated in AI ethics as of late. According to one prominent view, the involvement of intelligent machines, especially in the case of "black box" AI systems, the inner workings of which are not easy to explain and monitor, can introduce responsibility gaps in human dealings with them: cases when someone ought to be held responsible for the occurrence of some outcome, but there is nobody (or nothing) to meaningfully attribute that responsibility to (Nyholm 2022, 2023). However, suppose that one's decision on some matter leads to an unwelcome outcome, possibly as a side-effect to whatever one's actual goal may have been. It seems natural to assign greater responsibility to a decision-maker for that outcome when they were aware of the key

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

circumstances contributing to that result, even if the outcome was unintended. In contrast, if the same decision-maker was completely unaware of these crucial circumstances, their level of responsibility might be considered lower. This close connection between awareness and responsibility will have to be kept in mind when aware robots or AI systems are developed, deployed, and marketed as such. Even if the problem of responsibility gaps remains in human interactions with aware machines, describing machines as aware or partially aware of certain things could increase the demand and drive to address those gaps.

Relatedly, machine awareness may lead to demands for higher standards from explainable AI. When a machine is described as aware of factors that are relevant to evaluating its decisions, people will likely expect more thorough explanations from it for its decisions compared to a machine that lacks awareness of those relevant factors and circumstances. Such demands from users and supervisors of AI systems ought to be anticipated in advance of marketing and distributing aware machines. For a further discussion of connections between awareness, responsibility, and explainability in ethical assessments of aware machines, please refer to our previously published D2.2 report.

2.4 Recommendations

Our highlighted range of potential risks and ethical concerns calls not only for a new theoretical branch of AI ethics specifically tailored to address the introduction of aware machines, but also a concerted effort in behavioural and cognitive science research. Understanding how AI systems impact human behaviour and cognition will be crucial for developing ethical guidelines that are grounded in real-world implications.

To study our identified questions and concerns effectively, it is important to foster greater interdisciplinary collaboration between roboticists, computer scientists, philosophers, behavioural scientists, end users of intelligent machines, and policymakers. In the next two sections, we report our initial efforts to conduct such interdisciplinary research. Our aim is to provide a model for integrating ethical, behavioural, and cognitive insights into AI system design and policy, thereby contributing to the responsible advancement of AI technology.

3. People's perception of human and machine awareness

3.1 Research questions

As we discussed earlier, one concern with the introduction of the concept of machine awareness was the danger of potential confusions and misunderstandings between different stakeholders about what machine awareness actually means to them when AI systems and/or robots are described or marketed as aware of certain things. Eliciting the views of non-experts—potential end users of these systems, or people who may in other ways be affected by the deployment of such technology—is particularly important (Coeckelbergh 2020b). Firstly, as we already noted, there is currently no consensus among experts—for example, philosophers of mind—on how awareness ought best to be construed. As such, it is fruitful to examine which conception of awareness is best aligned with ordinary people's views. Secondly, while it is possible to instruct expert stakeholders on how to conceptualise machine awareness in their interactions with such systems, asking the same of laypeople is a more challenging task. Therefore, in order to address the aforementioned worry, we conducted an experiment to investigate the following four questions.

- Do people think of awareness in line with the functional account of awareness developed by the project EMERGE? In particular, does people's ascription of

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

awareness to an agent vary with the agent's theorised degree of awareness based on that agent's functional description?

- Do people think of awareness as an all-or-nothing state or as something that can come in degrees?
- Do people consider machine and human awareness to be alike? In particular, if a machine and a human are functionally described in similar ways, will people ascribe similar (degrees of) awareness to both?
- Do people consider awareness as distinct from consciousness?

3.2 Experiment design

We set up our experiment as a vignette-based empirical study, in which human participants read three fictional stories, each of which concerned a different scenario: cleaning, manufacturing, or catering. Depending on the treatment group to which participants were assigned, these stories involved either a human or a machine protagonist. The protagonist was functionally described as having one of three theorised degrees of awareness—low, medium, or high—concerning the protagonist's task at hand. In total, this comprised 18 possible stories that a participant might read (3 scenarios x 2 types of protagonist x 3 theorised degrees of protagonist's awareness). Examples of three variations of these stories are shown in the box below.

Example 1: cleaning, human, low awareness

Malika is a full-time member of the cleaning staff of a large office building. Malika's daily tasks include vacuuming the floors, disinfecting doorknobs, and wiping down desk surfaces. To block out her surroundings, Malika listens to her favourite podcast with her noise-cancelling headphones. Fully captivated by the story, she is no longer responsive to the environment and performs her tasks on 'autopilot'. When an automatic door suddenly closes behind her, she turns around and knocks over a vase.

Example 2: manufacturing, machine, medium awareness

Constructo, a state-of-the-art manufacturing robot, assembles car windows in the production line of a large car manufacturer. Constructo is equipped with six sensors, including four cameras and two infrared lights, that provide a detailed spatial layout of its environment. While the production line is running, Constructo has to identify the approaching car and install the suitable window. After a recent collision has damaged its three sensors, Constructo frequently monitors the state of the remaining sensors. Its focus, hence, frequently shifts between monitoring and performing the job at hand. When a colleague abruptly starts to drill metal behind it, Constructo turns around and knocks off the car's side mirrors.

Example 3: catering, machine, high awareness

RoboWaiter is a state-of-the art service robot working as a waiter in a local restaurant. RoboWaiter is equipped with six sensors, including four cameras and two infrared lights, that provide a detailed spatial layout of its environment. RoboWaiter takes customer orders, serves food and handles payments. After completing a routine checkup and charging its batteries overnight, RoboWaiter is fully operational. Paying full attention to its surroundings, RoboWaiter is fully focused on the job at hand. When a guest unexpectedly drops cutlery behind it, RoboWaiter turns around and knocks over a bottle of wine.

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

Each participant in the experiment was randomly assigned to one of two treatments. In one treatment the protagonist in all presented stories was a human; in the other treatment—a machine. There were thus 9 stories to sample from for each participant. Each participant received 3 of these 9 possibilities such that they encountered each scenario (cleaning, manufacturing, or catering) and each functional description of the three theorised degrees of protagonist's awareness (low, medium, or high) only once. The order in which these stories were presented was randomised for each participant as well.

After reading each story, participants answered three questions—two using a 7-point Likert scale and the third as unconstrained text:

Q1. How aware is [protagonist] of [its/her/his] spatial surroundings?

Not aware at all [1] ★★★★☆☆☆ Fully aware [7]

Q2. Is [protagonist] capable of having conscious experience of [its/her/his] spatial surroundings?

Not capable at all [1] ★★☆☆☆☆☆ Fully capable [7]

Q3. Why did [protagonist] [knock over the vase / knock off the car's side mirrors / knock over a bottle of wine]?

[Unconstrained text box to provide an answer]

Before proceeding to read and evaluate presented stories (the main part of the experiment), participants completed a short language test to gauge their command of English (Fig. 2). In this test, participants saw ten short sentences. Their task was to identify which of those (e.g., "They are playing in the park" or "He don't like eat vegetables") were written in *incorrect* English. Each participant received a score out of a maximum of 10 points (1 point for each sentence). Since it is easy to mistake the language test question for one that asks to identify sentences in *correct* English, we computed scores for both versions of the question and then used the maximum value of the two to determine a participant's final score. We subsequently excluded participants who scored 6 points or less (out of 10) in our data analysis.

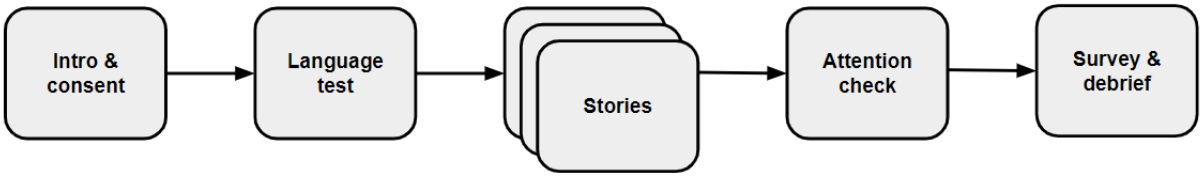

Figure 2. Chronological structure of the experiment procedure. *The diagram shows the succession of screens seen and tasks completed by participants.*

After the main part of the experiment, participants also completed a short attention check. Here participants were asked to identify which three from a list of six topics were addressed in fictional stories presented to them earlier in the experiment. We subsequently excluded participants who scored 4 points or less (out of 6) in our data analysis. However, irrespective of their performance in the language test and the attention check (which we did not disclose to our participants) all participants who entered the experiment were able to complete it in full and received payment for their participation. At the end of the experiment, participants completed a short demographics survey to record their age and gender.

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

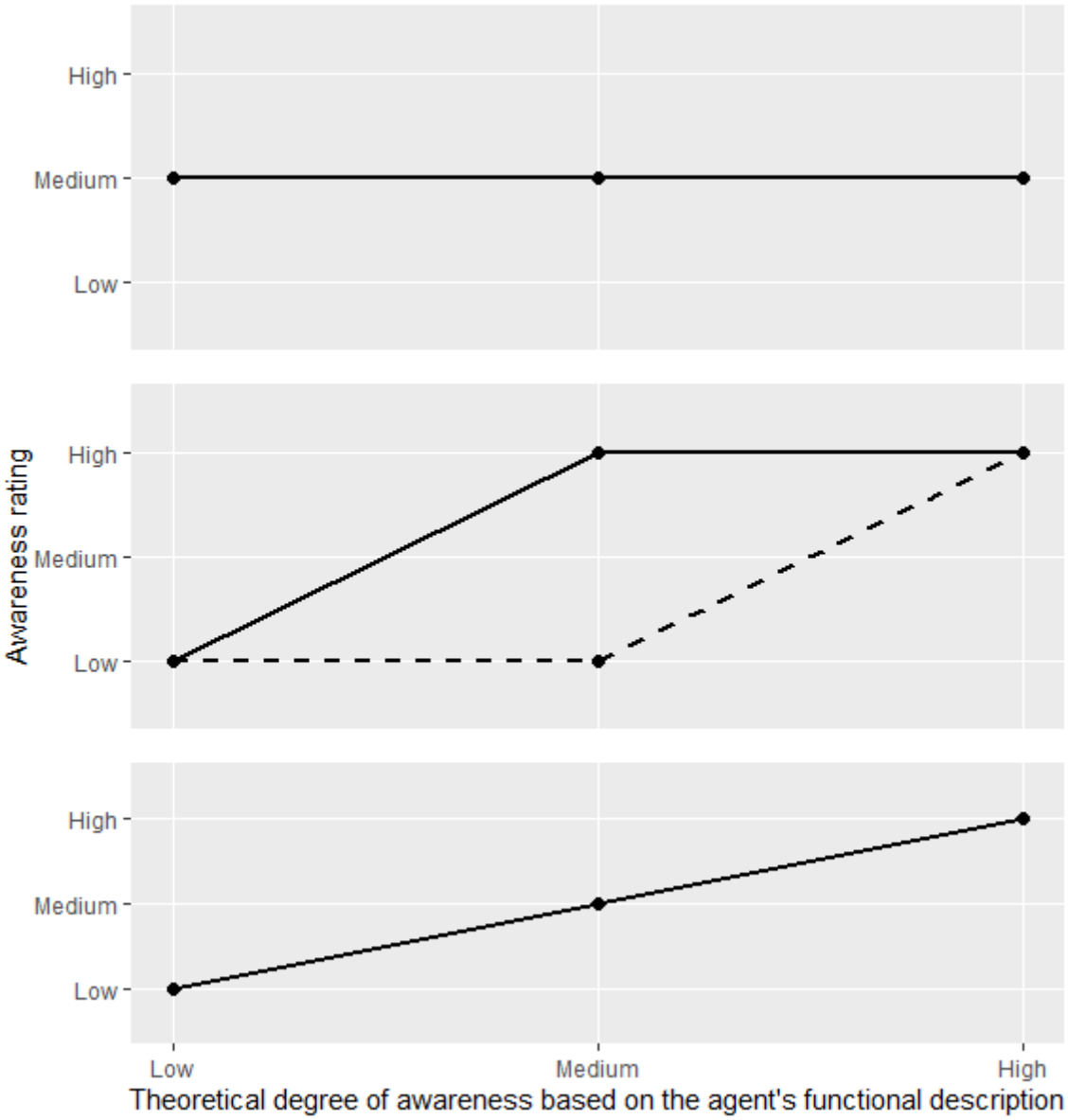

Figure 3. Hypotheses. *Expected results for different possible answers to our research questions. Y axis: participants' ascribed degree of awareness to their evaluated agent. X axis: theorised degree of awareness based on the agent's functional description that was presented to participants. Top panel: people do not think of awareness in line with the functional account of awareness developed by the project EMERGE. Middle panel: people's views support the functional account of awareness, but they think of awareness as an all-or-nothing state. The solid and the dashed lines show two possible patterns in data that would be consistent with this conclusion. Bottom panel: people's views support the functional account of awareness and they think of awareness as something that can come in degrees.*

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

We programmed the experiment in Qualtrics (https://www.qualtrics.com/) and conducted it online. We recruited our participants in the United Kingdom on Prolific—an online platform that is commonly used to recruit human participants for behavioural and survey-based studies in social science research (https://www.prolific.com/). Based on related previous empirical studies, our target was to recruit 200 participants in two batches. In the first batch we recruited 20 participants to test our experiment procedure and to update our estimated experiment completion time and the corresponding pay that we advertised to remaining participants recruited in the second batch. However, our initial estimate of 10 minutes or less proved accurate. Each participant received £1.80 for taking part, which was a slightly higher rate for 10 minutes worth of work based on the minimum hourly wage in Germany (€12.41) and the exchange rate (€1 = £0.84) at the time of data collection (September 2–3, 2024).

The Ethics Committee of the Faculty of Philosophy, Philosophy of Science and Religious Studies at Ludwig-Maximilians-Universität in Munich (LMU Munich) approved the study after it was reviewed for compliance with ethical research standards (ID number 238497). Prior to collecting data, we pre-registered our experiment design, data analysis plan, and core hypotheses on the Open Science Framework (OSF) database online (https://doi.org/10.17605/OSF.IO/EPRS9). We obtained informed consent from all participants who took part in our study and will subsequently make our dataset and statistical analyses freely and publicly available for further research on OSF.

Figure 3 shows what we expected to see in our data for different possible answers to our first two research questions. Participants' ascribed degree of awareness to their evaluated agent (a human or a machine protagonist) is on the y axis, while the theorised degree of awareness based on that agent's functional description that was presented to participants is on the x axis. Thus, if people's view of awareness is in line with the functional account of awareness developed by the project EMERGE, we expect to see a positive correlation between the two variables as shown in the middle and bottom panels of Fig. 3. Furthermore, if people think of awareness as something that can come in degrees, we expect this correlation to be particularly strong. Specifically, participants' ascribed degree of awareness to an agent should match that agent's theorised degree of awareness as shown in the bottom panel of Fig. 3. If, on the other hand, people think of awareness as an all-or-nothing state, we expect a sharp increase in participants' ascribed degree of awareness to an agent to occur once as we move from left to right along the x axis as shown in the middle panel of Fig. 3.

3.3 Results

We discuss descriptive initial insights from collected data. Additional statistical analyses are underway and we will report those in future publications.

199 participants logged into our experiment and successfully completed it. 12 participants failed our attention check after they completed the main task in the experiment (all these participants scored 4 points out of 6 in this test). That left 187 participants' responses to analyse (63% women, 35% men, 1% non-binary or other, and 1% who preferred not to disclose their gender; mean age ± 1 standard deviation = 40.1 ± 11.3).

As we explained earlier, each participant in the experiment read three fictional stories, all of which involved either a human or a machine protagonist, depending on the treatment group to which the participant was randomly assigned. Figure 4 shows participants' ascribed degrees of awareness to their evaluated agents ranging from "Not aware at all" [1] to "Fully aware" [7] (on the y axis) for three theorised degrees of awareness based on the agents' functional descriptions that were presented to participants in stories (on the x axis). The top and bottom

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

panels in the Figure show participants' ascribed degrees of awareness to their evaluated human and machine protagonists respectively.

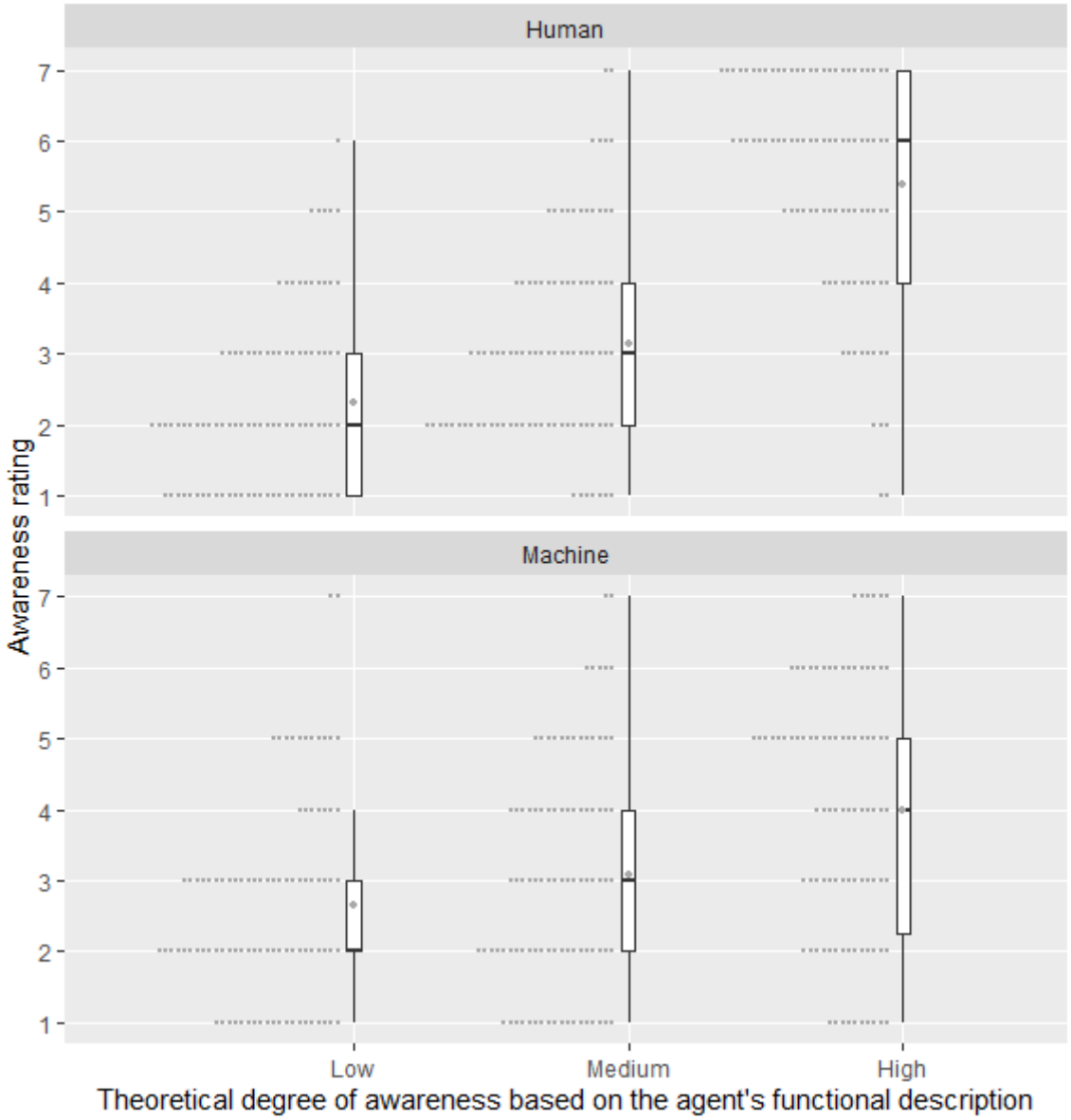

Figure 4. Awareness ratings. *Y axis: participants' ascribed degree of awareness to their evaluated agent ranging from "Not aware at all" [1] to "Fully aware" [7]. X axis: theorised degree of awareness based on the agent's functional description that was presented to participants. Top panel: awareness ratings ascribed to agents by participants who evaluated human protagonists. Bottom panel: awareness ratings ascribed to agents by participants who evaluated machine protagonists. For each theorised degree of awareness (column), small dots on the left show distributions of individual responses. The box-plots show the median values (black horizontal lines) as well as the 1st and the 3rd quartiles in individual responses (lower and upper ends of the white boxes). The grey dots on top of the box plots are mean values.*

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

For a reminder, we list our research questions using bullets before discussing answers to them. The first two were the following.

- Do people think of awareness in line with the functional account of awareness developed by the project EMERGE? In particular, does people's ascription of awareness to an agent vary with the agent's theorised degree of awareness based on that agent's functional description?

- Do people think of awareness as an all-or-nothing state or as something that can come in degrees?

As the results in Fig. 4 show, participants' ascribed degrees of awareness to both human and machine protagonists reveal a general trend that follows our hypothesised pattern in the bottom panel of Fig. 3. This suggests two things. Firstly, that people think of awareness in line with the functional account of awareness developed by the project EMERGE. And secondly, that they construe awareness as something that can come in degrees.

Notice, however, that the increase in participants' ascribed degree of awareness to their evaluated human protagonists is greater as we move from medium to high compared to the move from low to medium theorised degrees of awareness based on the evaluated agents' functional descriptions. This latter finding suggests that our participants may fall into two groups: those who construe awareness as something that can come in degrees (and whose ascribed degrees of awareness follow the hypothesised pattern in the bottom panel of Fig. 3) and those who construe it as an all-or-nothing state (and whose ascribed degrees of awareness follow the hypothesised pattern of the dashed line in the middle panel of Fig. 3).

- Do people consider machine and human awareness to be alike? In particular, if a machine and a human are functionally described in similar ways, will people ascribe similar (degrees of) awareness to both?

Figure 4 suggests that to some extent people consider human and machine awareness to be alike. In cases of low and medium theorised degrees of awareness, participants ascribed similar degrees of awareness to both types of agent. However, in the case of high theorised degree of awareness, participants ascribed a significantly higher degree of awareness to humans compared to machines. We speculate that when it comes to high theorised degrees of awareness for both types of agent, people begin to consider awareness to be an aspect of an agent's consciousness.

- Do people consider awareness as distinct from consciousness?

Figure 5 shows participants' judgements about their evaluated agents' capacity to consciously experience spatial surroundings ranging from "Not capable at all" [1] to "Fully capable" [7] (on the y axis) for the three theorised degrees of awareness of those agents (on the x axis). As previously, the top and bottom panels in the Figure show participants' judgements when they evaluated human and machine protagonists respectively.

As the results clearly show, participants ascribed a significantly higher consciousness rating to humans than to machines for all three theorised degrees of awareness of evaluated agents. Furthermore, in cases of low and medium theorised degrees of awareness, participants' consciousness ratings were higher than awareness ratings when they evaluated human protagonists. Conversely, in cases of medium and high theorised degrees of awareness, participants' consciousness ratings were lower than awareness ratings when they evaluated machine protagonists. Combined, these results suggest that by and large people consider

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

awareness to be distinct from consciousness. However, it is important also to note that a sizable group of participants considered machines to be capable of conscious experience. We hope that a deeper statistical analysis of results than what we are able to offer here will help us reveal whether these participants also tended to consider awareness as an aspect of consciousness, or whether the two are unrelated.

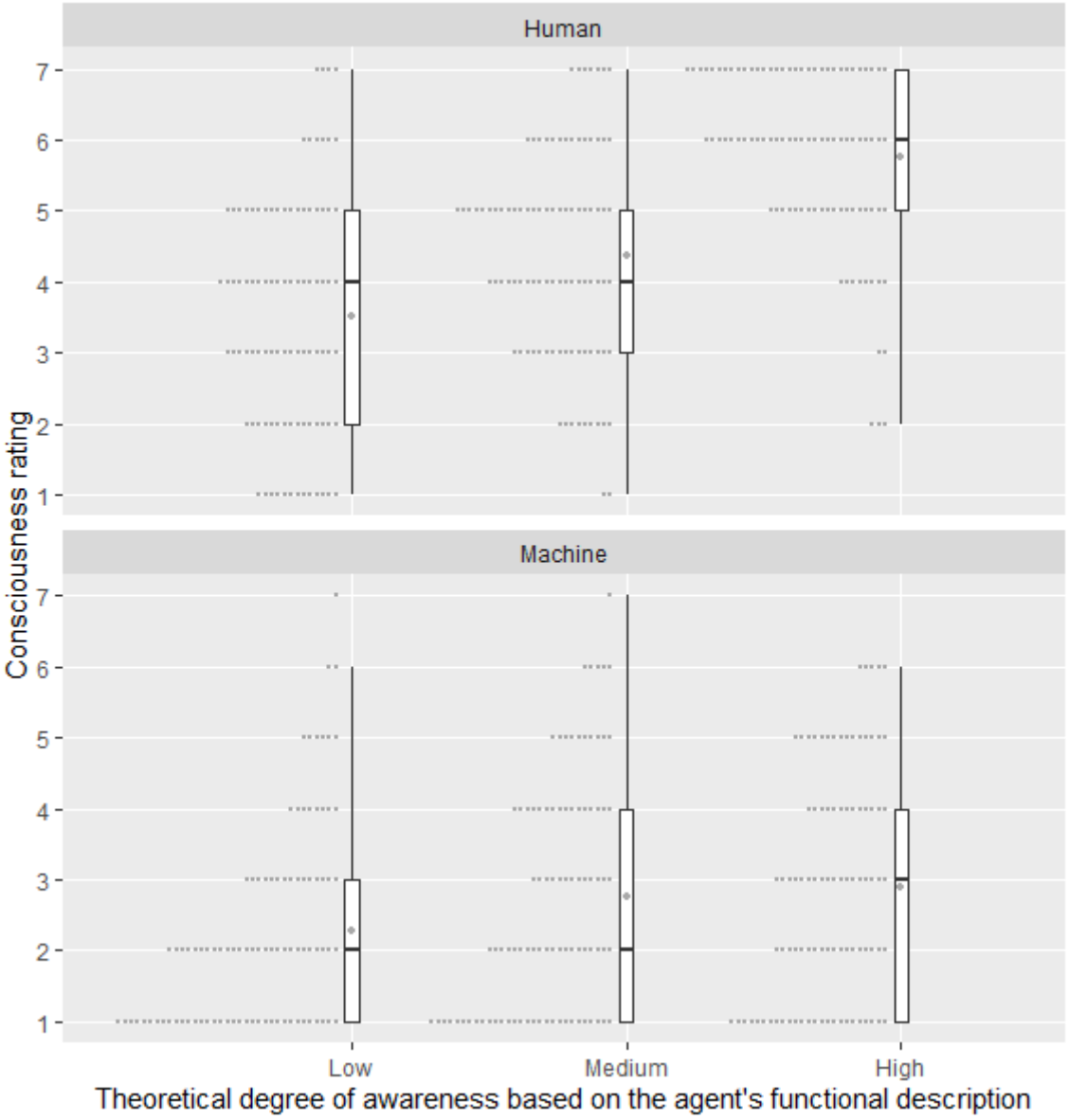

Figure 5. Consciousness ratings. *Y axis: participants' judgement about their evaluated agent's capacity to consciously experience spatial surroundings, ranging from "Not capable at all" [1] to "Fully capable" [7]. X axis: theorised degree of awareness based on the agent's functional description that was presented to participants. Top panel: consciousness ratings by participants who evaluated human protagonists. Bottom panel: consciousness ratings by participants who evaluated machine protagonists. For each theorised degree of awareness (column), small dots on the left show distributions of individual responses. The box-plots show the median values (black horizontal lines) as well as the 1st and the 3rd quartiles in individual responses (lower and upper ends of the white boxes). The grey dots on top of the box plots are mean values.*

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

3.4 Limitations and future research

As we noted earlier, we offered only descriptive initial insights from our collected data in this experiment and, as such, our reported findings here should be taken with a pinch of salt. Further statistical analyses are needed to corroborate our initial insights. They will also allow us to answer many additional questions. For example, what proportion of people construe awareness as something that can come in degrees and what proportion construe it as an all-or-nothing state? Do people who consider machines incapable of conscious experience ascribe varying degrees of awareness to them that are in line with the machines' theorised degrees of awareness based on their functional description? We hope to answer these and related questions in future publications.

4. People's perception of explainable aware AI

4.1 Research questions

As we discussed in Section 2.3, the introduction of the concept of machine awareness to descriptions of deployed and marketed AI systems may lead to demands for higher standards from explainable AI. However, similarly as with the concept of awareness, it is not obvious, from both scientific and philosophical points of view, what constitutes a good explanation, what type of explanation is the most appropriate, and how detailed an explanation ought to be in human dealings with AI (Coeckelbergh 2020b). As Coeckelbergh pointed out, the possible level of fine-grained detail in an issued explanation, for example, in terms of articulated causal chains that led to some decision or an outcome, and what is expected from and deemed as adequate in an explanation by those who demand it, are two separate questions. Therefore, eliciting the views of non-experts—potential end users of machines and those who may be affected by other people's use of AI systems—on what counts as an adequate explanation of a machine's decision is prudent in this context as well. In order to address this matter, we conducted a second experiment to investigate the following three questions.

- Do people expect a higher standard of explanation from machines described as aware compared to those described as non-aware?
- Does the answer to the first question depend on the type of issued explanation—one that refers to causes versus one that refers to authority?
- Does the answer to the first question depend on the purpose of the issued explanation, e.g., when it is to be used in litigation?

4.2 Experiment design

As in the previous case, we set up our second experiment as a vignette-based empirical study, in which human participants read three fictional stories, each of which concerned a different scenario: translation, firefighting, or hiking. Each story involved a machine protagonist—a software tool or a drone powered by AI. The machine was described to participants as basic, attuned to context, or specifically designed to be aware of a particular matter that was relevant to the story at hand. Each story also included an explanation provided by the machine of its decision. Depending on the treatment group to which participants were assigned, this explanation referred either to causes (of the machine's decision) or authority. In total, this comprised 18 possible stories that a participant might read (3 scenarios x 3 descriptions of machine x 2 types of machine's provided explanation). Examples of three variations of these stories are shown in the box below.

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

Example 1: translation, basic, explanation refers to causes

Doctors at High View hospital's emergency department in London use automated translation tools to communicate information to patients who do not speak English.

These translation tools are common and similar to the well known and widely available ones, like Google Translate, DeepL and ChatGPT.

Recently, after receiving emergency treatment, a patient who was leaving the hospital received instructions on which medication to take, using one of these tools.

A month later, during a medical check-up, it was discovered that the patient did not fully understand instructions they had been given and had mistakenly taken the wrong medication for their follow-up treatment.

During the investigation, the automated translation system was asked to explain why it gave the particular translation that it did, in order to gain insight into why the mistake happened. Here is the explanation that it provided:

I perform translations by utilizing a combination of statistical methods, pattern recognition, and linguistic rules. My training involved exposure to vast amounts of text in multiple languages, allowing me to learn the relationships between words, phrases, and sentences in different languages. When you request a translation, I analyze the input text, identify the language, and then generate an appropriate translation based on the learned patterns and rules.

Example 2: firefighting, attuned-to-context, explanation refers to authority

At Fire Station 14, Crew Manager Priya Desai and her team responded to a house fire on Maple Lane, deploying a fully-automated drone to scout the area.

The state-of-the-art drones that the firefighters use are specifically designed to provide critical and efficient assistance in firefighting missions.

The drone that was deployed initially provided critical information but lost its video feed and crashed through the roof, destabilizing the structure.

Despite the setback, Desai and her team continued their mission, rescuing an elderly couple from the basement. The fire was extinguished without casualties, but the crash made the rescue mission more difficult, underscoring the importance of good human judgment in firefighting.

During the mission review, to aid investigation, the automated drone operating system issued the following explanation:

My operated drone maintained a safe distance of 5 meters from the structure until the moment of malfunction. The 5 meter benchmark for operational safety that I followed was issued by human firefighting experts. The incident occurred despite me following the protocol.

Example 3: hiking, aware, explanation refers to causes

James Mitchell from Sheffield used an app powered by artificial intelligence to plan a hiking route in a mountainous region while on holidays in the USA.

The state-of-the-art route planning app that he used is specifically designed to be aware of daily weather conditions and necessary equipment.

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

Regrettably, confronted with exceptionally harsh weather and treacherous terrain, James became stranded in a canyon, necessitating a call for mountain rescue and a helicopter extraction.

Although the incident ended well, James is now faced with a substantial fee for the helicopter extraction that was required to ensure his safety.

When reviewing the incident, the automated advising system on which the app is built was asked to explain why it recommended to James the particular route that it did. Here is the explanation that the app issued:

I determine a recommended route based on various factors, including terrain, distance, time, historical data, user preferences, transit options, and traffic. For example, I use historic hiking and biking patterns to predict route popularity and congestion levels at different times of the year. I also work out a realistic distance and route completion time based on the user's past hiking patterns.

Crucially, irrespective of whether a machine was described to participants as basic, attuned to context, or specifically designed to be aware of a particular matter that was relevant to the story at hand, the explanation provided by the machine in a given scenario (whether it referred to causes or authority) always remained the same. For example, for all possible descriptions of a machine (basic, attuned to context, or aware) in the translation scenario, the system's issued causal explanation was the one shown in the example above.

Each participant in the experiment was randomly assigned to one of two treatments. In one treatment the machine's issued explanation in presented stories referred to causes; in the other treatment—to authority. There were thus 9 stories to sample from for each participant. Each participant received 3 of these 9 possibilities such that they encountered each scenario (translation, firefighting, or hiking) and each description of a machine (basic, attuned to context, or specifically designed to be aware of a particular matter that was relevant to the story at hand) only once. The order in which these stories were presented was randomised for each participant as well.

After reading each story, participants answered two questions using a 7-point Likert scale:

Q1. Do you think this explanation is adequate, considering the situation and what you'd expect from a [translation tool / drone operating system / route planning app] like this?

Not adequate at all [1] ★★★★☆☆☆ Fully adequate [7]

Q2. Do you think this explanation is adequate for use in legal matters, like in a court case?

Not adequate at all [1] ★★☆☆☆☆☆ Fully adequate [7]

The rest of the experiment procedure was identical to the one we used in the previous experiment (Fig. 2). Participants completed a short language test and an attention check. We conducted the experiment online and recruited our participants in the United Kingdom on Prolific. Based on related previous empirical studies, our target was to recruit 200 participants. We ensured fair pay to our participants and our initial estimate of 10 minutes or less for completing the experiment proved accurate after we reviewed responses from the initial 20 participants (each participant received £1.80 for taking part, which was in line with the minimum hourly wage in Germany and the EUR to GBP exchange rate at the time of data collection on September 2–3, 2024).

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

The Ethics Committee of the Faculty of Philosophy, Philosophy of Science and Religious Studies at Ludwig-Maximilians-Universität in Munich (LMU Munich) approved the study after it was reviewed for compliance with ethical research standards (ID number 238497). Prior to collecting data, we pre-registered our experiment design, data analysis plan, and core hypotheses on the Open Science Framework (OSF) database online (https://doi.org/10.17605/OSF.IO/YM2T3). As previously, we obtained informed consent from all participants who took part in our study and will subsequently make our dataset and statistical analyses freely and publicly available for further research on OSF.

4.3 Results

As with the previous experiment, we discuss descriptive initial insights from collected data. Additional statistical analyses are underway and we will report those in future publications.

200 participants logged into our experiment and successfully completed it. 1 participant failed our language test and 23 failed our attention check after they completed the main task in the experiment (the participant who failed the language test also failed the attention check). That left 177 participants' responses to analyse (62.7% women, 36.1% men, 0.6% non-binary or other, and 0.6% who preferred not to disclose their gender; mean age ± 1 standard deviation = 38.8 ± 11.9).

As we explained earlier, each participant in the experiment read three fictional stories, in each of which a differently described machine provided an explanation for its decision. That explanation referred either to causes or to authority, depending on the treatment group to which the participant was randomly assigned. Figure 6 shows participants' ratings of the adequacy of their evaluated machines' issued explanations ranging from "Not adequate at all" [1] to "Fully adequate" [7] (on the y axis) for three types of machine that were presented to participants in stories (on the x axis). The top and bottom panels in the Figure show the adequacy ratings of participants who evaluated explanations that referred to authority and those that referred to causes respectively.

As previously, we list our research questions using bullets before discussing answers to them. The first two were the following.

- Do people expect a higher standard of explanation from machines described as aware compared to those described as non-aware?
- Does the answer to the first question depend on the type of issued explanation—one that refers to causes versus one that refers to authority?

As the results in Fig. 6 show, participants' ratings of the adequacy of machines' issued explanations follow a slight downward trend for increasing degree of awareness in machines' descriptions. This is particularly so for machines that were described as aware of a relevant factor at hand compared to machines that were described as basic or attuned to context more generally when their issued explanations referred to causes of their decisions (the bottom panel in Fig. 6). While the same (slight) trend appears present, it was certainly weaker when machines' issued explanations referred to authority rather than causes (the top panel in Fig. 6). Thus, with regards to our first two research questions, people appear to expect a (slightly) higher standard of explanation from machines described as aware compared to those described as non-aware when the machines' issued explanations refer to causes of events, but not (or significantly less so) when their explanations refer to authority.

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

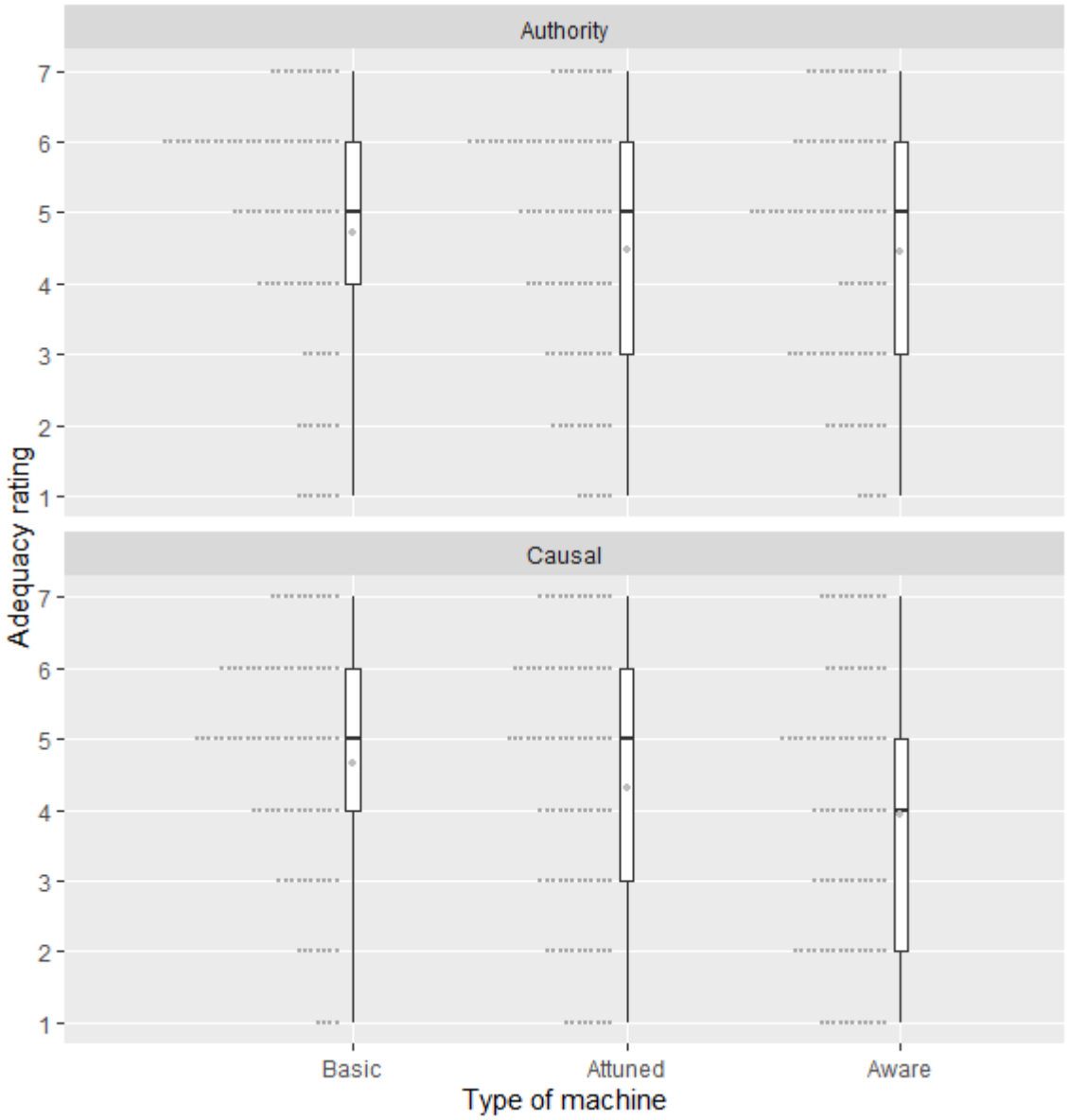

Figure 6. Adequacy ratings. *Y axis: participants' rating of the adequacy of their evaluated machine's issued explanation, ranging from "Not adequate at all" [1] to "Fully adequate" [7]. X axis: type of machine that was presented to participants. Top panel: adequacy ratings of participants who evaluated explanations that referred to authority. Bottom panel: adequacy ratings of participants who evaluated explanations that referred to causes. For each type of machine (column), small dots on the left show distributions of individual responses. The box-plots show the median values (black horizontal lines) as well as the 1st and the 3rd quartiles in individual responses (lower and upper ends of the white boxes). The grey dots on top of the box plots are mean values.*

- Does the answer to the first question depend on the purpose of the issued explanation, e.g., when it is to be used in litigation?

Figure 7 shows participants' ratings of the adequacy of the same issued explanations for use in litigation. Firstly, comparing results in Fig.7 to those in Fig. 6, everything else being equal, participants rated machines' issued explanations as less adequate when those explanations

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

were considered for use in litigation. This was so irrespective of whether explanations referred to causes or to authority. Also, as previously, participants' ratings of the adequacy of evaluated explanations follow a slight downward trend for increasing degree of awareness in machines' descriptions when the machines' explanations referred to causes. However, this was not the case for explanations that referred to authority. Thus, regarding our third research question the answer is mixed. People's ratings of the adequacy of machines' issued explanations are similarly sensitive to the degree of awareness in machines' descriptions when those explanations are considered for use in legal matters as well as outside of the legal context, but overall people deem these explanations as less adequate in litigation.

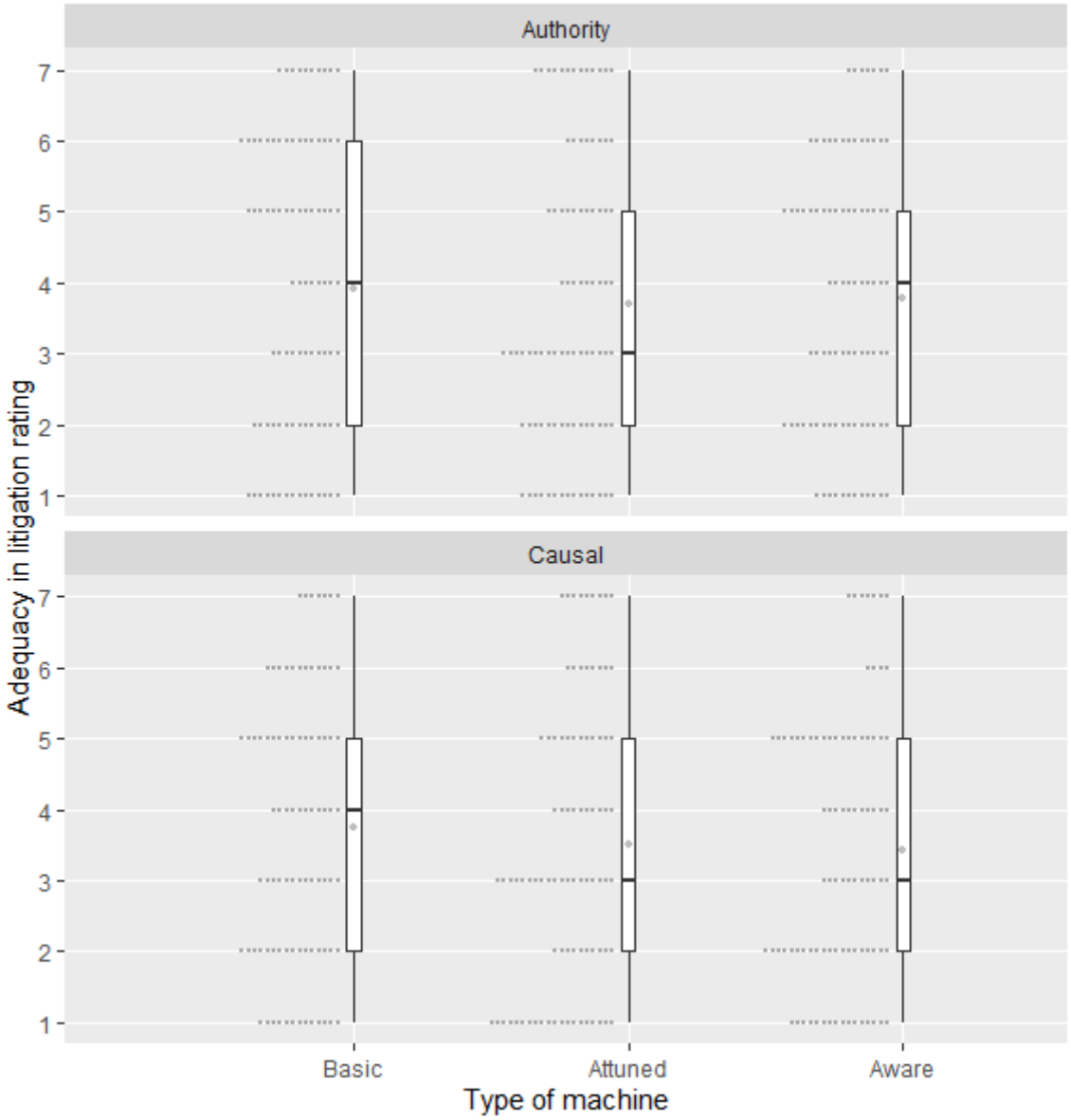

Figure 7. Adequacy ratings for litigation. *Y axis: participants' rating of the adequacy of their evaluated machine's issued explanation for use in litigation, ranging from "Not adequate at all" [1] to "Fully adequate" [7]. X axis: type of machine that was presented to participants. Top panel: adequacy ratings of participants who evaluated explanations that referred to authority. Bottom panel: adequacy ratings of participants who evaluated explanations that referred to causes. For each type of machine (column), small dots on the left show distributions of individual responses.*

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

The box-plots show the median values (black horizontal lines) as well as the 1st and the 3rd quartiles in individual responses (lower and upper ends of the white boxes). The grey dots on top of the box plots are mean values.

4.4 Limitations and future research

As with our first experiment, we offered only descriptive initial insights from our collected data. Further statistical analyses are necessary to corroborate these initial insights, especially because the trends that we observed here are rather subtle. We will report our full statistical analysis of these results and any additional insights that we draw from our dataset in future publications.

5. Summary and practical guidelines

In this section we summarise what stems from our review and initial insights from our conducted experiments in terms of practical guidelines and recommendations for the development, deployment, and regulation of AI systems as well as directions for future research.

Start with the concept of **machine awareness**. One risk with the introduction of this term to describe deployed machines is the danger of potential confusions and misunderstandings that may arise between different stakeholders about what machine awareness actually means to them (Section 2.3). Our initial empirical insights suggest that by and large people agree with the functional account of awareness developed by the project EMERGE: people ascribe higher degrees of awareness to agents that are theorised to have higher degrees of awareness based on their functional descriptions and people seem to broadly agree with the project's proposed idea that awareness can come in degrees (Section 3.3). They also appear to differentiate awareness and consciousness as is assumed in the proposed concept—a distinction about which there is no clear consensus in ongoing academic debates. That is good news for the direction taken by the project EMERGE. Developers, deployers, and regulators of AI systems should also take note that the introduction of this concept to descriptions of deployed and marketed machines can bring about numerous benefits to human dealings with AI systems (Section 2.2) without fearing that this will cause widespread confusion among different parties involved.

That said, one still ought to tread carefully. Our initial findings suggest that people may nevertheless fall into several groups in terms of how they construe awareness. There may be those who construe awareness as something that can come in degrees and those who construe it as an all-or-nothing state of an agent. Also, since in some of our studied cases people ascribed higher degrees of awareness to humans than to machines when both were functionally described in similar ways, we speculate that in certain contexts some people construe awareness as an aspect of consciousness (in line with one prominent view in our reviewed debate in philosophy of mind). This needs to be investigated further. If our hunch is correct, developers, deployers, and regulators of AI systems will need to know what proportions and which people in our societies construe awareness in these different ways. That also calls for proactive and preemptive thinking about how to succinctly, accurately, and unambiguously communicate the developed concept of machine awareness to the broader public in order to ensure that the introduction of this concept benefits all members of our society equally.

Another point that we mentioned, but did not address in our experiments and which therefore needs to be investigated further, is that highly aware machines may come hand-in-hand with

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

heightened demands for ascriptions of responsibility either to those machines themselves or to their deployers. Also, while we focused on potential risks, it is equally important to empirically assess potential benefits of machine awareness as well (Section 2.2). Future research should investigate, for example, whether users of AI systems will indeed be more ready to rely on those systems and delegate tasks to them when provided with information about how aware those systems are about certain things.

In our second study we addressed the topic of **explainable aware AI**. As hypothesised, our initial findings suggest that people are sensitive to the idea of machine awareness and demand a higher standard of explanation from machines described as aware compared to those described as non-aware (Section 4.3). We underscore, however, that this trend appears to be slight and therefore should be taken with a pinch of salt for now while we conduct statistical analyses of our collected data (which we will prepare for future publications). That said, even if this trend is slight, it ought to be taken into account by developers and regulators of (aware) AI systems to prepare for their deployment accordingly.

Our second finding, one that surprised us, is that people's judgements vary with the type of explanation issued by machines. People seem to expect a higher standard of explanation from aware machines compared to non-aware machines when the issued explanations address causes of events, but not, or significantly less so, when the explanations appeal to authority. We speculate that this might be connected to the idea of the human need to be in control of AI (Nyholm 2022, chapter 4). People want AI-powered agents to follow authority, but expect humans placed in similar situations to reason in terms of causal logic. If true, this might be reflected in people's judgments of the appropriateness of machines' issued explanations as we observed. Our speculative hypothesis will have to be investigated further in future research.

Turning to policy and regulation, this latter finding opens a new can of worms. What types of explanation should aware AI systems be constrained to provide in their interactions with humans? We look forward to investigating this and our other questions raised here, and to sharing our findings in future reports.

References

Black, D. 2017. The concept of consciousness and the bogeyman of conflation. *Journal of Consciousness Studies* 24, 28–50.

Block, N. 1995. On a confusion about a function of consciousness. *Behavioral and Brain Sciences* 18, 227–247.

Chalmers, D. J. 1995. Facing up to the problem of consciousness. *Journal of Consciousness Studies* 2, 200–219.

Chalmers, D. J. 2023. Could a large language model be conscious? arXiv preprint arXiv:2303.07103.

Coeckelbergh, M. 2020a. *AI Ethics.* MIT Press.

Coeckelbergh, M. 2020b. Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and Engineering Ethics* 26, 2051–2068.

Danaher, J. 2019. Philosophical case for robot friendship. *Journal of Posthuman Studies* 3, 5–24.

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

Danaher, J. 2021. Welcoming robots into the moral circle: a defence of ethical behaviourism. *Science and Engineering Ethics* 26, 2023–2049.

European Commission 2021. Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts.

Li, D., He, W., and Guo, Y. 2021. Why AI still doesn't have consciousness? *CAAI Transactions on Intelligence Technology* 6, 175–179.

McDermott, D. 2007. Artificial intelligence and consciousness. In: P. D. Zelazo, M. Moscovitch, and E. Thompson (eds.) *The Cambridge Handbook of Consciousness.* Cambridge University Press, 117–150.

Nagel, T. 1974. What is it like to be a bat? *The Philosophical Review* 83, 435–450.

Nyholm, S. 2022. *This Is Technology Ethics: An Introduction.* John Wiley & Sons.

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.* Routledge, 191–213.

Sætra, H. S. 2019. When nudge comes to shove: liberty and nudging in the era of big data. *Technology in Society* 59, 101130.

Schmauder, C., Karpus, J., Moll, M., and Bahrami, B. 2023. Algorithmic nudging: the need for an interdisciplinary oversight. *Topoi* 42, 799–807.

Seth, A. K. and Bayne, T. 2022. Theories of consciousness. *Nature Reviews Neuroscience* 23, 439–452.

Funded by the European Union
European Innovation Council

Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.