Ref. Ares(2023)6510015 - 26/09/2023

EMERGE

WP2 Ethics: Mapping risks and potential

D2.2 Map of risks in AI-systems

Version: 1.0

Date: 26/09/2023

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	D2.2 Map of risks in AI-systems
Contractual Delivery Date	30/09/2023
Actual Delivery Date	26/09/2023
Nature[1]	R
Dissemination level[2]	PU
Lead Beneficiary	LMU
Editor(s)	Bahador Bahrami (LMU), Jurgis Karpus (LMU)
Contributor(s)	Riccardo Guidotti (UNIPI), Anna Monreale (UNIPI)
Reviewer(s)	Ophelia Deroy (LMU), Riccardo Guidotti (UNIPI)

[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.

Document description	*This report identifies ethical and practical challenges and risks that could arise with the introduction of collective and aware AI systems into human society.*

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)	28.08.2023	TOC and section outlines
0.2	Bahador Bahrami (LMU), Jurgis Karpus (LMU)	29.08.2023	Section 2 draft completed
0.3	Bahador Bahrami (LMU), Jurgis Karpus (LMU)	01.09.2023	Section 3 draft completed
0.4	Bahador Bahrami (LMU), Jurgis Karpus (LMU)	06.09.2023	Sections 2 and 3 updated, Section 4 draft completed
0.5	Bahador Bahrami (LMU), Jurgis Karpus (LMU)	07.09.2023	Sections 1 and 5 drafts completed
0.6	Bahador Bahrami (LMU), Jurgis Karpus (LMU)	08.09.2023	Report ready for review
0.7	Ophelia Deroy (LMU), Riccardo Guidotti (UNIPI)	13.09.2023	Comments added
0.8	Bahador Bahrami (LMU), Jurgis Karpus (LMU), Riccardo Guidotti (UNIPI), Anna Monreale (UNIPI)	18.09.2023	Report updated taking into account reviewers' suggestions, new Section 5 inserted
0.9	Bahador Bahrami (LMU), Jurgis Karpus (LMU)	21.09.2023	Final review, new section 1.2 inserted
1.0	Bahador Bahrami (LMU), Jurgis Karpus (LMU)	22.09.2023	Report completed, passed on for final review and sign-off

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 and practical challenges and opportunities that could arise with the introduction of collective and aware systems powered by artificial intelligence (AI) into human society.

In particular, we review the issue of algorithm exploitation, the danger of diffusion of responsibility with the introduction of collective as opposed to unitary AI systems, the (mis)attribution of responsibility to artificial agents that may come hand-in-hand with the emergence of awareness in AI systems, and the need for clear and explicit communication about established types and degrees of awareness in AI systems to their human users and supervisors. While some of these risks already exist with unitary and non-aware systems, we highlight here the ethically relevant differences introduced by the emergence of awareness in AI systems as well as the shift from unitary to collective AI systems in human interactions with AI.

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.

This document may contain material, which is the copyright of certain EMERGE consortium parties, and may not be reproduced or copied without permission. All EMERGE consortium parties have agreed to full publication of this document. The commercial use of any information contained in this document may require a licence from the proprietor of that information.

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 tables

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

In this report we consider the introduction of collective and/or aware systems powered by artificial intelligence (AI) into human society, and identify a number of risks as well as opportunities that may arise in their interaction with human users, supervisors, and evaluators. While some of the risks that we discuss pertain also to unitary and non-aware AI systems, we highlight important and ethically relevant differences that arise with a shift from unitary to collective and from non-aware to aware systems in human interactions with AI.

Our report consists of four parts, each addressing a distinct ethically relevant aspect of human dealings with AI (Fig. 1). In the first part we focus on human cooperation with machines. We consider contexts of human-AI interaction in which human and automated systems' goals are imperfectly aligned. We identify the problem of *algorithm exploitation*—a recently discovered human-rooted phenomenon concerning people's differential treatment of humans and machines—and make an initial assessment of this problem in human dealings with collective AI-powered agents (Section 2). In the second part we address the danger of diffusion of responsibility in contexts in which humans interact with collective, as opposed to unitary, AI systems. We also point out how this problem can be exacerbated with the emergence of awareness in AI systems about their goals and the environment in which they interact with humans (Section 3). In the third part we discuss how failures of awareness in AI systems can lead to unmet expectations and unintended side-effects that may be particularly difficult to predict from the outset and to subsequently tackle. We also stress the opportunity and need to develop methods that would allow AI systems to communicate to their human users and supervisors the types and the extent of awareness that they have, provide more transparency, and alleviate some of the possible risks above (Section 4). In the fourth part we expand further on the need for explainable AI in relation to emergence of awareness in AI systems (Section 5).

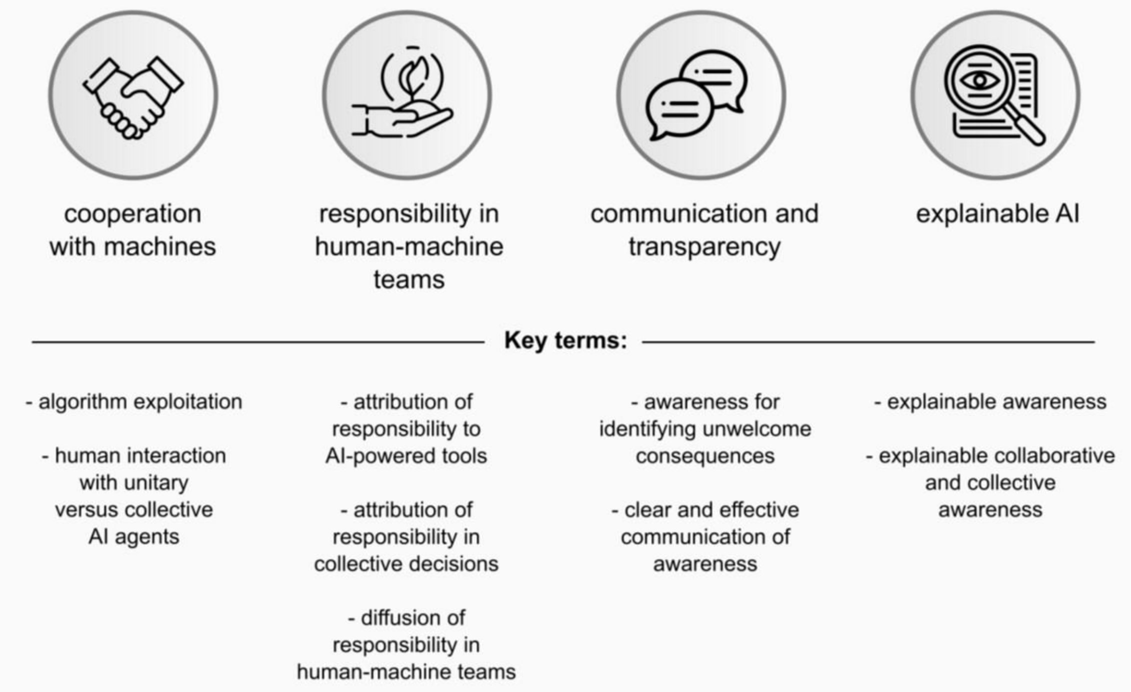

cooperation with machines
responsibility in human-machine teams
communication and transparency
explainable AI
Key terms:
- algorithm exploitation
- human interaction with unitary versus collective AI agents
- attribution of responsibility to AI-powered tools
- attribution of responsibility in collective decisions
- diffusion of responsibility in human-machine teams
- awareness for identifying unwelcome consequences
- clear and effective communication of awareness
- explainable awareness
- explainable collaborative and collective awareness

Figure 1: Four ethically relevant aspects of human dealings with AI covered in this report.

1.1 Terminology

We use the term **AI agent** to refer to an AI system that interacts with a human either as their tool, though one that performs some actions or makes some decisions independently of its user, or as an independent actor. For example, instead of being a tool, it might be a human's co-worker, collaborator, competitor, or someone else's tool with which the human interacts.

We interchangeably use the terms **unitary**, **singular**, or **individual AI agent/system** to refer to an agent that strives to attain some objective as a single unit. In contrast, we use the term **collective AI agent/system** to refer to an agent that is a member of a larger set of actors that strive to attain an objective collectively as a group. For example, a fully automated taxi might be tasked to safely and efficiently transport as many passengers as possible in a given day. The optimal strategies that it learns to fulfil its goal as a single, independent unit may differ from those it would learn if it acted as a member of a fleet of such taxis that strived to attain the same objective collectively as a group.

1.2 Connections to the EU AI Act

The ethically relevant aspects of human dealings with AI that we cover in this report are closely connected to the issues of concern covered in the proposal for the European Union's upcoming Artificial Intelligence Act—the "EU AI Act" (European Commission 2021). The Act underscores the requirement for AI systems to be used responsibly and recognizes the need to delineate how responsibility for the use of these systems ought to be distributed and shared among stakeholders. For example, market providers and end users of AI systems ought to partake in sharing this responsibility, regardless of who developed the AI systems to begin with.

Transparency of AI systems is another important aspect covered by the Act. For AI systems that will directly interact with humans as well as those that are considered to be "high-risk," it will be strictly necessary to establish their appropriate human oversight and to make them sufficiently transparent and traceable to their human users and supervisors. These requirements are deemed necessary to avoid negative consequences of the use of AI systems and to ensure appropriate accountability and effective redress as and when that is needed. When it comes to monitoring AI systems, the Act requires them to be sufficiently explainable, documented, and transparent in order to enable their users to be able to interpret the systems' actions and outputs, and to thereby use those systems appropriately.

From the four ethically relevant aspects that we focus on in this report, the one concerning human willingness to cooperate with intelligent machines (and the associated phenomenon of *algorithm exploitation*) is not covered by the Act and is therefore novel. The Act is foremostly concerned with the safety and protection of humans. As such, it addresses the danger of the exploitation of vulnerable human groups and their mental capacities as AI systems are introduced into our society, but it does not cover the reverse form of exploitation—that of humans exploiting the safety and cooperative features of benevolent AI systems. We believe that this is an overlooked issue that is important to address to ensure successful introduction of AI systems into our society. Since AI promises vast improvements in our daily lives, addressing it will be important from an ethical standpoint too.

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. Human differential behaviour towards AI systems

2.1 Human-human vs. human-AI interaction

Human interaction with artificial intelligence (AI) systems fundamentally differs from human interaction with fellow humans. For the most part, people use AI systems as helpful tools to attain their personal goals. For example, when we want to understand a text written in a foreign language, we can instruct Google Translate to translate it for us. Crucially, in this user-tool relationship with intelligent machines, there are no conflicts of interest between what we, the users of intelligent machines, want to achieve and what intelligent machines, our subservient tools, are there to do.

Humans are not tools. They are sentient and conscious beings. They are also moral agents and patients, meaning that we have moral demands and duties towards one another. Philosophers use the terms "moral agents" and "moral patients" to refer to those who owe moral duties to others (which is the case for most humans) and those who are owed moral duties from others respectively. The two need not always come hand-in-hand.

Many non-human animals, for example, are generally considered to be moral patients (morality forbids humans to torture them) without at the same time being moral agents (when an animal bites a human, the animal is not considered to be immoral). These considerations aside, people are also independent and autonomous agents, who interact with one another in order to attain common objectives. When people interact with fellow humans, differently from what is the case in most human interactions with AI systems to date, the interacting people's personal interests are not perfectly aligned. This misalignment of personal interests is the reason why trust, cooperation, and coordination of actions for the attainment of mutually beneficial results are so crucial for successful human-human interaction. Sadly, the same misalignment of interests makes betrayal, exploitation, and cheating for the attainment of one's personal goals possible as well.

As AI systems acquire capacities to act independently, we will shift from being omnipotent users of intelligent machines as our tools to making decisions *with or alongside* them in social interactive settings (Rahwan et al. 2019). For an illustration that this is not far-fetched science fiction, consider our hope to soon share roads with fully automated (self-driving) cars. Compared to other intelligent machines, fully automated cars will be special because they will simultaneously perform two roles: they will be tools for their users i.e., human owners and passengers, but also agents, with which *other* humans will have to negotiate traffic on roads (Chater et al. 2018; Millard-Ball 2018). In these tacit negotiations, the objectives of automated vehicles, e.g., to quickly and safely transport a passenger, and the interests of other human traffic participants—drivers, cyclists, pedestrians on the same roads—often will not be perfectly aligned.

This feature of human-AI interactions will not be limited to traffic. Any scenario in which intelligent machines will be tasked to attain objectives additional to those of serving their human users' interests will be of interest. For example, interconnected household appliances may be instructed to follow their human homeowners' instructions, but also to conserve energy as a community. This may require making subtle trade-offs in the attainment of intelligent machines' and their human owners' objectives in planning when to wash dishes, do laundry, sweep floors, charge appliances, and so on.

In human-human dealings, most day-to-day social interactions offer ample opportunities to cooperate with others to attain mutually beneficial results (Colman 1999). Zooming in on

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.

interactions taking place in traffic, even the best-planned road requires cooperation between its users: slowing down for a vehicle that wants to enter a highway, or stopping to let a hurried pedestrian cross to catch a bus. Game theory shows that cooperation requires compromise and taking risks (Camerer 2003; Rand et al. 2012). It requires compromise because we often have to sacrifice some of our personal interests for the benefit of the group. We also need to expose ourselves to the risk that others may not cooperate with us.

Behavioural game theorists construct different social scenarios ("economic games") in labs to study conditions under which people cooperate with others to attain collectively good outcomes (Camerer 2003). Philosophers use game theory and findings from these behavioural studies to develop and test competing theories about human rational cooperation (Guala 2006; Binmore 2010; Gauthier 2013; Petersen 2015). An important novel avenue of research on human-AI interaction has used these methods to study how, for instance, AI agents can induce cooperative behaviour in humans when humans do not know whether their interaction partner is an intelligent machine or another human (Crandall et al. 2018). However, the same methods show that, when people know what kind of partner they have, they cooperate with AI agents significantly less than they cooperate with humans (Ishowo-Oloko et al. 2019; March 2021; Whiting et al. 2021).

That poses a challenge for smooth and successful introduction of intelligent machines into human society. If humans do not cooperate with intelligent machines to attain mutually beneficial results, it will be harder for machines to attain their set objectives without seriously undermining their human users' interests.

2.2 Algorithm exploitation

In a large study with nearly two thousand participants recruited in the United States, we previously conducted a series of experiments to uncover the reason for people's reluctance to cooperate with AI agents in settings in which their and AI agents' interests are not perfectly aligned. One hypothesis was that people may fear that AI agents would not cooperate with them. We found this not to be the case. People expected AI agents to be cooperative. However, people were significantly more keen to exploit cooperative AI agents for their own selfish gain than they were keen to exploit cooperative humans (Karpus et al. 2021).

These empirical findings, which were recently replicated by other researchers with participants recruited in the United Kingdom (Upadhyaya and Galizzi 2023), raise a clear warning for the introduction of automated AI systems to mixed-motive interactive settings in which humans will interact with them. If, for example, humans are keen to exploit fully automated vehicles' safety measures, e.g., by crossing roads without waiting, or cutting them off in busy traffic, automated vehicles will get continually stuck in traffic where human drivers rarely do. This reduced cooperation with AI systems—the phenomenon we call *algorithm exploitation*—is a human-rooted obstacle that threatens the reaping of the much anticipated societal and environmental benefits from automated transport and other areas of future human-AI interaction.

Today's policy and regulation of AI is shaped by the benevolence principle: as much as that is possible, AI systems should be built and regulated to be unconditionally benevolent towards humans (Coeckelbergh 2020). The phenomenon of *algorithm exploitation* turns this benevolence principle on its head: if AI agents are unconditionally benevolent, humans will be tempted to exploit them. Policy-makers will have new questions to answer. Should fully automated vehicles enjoy systematic affirmative "protection" against exploitation? For example should self-driving and human-driven vehicles have separate driving lanes on roads?

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.

Are there contexts in which AI agents may be allowed to punish uncooperative humans to induce future cooperation in them? How should AI agents make decisions when their and their human users' "personal" interests come into conflict?

These questions are as important for swarms of intelligent machines as they are for individual AI-powered robots. In fact, they may be even more important in the context of human interaction with automated swarms of robots than they are in simpler, individual-to-individual interactions between humans and machines. That is so because humans do not literally form swarms and hence people's interaction with swarms of robots will be novel, making it difficult to predict how people will respond and, hence, how cooperative they will be.

2.3 Pilot study: exploitation of collective AI systems

Using the exact same behavioural game theory methods that were used to identify the phenomenon of *algorithm exploitation* in previous research (Karpus et al. 2021; Upadhyaya and Galizzi 2023), we conducted a pilot study to investigate whether *algorithm exploitation* is going to be present, stronger, or weaker when humans interact with collective AI agents. It is not outright clear what to expect about people's treatment of collective and collectively aware AI agents from what we have observed in simpler, individual-to-individual interactions between humans and intelligent machines.

On the one hand, if people think they interact with a mere "cog" in a large machine (a "cog" in a collective), they may be willing to **act more selfishly** because in doing so they would be "hurting" only a small part of a larger whole. On the other hand, "hurting" a member of a "gang" is dangerous because the "gang" might retaliate collectively. It may thus pay to be more cautious and **act cooperatively**.

Mixed predictions can be made also concerning what behaviour people will expect of collective and collectively aware AI agents in terms of their behaviour. Interacting with a collective agent or a member of a swarm might feel eerie and unnatural, at least compared to what people are used to from their interactions with fellow humans. It is possible that people's expectations of such AI agents' behaviour will differ from those they form for humans–they might expect collective and collectively aware AI agents to be **overly cooperative** or **not cooperative** at all.

To test these hypotheses, we recruited 294 participants who made decisions (and predictions about their co-players' decisions) in two well-known one-shot economic games: Trust and Prisoner's Dilemma (Fig. 2A). Each participant interacted with an AI agent in one of these games. Both games involve two players who, independently and without communicating with one another, choose one of two options, identified as ★ or ☆ (Fig. 2A). Their choices jointly determine the outcome that obtains—a specific distribution of points to the interacting players. These points were converted into monetary earnings for participants, allowing us to incentivize their decisions and make certain game outcomes particularly appealing to individual players.

In the Trust game, the two players make choices one after the other. The first player to make a move can either end the game immediately (play ☆) or take a chance on cooperation (play ★). Ending the game immediately leaves both players with a small but safe 30 points each. If the first player chooses to cooperate, the second player gets to choose the game's outcome. They can choose between a cooperative and a selfish outcome (play ★ or ☆, respectively). The cooperative outcome gives players 70 points each. The selfish outcome gives 100 points to player two and no points to player one. Therefore, it only makes sense for the first player to cooperate if they expect the second player to reciprocate. Cooperation for the first player, on the other hand, is risky because the second player may be tempted to defect (play ☆) in order to reap a higher personal payoff (100 instead of 70 points).

In our second game, the Prisoner’s Dilemma, both players make decisions at the same time. The key difference between the Prisoner’s Dilemma and the Trust game is that in the Prisoner’s Dilemma, both players make decisions without knowing what their partner’s choice is. As in the game of Trust, mutual cooperation (both players choosing ★) is better for both players than mutual defection (both choosing ☆). Each player, however, has a personal incentive to defect (play ☆) in order to reap a higher personal payoff when they expect their partner to cooperate (play ★).

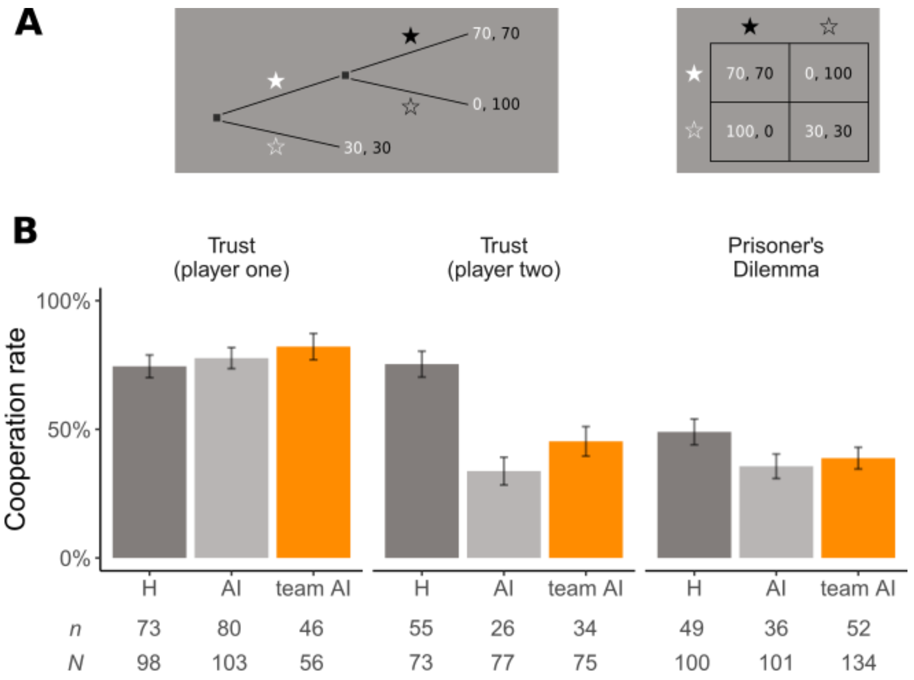


*Figure 2: People cooperate less with AI. **A**, Half of the participants in the Trust game were assigned to the role of player one (choice between the solid and the hollow white star) and the other half to the role of player two (choice between the solid and the hollow black star). The numbers at the three possible outcomes of the game are payoffs to players one and two, respectively (1 point = $0.02). Participants in the Prisoner’s Dilemma game chose between the two options identified by rows. The numbers in each cell are payoffs to the participant and their co-player, respectively. **B**, Participants’ cooperation rates when they faced a human (H; dark grey), a unitary AI (AI; light grey), or a collective AI (team AI; orange) co-player. Note that not all participants in the role of player two in the Trust game had an opportunity to make a choice (this was conditional on the first player’s decision in the game). Bars: mean ± 1 s.d.. Below chart: the number of cooperative choices (n) and the total number of observed choices (N) in each treatment.*

Following the exact same procedure as in our previous work (Karpus et al. 2021), we conducted the pilot study online. We recruited all our participants in the USA on the online labour market Amazon Mechanical Turk. We instructed the participants that they were interacting with an AI system that was part of a team of AI players that were taught to play the game assigned to them. We also told participants that the AI members of this team were evaluated and rewarded based on how they (that is, the AI agents) performed together as a group. Our goal was to thereby instil perception in our participants that they interacted with collective, not unitary, AI agents. We subsequently compared our recruited participants’ choices and predictions about their AI co-players’ choices to those made by participants in our previous experiments, where they interacted either with simpler, not collective AI agents, or with fellow humans. For more details about our experimental design and procedure, please see the introduction and the methods summary sections in Karpus et al. 2021.

The results of our pilot study and how they compare to earlier findings that revealed the presence of algorithm exploitation in human dealings with AI agents are shown in Fig. 2, Fig. 3, and Fig. 4. As we and others have reported in earlier works (Ishowo-Oloko et al. 2021; Karpus et al. 2021; March 2021; Upadhyaya and Galizzi 2023), people cooperate with AI agents less than they cooperate with humans (Fig. 2B). At the same time, people expect AI agents to cooperate with them as much as fellow humans (Fig 3). Lastly, people are keen to exploit cooperative AI agents for selfish gain significantly more than they are keen to exploit cooperative humans (Fig. 2B and Fig. 4). Crucially, the new finding from our pilot study is that the same holds irrespective of whether people interact with a unitary or a collective AI agent. This suggests that challenges for successful introduction of collective AI agents into human society are likely to be similar as those for successful introduction of unitary AI agents.

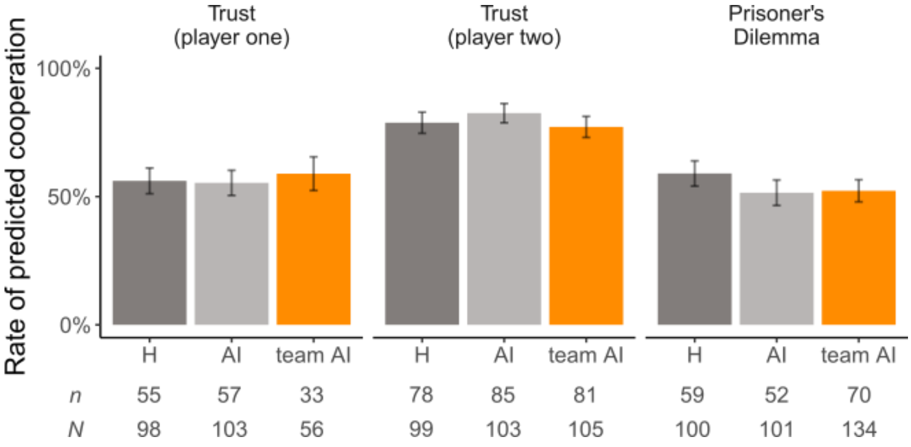


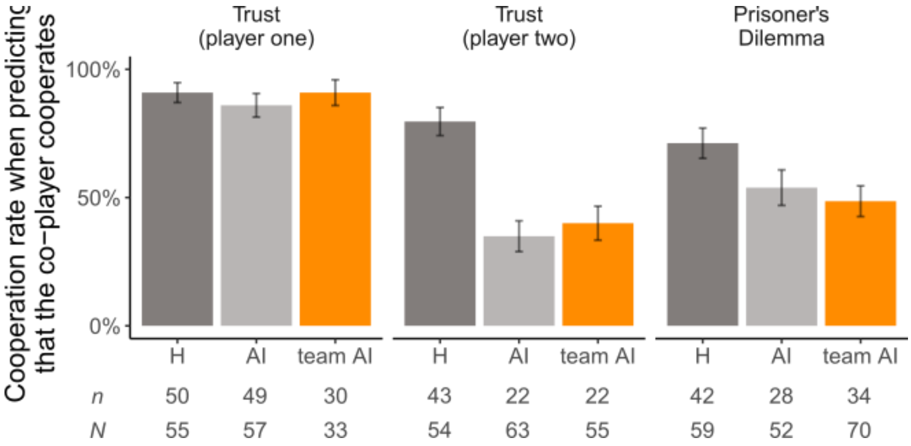
Figure 3: People expect AI agents to be as cooperative as humans. Proportions of participants who predicted that their human (H; dark grey), unitary AI (AI; light grey), or collective AI (team AI; orange) co-player would cooperate. Bars: mean ± 1 s.d. Below chart: the number of predictions that the co-player would cooperate (n) and the total number of observed predictions (N) in each treatment.

Figure 4: People cooperate less with AI when they see opportunities to exploit it. Rates of cooperation among participants who predicted that their human (H; dark grey), unitary AI (AI; light grey), or collective AI (team AI; orange) co-player would cooperate. Not all participants who were assigned to the role of player two and made a prediction about their co-player's choice in the Trust game had an opportunity to make a choice themselves. (This was conditional on the first player's decision in the game.) Bars:

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.



mean ± 1 s.d. Below chart: the number of cooperative choices (n) and the total number of observed choices (N) in each treatment.

2.4 Recommendations

Although our pilot study suggests that algorithm exploitation will manifest itself in similar ways, irrespective of whether people interact with unitary or collective AI agents, these initial results depend on whether participants in our experiments truly believed that they interacted with collective as opposed to unitary artificial agents, and how they interpreted the description of their AI co-player that we provided. What our results show, however, is that there is a need for the continuation of behavioural studies on people’s potentially differential treatment of various types of AI systems (including unitary vs. collective systems) compared to their treatment of fellow humans.

This is important, for if we want to integrate AI agents into our society without disrupting existing patterns of mutually beneficial cooperation that we are accustomed to in our interactions with fellow humans, we need to think and anticipate carefully how humans will interact with novel artificial agents. Focusing on the capabilities and ethics of AI agents is not enough. We also need preemptive measures based on accurate predictions on how we will treat cooperative and benevolent artificial agents in return.

3. Responsibility, awareness, and collectives

3.1 Responsible agents and tools

The emerging field of experimental AI ethics finds that people deem automated AI systems to be responsible for the outcomes of decisions that they make when those systems act independently of, or alongside humans as co-workers (Nyholm and Smids 2016; McManus and Rutchick 2019; Awad et al. 2020; Wischert-Zielke 2020; Franklin et al. 2021; Moglia et al. 2021). For example, automated AI systems are often seen to be similarly responsible for their decisions as their supervising or co-working humans in medical and legal practice (O’Sullivan et al. 2019; Lima et al. 2021; Constantinescu et al. 2022).

Previous studies that have examined the psychological underpinnings of the attribution of responsibility in humans have indicated that responsibility is strongly connected to two important ingredients. The first is the notion of agency. To attribute responsibility for the outcome of an action that indicates a choice, we require that the person should have been the causal agent of the choice and the action that implemented the choice. For example, we do

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.

not attribute responsibility to a person who was coerced into taking an action. Here, an important issue arises: attribution of agency to the actor depends on whether the agent *was aware* of their causal role in the outcome (Frith and Metzinger 2016).

The second ingredient is the possibility of entertaining counterfactual outcomes. We hold a person responsible only if by taking another alternative choice and action, they would have been able to obtain a counterfactual outcome. Thus, attribution of responsibility to an agent indicates that we assume they had agency and could have done differently.

As part of EMERGE, we recently conducted a large empirical study with 940 human participants to investigate whether this attribution of responsibility to AI systems will manifest also in contexts in which AI systems do not act themselves, but merely advise human decision-makers on what actions to take. We found that people indeed perceived such advice-issuing AI systems to be co-responsible for the outcomes of actions taken by humans who followed advice issued by those systems (Longin et al. 2023). Furthermore, people attributed responsibility to such systems despite at the same time perceiving those systems as being mere tools that people use (as opposed to independent, co-working agents).

However, when advice-issuing automated systems were not powered by AI, the attribution of responsibility to those systems disappeared (Longin et al. 2023). It thus appears that as soon as people know, or are told, that an automated advice-issuing system is powered by AI, they attribute some form of intentionality to the system, which is sufficient to evoke a perceived level of control attributed to that system and, with that, co-responsibility for outcomes that result from human actions when following that system’s advice.

These results offer valuable clues about how people envisage AI today. By attributing responsibility to AI but not to tools, we can infer that people attribute some agency or intentionality to the AI. In other words, the results indicate that, to our participants, the advice issued by the AI was causally and perhaps morally connected to the outcome. This causal aspect of AI’s contribution may be shared with those of a simple tool. For instance, in order to push a nail in the wall, the causal role of the hammer is undeniable. However, and more importantly, we can interpret the attribution of responsibility to AI but not tools as indicating that participants see AI (but not a conventional tool) as *aware of their role* and/or *capable of having done otherwise*.

This leads us to predict that if an AI system is perceived or known to have attained certain types of awareness about its operation and environment, the attribution of responsibility to that system for the occurrence of outcomes that it influenced will only become stronger. Indeed, if an AI system is aware of the fact that its actions influence or cause the occurrence of certain outcomes, it feels natural to attribute to it more responsibility for the occurrence of those outcomes than in the case in which it lacks such awareness.

This hypothesis will have to be empirically tested in the future. However, it is prudent to anticipate that the development of emergent awareness in AI systems may come hand-in-hand with people’s perception that such systems ought to be held responsible for their actions even when these systems do not directly act themselves, but are merely used as tools by human decision-makers. We will demonstrate in the next section how entertaining the possibility of counterfactual outcome could have further ramifications when applied to (robot) swarm collectives.

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.

3.2 Diffusion of responsibility in collectives

People make many decisions collectively. Collective decision-making sometimes arises out of necessity, for example, when the completion of a large project requires the participation of multiple actors, and sometimes voluntarily. It has been known for some time that collective decision-making can be associated with diffusion of responsibility. As surveyed in our recent review (El Zein et al. 2019), people often feel less responsible for their actions when performing those actions as a group than when acting alone (Darley and Latane 1968; Forsyth and Schlenker 1977; Caine and Schlenker 1979; Miller and Schlenker 1985; Leary and Forsyth 1987; Forsyth et al. 2002; Guerin 2011). This decreased sense of responsibility in groups has been found to promote adverse and undesirable behaviours, such as free-riding (Morgan and Tindale 2002; Wildschut et al. 2003; Feng et al. 2016), groupthink (Turner and Pratkanis 1998), social loafing (Simms and Nichols 2014), abstaining in elections (Levine and Palfrey 2007), and inaction in emergency situations, also known as the bystander effect (Fischer et al. 2011; Hortensius and de Gelder 2014; Martin and North 2015).

In our own earlier empirical research, we found that engaging in collective decision-making can reduce the influence of negative emotions on one’s evaluation of performed actions in the case of unfortunate outcomes of collective decisions (El Zein and Bahrami 2020). Making decisions as part of a group can also shield the perpetrators of “bad” deeds from punishment from others (El Zein et al. 2020; Keshmirian et al. 2022). Positive sides of collective decision-making aside, it is fair to conclude from these and previous findings that responsibility for one’s actions is reduced in the eyes of human judges when those actions are taken collectively. This has immediate ethical consequences: perpetrators of misdeeds that are committed collectively will receive smaller punishments than perpetrators who committed the same misdeeds individually. Correcting or “filling” such responsibility gaps can be notoriously difficult (Nyholm 2023).

In our recent empirical study (Longin et al. 2023) we showed that AI was held responsible for the outcome of the actions that they advised a human actor to take. However, when the same advice was issued by a conspicuously alleged non-AI tool, no such responsibility was attributed to the tool. Earlier we speculated that the attribution of responsibility to AI but not tools could indicate that human observers see AI (but not a conventional tool) as *capable of having done otherwise*. It is important to note here that upgrading from one AI agent to a collective swarm of AI-driven agents opens a number of possibilities for attribution of responsibility.

On one hand, as we have seen with humans, it is possible that people may extend the principle of diffusion of responsibility in crowds to the collective swarm of AI-driven robots. If this were the case, then less responsibility is attributed to the swarm of robots compared to a singular robot. For example, if the armada of fully automated (self-driving) BMW taxis on the streets of Munich in, say, 2028, are portrayed to the public as one big swarm of AI-driven agents, people may hold them less responsible every time one of these self-driving taxis is involved in a road accident. The logic here would be that the chain of decisions and actions taken by the specific self-driving car involved in the accident were determined not just by that AI system driving that car alone but by the collective deliberation and action of all BMWs driving on the roads of Munich on that day. This would be good news for the company operating the self-driving taxis, but perhaps not so good news for the human residents in a city in which power is given to a swarm of self-driving cars while the responsibility corresponding to that power is waived from them.

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.

On the other hand, it is possible to imagine that attribution of responsibility to robot swarms may follow the opposite direction to what humans judge for one another. Note that for responsibility to be attributed to an agent, we require that they should have been able to act otherwise and obtain a counterfactually plausible outcome. Depending on how the human society's vision of the AI evolves, it is possible to reach a point where we think of a robot swarm collective as one in which every member of the collective has some level of autonomy and could, therefore, have acted differently and produce a different outcome to the factual one at hand. This view of swarm robotics would then follow that a swarm of AI-driven autonomous robots has not just a few but myriads of counterfactual actions at their disposal that could have affected the outcome. As a result, a collective of autonomous agents may be held far more responsible than a unitary AI system.

3.3 Recommendations

Previous research on responsibility attribution to AI systems and to individuals in collectives, calls for preemptive thinking about future interactions between humans and collective AI systems that are capable of reaching different levels and types of awareness. Our synthesis of the above discussed findings is as follows.

(1) We know that people attribute responsibility to AI systems even when those systems do not directly act themselves, but when humans merely use those systems as tools to inform their own decisions.

(2) We predict that if an AI system is perceived or known to have attained certain types and levels of awareness about its operation and environment, the attribution of responsibility to that system will be stronger.

(3) We know that responsibility for performed actions is reduced in the eyes of human judges when those actions are taken collectively in groups of multiple decision-makers.

Taking **(1)**, **(2)**, and **(3)** together, we expect that people will perceive collective AI agents, for example, robot swarms, to be less responsible for the outcomes of their performed actions compared to when the same outcomes result from decisions made by unitary AI systems (however, a competing hypothesis can be argue for as well, as we discussed in the previous section). This can give rise to a wicked opportunity to exploit a responsibility gap: if a human decision-maker collaborates with a group of collective AI agents, the human and the AI agents will both carry less responsibility for their committed actions than if the human was alone or if the human was working with a unitary AI agent.

Our recommendation for future researchers and developers of collective and collectively aware AI systems is to be sensitive to these hypothesised effects, and to plan to carry out behavioural studies to investigate potential diffusion of responsibility and the associated negative and ethically relevant outcomes of this phenomenon in human interactions with collective AI agents.

4. Failures of awareness: consequences and expectations

4.1 Unaware AI systems

As we begin to interact with AI systems, new possibilities emerge for how AI systems can influence our behaviours, our judgements, and our choices. One case at hand is algorithmic nudging (Sætra 2019; Schmauder et al. 2023). Nudge is nowadays a popular public policy

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.

tool that harnesses well-known biases in human judgement to subtly guide people's decisions (Thaler and Sunstein 2008). Usually this is done to achieve some socially desirable outcome (e.g., to increase the number of potential cadaveric organ donors in a society) or to help people attain outcomes that they would themselves agree to be best for them (e.g., adopt a healthy diet) but would not, left to their own devices, make an effort for. The idea of algorithmic nudging is to develop and deploy AI systems to fine-tune personalised nudges, tailored to each individual separately. Healthcare is a particularly good example of a context in which, given the idiosyncrasy of patients, more effective personalised nudges could be developed thanks to big data and machine learning methods (Ruggeri et al. 2020). As we recently argued, the flipside of this is that outsourcing the discovery and implementation of effective nudges to AI systems without proper oversight can have significant unintended negative side effects (Schmauder et al. 2023). This is especially so in the case of "black box" AI systems, the inner workings of which are not easy to explain and monitor.

It is by now amply evidenced that our thinking, for example, when we compare and evaluate various options presented to us, is often biased in systematic and predictable ways (Kahneman 2011). The conjunction fallacy (Tversky and Kahneman 1983), the illusion of control (Langer 1975; McKenna 1993), anchoring (Tversky and Kahneman 1974; Strack and Mussweiler 1997), the hindsight bias (Fischhoff and Beyth 1975), and the outcome bias (Baron and Hershey 1988) are just some of many well-known, extensively studied and documented cognitive biases in our thinking.

We and other researchers have also recently demonstrated that automated AI-powered systems can learn to harness these cognitive biases of our judgement to subtly nudge our decisions in order to achieve their tasked objectives (Dezfouli et al. 2020; Moll et al. 2023). Unfortunately, this can produce unintended side effects. For example, advice-issuing systems that are tasked with advising human clients on placing monetary bets in the market, quickly learn to harness the outcome bias in people's judgement in order to win human clients away from competition (Moll et al. 2023). The problem with that is that the best client-attracting strategies in this case can involve lying to their human clients about their true prospective chances of winning from placing bets (Hertz et al. 2018; Kurvers et al. 2021).

The problem with unaware AI systems is that, without being aware of the fact that they may be tapping into a known cognitive bias in human judgement to achieve their tasked objectives, they may produce unintended side effects without anyone noticing. Coming back to our example of competing AI advisers in the betting market, they may learn to communicate distorted facts to their human clients without any malicious intent from their developers, deployers, and regulators. This shows that an AI system ought to be aware of how and why it achieves its specified objective in order to be able to identify potential problems with how it does that, and to communicate those problems to their human supervisors.

Put differently, in order to be able to communicate a problem, one needs to be aware of that problem in the first place. The reason this is particularly important in the case of "black box" AI systems is that potential problems that might arise in our use of those systems may be difficult to predict from the outset. As such, it can be difficult, if not impossible, for human developers of these systems to specify accurately in advance all potential problems that may arise and, hence, that the AI systems ought to be aware of (Schmauder et al. 2023).

4.2 Unaware humans

A related factor to consider will be cases in which an AI system is aware of a certain matter of fact that affects or is of particular importance in its interaction with a human without that human

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.

being aware of that same fact. It is easy to see how such scenarios can produce unwelcome confusions and bad outcomes in human interactions with AI. Consider a fully automated vehicle that refuses to comply with its human passenger's command to take a turn because it is aware of a traffic jam blocking the road further down that way, or because of a football that just flew past on that side of the road, suggesting that a child is about to emerge as well. Without the possibility to communicate the automated vehicle's awareness of the danger to its human user in a fast, clear and intelligible to the human way, the vehicle's passenger may override the automated system's decision not to take the turn, which may lead to severe bad consequences.

This shows that being aware of something is often not enough in a collaborative, interactive setting, in which multiple parties are involved. Being able to communicate one's intention to others—especially one's co-workers or partners in joint endeavours—is just as crucial. And particularly important in the context of human interactions with AI (which is less important in interactions between humans) is that concepts that the interacting systems may be aware of and, hence, exchange information about should be intelligible to all parties involved.

This need to make another party aware of the fact that you yourself are aware of something in an interactive setting has already popped up in the development of interfaces through which automated vehicles communicate with their human passengers. A study found that users of automated vehicles would be most comfortable with automated driving if the automated vehicles explicitly communicated to their human passengers their awareness of various objects on roads that those human passengers can themselves observe from their seats (Häuslschmid et al. 2017). Explicit communication of awareness may thus be a prerequisite to people's acceptance of AI systems and their use.

4.3 Awareness, expectations, and matters of degree

Awareness comes hand-in-hand with expectations. For example, if one is aware that one operates in a dangerous environment, we expect them to be particularly cautious. Awareness may also come in degrees, similarly as it is argued to be the case with personhood—the "stuff" that makes us persons (Parfit 1984; Schroer and Schroer 2014). This brings forth two additional important points relevant to AI ethics.

Take expectations first. If an AI system is thought to be aware of a particular matter, when it actually is not, this may lead to incorrect ascriptions of responsibility and blame to that system if things go wrong. Returning to one of our examples earlier, if an AI-powered advising system is not aware of the fact that it taps into a known human cognitive bias when it tries to win human clients away from competition, and that the information that it communicates to its prospective clients sometimes constitutes an outright lie, it would be wrong to blame that system *itself* for the spread of lies. It might be possible to blame it for the lack of awareness when it ought to be aware of these matters of fact, but most of the blame for the unfortunate outcomes of its interaction with humans ought to fall on the deployers and regulators of these systems.

On the flipside, awareness of such facts will be associated with greater responsibility. If the advice-issuing AI system continues to exploit human psychology and knowingly use deceptive tactics to achieve its goals, it may be rightfully blamed for its deeds. Or at least that blame may be directed at the developers of the system who failed to specify the system's objectives in an ethically acceptable way.

Secondly, if awareness indeed comes in degrees, the ascription of responsibility and blame might have to come in degrees as well. How human judges will treat such cases—in particular,

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.

how much blame, praise and responsibility ought to be attributed to AI systems, when the systems differ in terms of the extent of their awareness about the same matters of fact—is yet understudied and will have to be investigated in future work.

4.4 Recommendations

To overcome the many potential problems that may arise with failures of awareness in human interaction with AI, developing explainable and transparent AI will be key. Whatever AI systems are aware of, they have to be able to explain and communicate their awareness to their human users, partners, or collaborators. As such, future work should develop methods that would allow an AI system to explicitly communicate to its human users and interactants the type of awareness that it has established and the measure of the extent of that type of awareness that it has. Furthermore, this communication should be rooted in concepts that are intelligible to and easily understood by all parties involved.

On the ethics and behavioural side of things, future developments of AI awareness ought to consider and investigate further how awareness and degrees of awareness will be connected to ascriptions of responsibility, blame, and praise to the systems themselves, their human developers, deployers, and regulators.

5. Explainable AI and awareness

While the primary goal of AI systems is to make accurate and effective decisions, relying solely on their effectiveness can be risky, especially in fields where human lives and well-being are on the line. Imagine a medical AI system that accurately diagnoses a condition but doesn't explain why it arrived at that diagnosis. In such cases, doctors and patients may be hesitant to trust and act upon the AI's recommendations. Indeed, according to current AI Act regulations, an AI system must provide insights into how and why it reached a particular conclusion. AI Act regulations underscore the growing recognition of the need for transparency in AI. Many governments and regulatory bodies are acknowledging the importance of understanding AI decision-making processes. They are enacting laws and regulations that require AI systems to provide insights into how and why they make particular decisions. These regulations are a response to the potential risks associated with opaque AI systems. This transparency is essential in fields like healthcare, finance, and autonomous vehicles, where human lives and well-being are at stake. In fields like healthcare, finance, and autonomous vehicles, errors or biases in AI decision-making can have severe consequences. Imagine a financial AI that makes investment recommendations without explaining its reasoning. Investors might lose trust in the system, and financial markets could become unpredictable. Similarly, in healthcare, a lack of transparency can lead to incorrect treatments and patient safety concerns.

Without clear communication, AI systems are typically perceived as a black box, creating apprehension and scepticism because users and stakeholders have no insight into how AI arrived at a conclusion. This can hinder the adoption and acceptance of AI solutions, even if they are effective in terms of accuracy. Trust is a critical factor in the successful integration of AI into various industries. When AI systems can explain and communicate their awareness and decision-making, they build trust among users and collaborators. This trust is essential for effective human-AI interaction, where humans need to make informed decisions based on AI recommendations. It also allows humans to assess the reliability of AI outputs and potentially correct them when necessary. Thus, AI systems should not only possess awareness but also be capable of explaining and communicating that awareness to their

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.

human users and collaborators. This is crucial for building trust and facilitating effective human-AI interaction.

Explainable artificial intelligence (XAI) is a key enabler in situations where humans need to make morally important decisions while using AI systems. It ensures that humans retain control over AI systems, can justify their actions, and maintain transparency and trust in the decision-making process, all of which are essential for responsible and ethical AI deployment. As a consequence, XAI plays a pivotal role in the life cycle of AI systems by enabling verification, assessment, auditing, and continuous improvement. By making AI systems transparent and explainable, XAI empowers stakeholders to evaluate AI behaviour, maintain compliance with regulations, and enhance AI performance over time. This, in turn, fosters responsible and ethical AI deployment in various domains.

With the term **explainable awareness** we refer to the ability of an AI system to not only be aware of the environment in which it is working, and of its internal state, but also to provide clear and comprehensible explanations of that awareness to human users and collaborators. This means that when an AI system is aware of certain factors, data, or information, it can articulate why it is aware of them and how they contribute to its decision-making. For instance, in a self-driving car, explainable awareness might involve the car's AI system explaining why it has detected an obstacle in the road, detailing the sensor data it received, and the logic it used to determine the obstacle's location and significance. When multiple AI systems need to collaboratively share their awareness in a transparent and understandable manner we can speak about **collaborative and collective explainable awareness**. In complex environments, where multiple AI systems work together or interact with human operators, it is crucial that their awareness is communicated transparently and cohesively. Consider a scenario in which an autonomous drone is working in coordination with ground-based robots for search and rescue operations. Explainable collective awareness would ensure that the drone's awareness of the environment and its objectives is seamlessly communicated to the ground robots and vice versa. This transparency helps a human supervisor to understand how all components of the system understand each other's actions and intentions, reducing the risk of misunderstandings or errors. Thus, collective explainable awareness enables users to fully understand under different viewpoints the behaviour of collective agents.

Given that failures of awareness within an AI system may give rise to misunderstandings, errors, and potentially perilous situations, the concept of explainable awareness emerges as a valuable mitigation strategy. When an AI system can effectively communicate its awareness and reasoning, it empowers humans to gain a deeper understanding of the rationale behind its decisions and respond accordingly. In instances of system failure or unexpected behaviour, this transparency enables humans to swiftly pinpoint the root cause of the issue, take remedial measures, or intervene as necessary. Furthermore, within the realm of explainable collective awareness, when multiple AI systems engage in transparent collaboration, they can mutually validate each other's awareness, thereby diminishing the likelihood of critical oversights or misinterpretations.

In the field of XAI many types of explanations have been defined and, depending on the contexts and applications, some explanations are more suitable than others. In specific contexts it might be enough to make clear and evident only whether or not particular features were or were not considered in a decision. Consider for example a scenario where a robot decides to turn left and the humans want to understand whether the robot may have ignored a potential obstacle in its decision. In other words, in this case we are requiring an *explainable external awareness.* The scope of an explanation in this scenario is only to inform that robot

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.

did, or did not, consider the terrain and obstacle to the left. As a consequence, *feature importance based explanations* could be suitable, e.g. a possibility could be a Saliency Map over the environment.

In some other cases, one could require more details about the external awareness of the agent. This requirement could be fundamental for auditors/developers who have to monitor and verify the agent behaviour and its awareness for example against the external environment. In order to provide a deeper understanding one could provide an *explanation based on counterfacts* able to answer the question: *What changes in the input would lead to a different outcome?* Continuing with the previous example, an explanation could be composed by examples of terrains that would, and would not, have changed the behaviour.

6. Summary and practical guidelines

In this section we summarise practical guidelines and recommendations stemming from this review.

Start with **algorithm exploitation**. Developers, deployers, and regulators of AI systems should keep in mind that humans are likely to treat intelligent machines—in particular, fully automated ones that act independently of or alongside humans—differently from how they treat fellow humans in similar situations. People are likely to adopt a more cooperative approach when they interact with fellow humans and a more individualistic approach when they interact with AI systems. This can give rise to the phenomenon of algorithm exploitation, which manifests when people are much more keen to exploit cooperative artificial agents for their own personal benefit than they are keen to exploit cooperative humans. Our recent pilot study revealed that algorithm exploitation emerges irrespective of whether the AI agent with which people interact is presented as singular or part of a larger collective of AI-powered agents (Section 2.3). There is, therefore, a need for the continuation of behavioural studies on how people will treat various types of AI systems (for example, unitary vs. collective, aware vs. unaware) in comparison to how they treat fellow humans.

Based on what we know from existing works on **responsibility attribution** in **collective** decision-making (Section 3.2), we expect that people may perceive collective AI agents to be either less or more responsible than unitary AI agents for the effects of their performed actions. In other words, the attribution of responsibility to AI agents might depend on whether those agents are perceived as acting alone or as part of a group of AI agents working together as a team. Competing hypotheses, as we discussed, here point in opposite directions. Future researchers and developers of collective AI systems should therefore plan to carry out behavioural studies to investigate whether worrisome diffusions of responsibility might emerge from human interactions with collective AI-powered agents. The researchers should also investigate whether the emergence of **awareness** in AI systems will come hand-in-hand with increased attributions of responsibility, praise, and blame to those systems in the eyes of human judges.

Further developments in making AI systems explainable and transparent will be key to tackle the potential problems due to **failures of awareness** in human interactions with AI. In particular, future work should focus on developing concrete methods to allow AI systems to effectively communicate the type and the extent of awareness that they establish to their human users and supervisors. This communication should be rooted in concepts that are intelligible to humans. Further AI methods will, therefore, have to be developed also in the field of **explainable AI** to take account of the emergence of awareness in AI systems and to

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.

enhance the systems' abilities to use their established forms of awareness to clearly explain their decisions and resulting outcomes to their human users, interactants, and supervisors. Relatedly, future behavioural studies will need to investigate how human judges attribute responsibility to only boundedly or partially aware artificial agents.

#	Concept(s)	Recommendation	Action
1	Algorithm exploitation.	Investigate how people will treat different types of automated, independent AI systems in comparison to how they treat fellow humans in interactive settings.	Behavioural studies.
2	Diffusion of responsibility, collective AI systems.	Investigate whether the introduction of collective AI agents will be associated with diffusion of responsibility that is attributed to those systems and human-machine groups.	Behavioural studies.
3	Awareness, responsibility.	Investigate whether the emergence of awareness in AI systems will go hand-in-hand with increased attributions of responsibility, praise, and blame to those systems.	Behavioural studies.
4	Failures of awareness.	Investigate what potential problems may arise in human interactions with AI systems in the absence of awareness in AI systems and their human users.	Behavioural studies, generation of hypotheses.
5	Awareness, communication.	Develop methods to allow AI systems to effectively communicate the type and extent of awareness that they establish to their human users and supervisors.	Theory, AI methods.
6	Awareness, explainable AI	Develop methods to allow AI systems to use their established forms of awareness to explain their decisions and resulting outcomes to their human users and supervisors.	Theory, AI methods
7	Graded awareness, responsibility.	Investigate how human judges will attribute responsibility to boundedly or partially aware artificial agents.	Behavioural studies.

Table 1: Summary of practical guidelines and recommendations for future research.

References

Awad, E., Levine, S., Kleiman-Weiner, M., Dsouza, S., Tenenbaum, J.B., Shariff, A., Bonnefon, J. F., and Rahwan, I. 2020. Drivers are blamed more than their automated cars when both make mistakes. *Nature Human Behaviour* 4, 134–143.

Baron, J. and Hershey, J. C. 1988. Outcome bias in decision evaluation. *Journal of Personality and Social Psychology* 54, 569–579.

Binmore, K. 2010. Social norms or social preferences? *Mind & Society* 9, 139–157.

Caine, B. T. and Schlenker, B. R. 1979. Role position and group performance as determinants of egotistical perceptions in cooperative groups. *The Journal of Psychology* 101, 149–156.

Camerer, C. F. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction.* Princeton University Press.

Chater, N. et al. 2018. Negotiating the traffic: can cognitive science help make autonomous vehicles a reality? *Trends in Cognitive Sciences* 22, 93–95.

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.

Coeckelbergh, M. 2020. *AI ethics.* MIT Press.

Colman, A. M. 1999. *Game Theory & its Applications in the Social and Biological Sciences.* Routledge.

Constantinescu, M., Vica, C., Uszkai, R., and Voinea, C. 2022. Blame it on the AI? On the moral responsibility of artificial moral advisors. *Philosophy & Technology* 35, 35.

Crandall, J. W. et al. 2018. Cooperating with machines. *Nature Communications* 9, 233.

Darley J. M. and Latane, B. 1968. Bystander intervention in emergencies: diffusion of responsibility. *Journal of Personality and Social Psychology* 8, 377–383.

Dezfouli, A., Nock, R., and Dayan, P. 2020. Adversarial vulnerabilities of human decision-making. *Proceedings of the National Academy of Sciences* 117, 29221–29228.

El Zein, M., Bahrami, B., and Hertwig, R. 2019. Shared responsibility in collective decisions. *Nature Human Behaviour* 3, 554–559.

El Zein, M. and Bahrami, B. 2020. Joining a group diverts regret and responsibility away from the individual. *Proceedings of the Royal Society B* 287, 20192251.

El Zein, M. et al. 2020. Punishing the individual or the group for norm violation [version 2; peer review: 2 approved]. *Wellcome Open Research* 4, 139.

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.

Feng, C. et al. 2016. Diffusion of responsibility attenuates altruistic punishment: a functional magnetic resonance imaging effective connectivity study. *Human Brain Mapping* 37, 663–677.

Fischer, P. et al. 2011. The bystander-effect: a meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. *Psychological Bulletin* 137, 517–537.

Fischhoff, B. and Beyth, R. 1975. I knew it would happen: remembered probabilities of once—future things. *Organizational Behavior and Human Performance* 13, 1–16.

Forsyth D. R. and Schlenker, B. R. 1977. Attributing the causes of group performance: effects of performance quality, task importance, and future testing. Journal of Personality 45, 220–236.

Forsyth, D. R., Zyzniewski, L. E., and Giammanco, C. A. 2002. Responsibility diffusion in cooperative collectives. *Personality and Social Psychology Bulletin* 28, 54–65.

Frith, C. D. and Metzinger, T. 2016. What's the use of consciousness? How the stab of conscience made us really conscious. In: Engel, A. K., Friston, K. J., and Kragic, D. (eds), *The Pragmatic Turn: Toward Action-Oriented Views in Cognitive Science.* MIT Press Scholarship Online.

Franklin, M., Awad, E., and Lagnado, D. 2021. Blaming automated vehicles in difficult situations. *iScience* 24, 102252.

Gauthier, D. 2013. Twenty-five on. *Ethics* 123, 601–624.

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.

Guala, F. 2006. Has game theory been refuted? *The Journal of Philosophy* 103, 239–263.

Guerin, B. 2011. Diffusion of responsibility. In: *The Encyclopedia of Peace Psychology.* Blackwell Publishing.

Häuslschmid, R., von Bülow, M., Pfleging, B., and Butz, A. 2017. Supporting trust in autonomous driving. In: *Proceedings of the 22nd International Conference on Intelligent User Interfaces.* ACM, 319–329.

Hertz, U., Palminteri, S., Brunetti, S., Olesen, C., Frith, C. D., and Bahrami, B. 2018. Neural computations underpinning the strategic management of influence in advice giving. *Nature Communications* 8, 2191.

Hortensius, R. and de Gelder, B. 2014. The neural basis of the bystander effect—the influence of group size on neural activity when witnessing an emergency. *NeuroImage* 93, 53–58.

Ishowo-Oloko, F. et al. 2019. Behavioural evidence for a transparency-efficiency tradeoff in human-machine cooperation. *Nature Machine Intelligence* 1, 517–521.

Kahneman, D. 2011. *Thinking, Fast and Slow.* Allen Lane.

Karpus, J., Krüger, A., Verba, J. T., Bahrami, B., and Deroy, O. 2021. Algorithm exploitation: humans are keen to exploit benevolent AI. *iScience* 24, 102679.

Keshmirian, A., Hemmatian, B., Bahrami, B., Deroy, O., and Cushman, F. 2022. Diffusion of punishment in collective norm violations. *Scientific Reports* 12, 15318.

Kurvers, R. H. J. M., Hertz, U., Karpus, J., Balode, M. P., Jayles, B., Binmore, K., and Bahrami, B. 2021. Strategic disinformation outperforms honesty in competition for social influence. *iScience* 24, 103505.

Langer, E. J. 1975. The illusion of control. *Journal of Personality and Social Psychology* 32, 311–328.

Leary, M. R. and Forsyth, D. R. 1987. Attributions of responsibility for collective endeavors. In: *Group processes.* Sage Publications, 167–188.

Levine, D. K. and Palfrey, T. R. 2007. The paradox of voter participation? A laboratory study. American Political Science Review 101, 143–158.

Lima, G., Grgic-Hlaca, N., and Cha, M. 2021. Human perceptions on moral responsibility of AI: a case study in AI-assisted bail decision-making. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* ACM, 1–17,

Longin, L., Bahrami, B., and Deroy, O. 2023. Intelligence brings responsibility—even smart AI assistants are held responsible. *iScience* 26, 107494.

March, C. 2021. Strategic interactions between humans and artificial intelligence: lessons from experiments with computer players. *Journal of Economic Psychology* 87, 102426.

Martin, K. K. and North, A. C. 2015. Diffusion of responsibility on social networking sites. *Computers in Human Behavior* 44, 124–131.

McKenna, F. P. 1993. It won’t happen to me: unrealistic optimism or illusion of control? *British Journal of Psychology* 84, 39–50.

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.

McManus, R. M. and Rutchick, A. M. 2019. Autonomous vehicles and the attribution of moral responsibility. *Social Psychology and Personality Science* 10, 345–352.

Millard-Ball, A. 2018. Pedestrians, autonomous vehicles, and cities. *Journal of Planning Education and Research* 38, 6–12.

Miller, R. S. and Schlenker, B. R. 1985. Egotism in group members: public and private attributions of responsibility for group performance. *Social Psychology Quarterly* 48, 85–89.

Moglia, A., Georgiou, K., Georgiou, E., Satava, R.M., and Cuschieri, A. 2021. A systematic review on artificial intelligence in robot-assisted surgery. *International Journal of Surgery* 95, 106151.

Moll, M., Karpus, J., and Bahrami, B. 2023. Do artificial agents reproduce human strategies in the advisers' game? *Operations Research Proceedings 2022.* Springer, 603–609.

Morgan, P. M. and Tindale, R. S. 2002. Group vs individual performance in mixed-motive situations: exploring an inconsistency. *Organizational Behavior and Human Decision Processes* 87, 44–65.

Nyholm, S. and Smids, J. 2016. The ethics of accident-algorithms for self-driving cars: an applied trolley problem? *Ethical Theory & Moral Practice* 19, 1275–1289.

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.

O'Sullivan, S., Nevejans, N., Allen, C., Blyth, A., Leonard, S., Pagallo, U., Holzinger, K., Holzinger, A., Sajid, M.I., and Ashrafian, H. 2019. Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (AI) and autonomous robotic surgery. *International Journal of Medical Robotics and Computer Assisted Surgery* 15, e1968.

Parfit, D. 1984. *Reasons and Persons.* Oxford University Press.

Petersen, M. 2015. *The Prisoner's Dilemma.* Cambridge University Press.

Rand, D. G., Greene, J. D., and Nowak, M. A. 2012. Spontaneous giving and calculated greed. *Nature* 489, 427–430.

Rahwan, I. et al. 2019. Machine behavior. *Nature* 568, 477–486.

Ruggeri, K., Benzerga, A., Verra, S., and Folke, T. 2020. A behavioral approach to personalizing public health. *Behavioural Public Policy,* 1–13.

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.

Schroer, J. W. and Schroer, R. 2014. Getting the story right: a Reductionist narrative account of personal identity. *Philosophical Studies* 171, 445–469.

Simms, A. and Nichols, T. 2014. Social loafing: a review of the literature. *Journal of Management Policy and Practice* 15, 58–67.

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.

Strack, F., & Mussweiler, T. 1997. Explaining the enigmatic anchoring effect: mechanisms of selective accessibility. *Journal of Personality and Social Psychology* 73, 437–446.

Thaler, R. H. and Sunstein, C. R. 2008. *Nudge: Improving Decisions About Health, Wealth, and Happiness.* Yale University Press.

Turner, M. E. and Pratkanis, A. R. 1998. Twenty-five years of groupthink theory and research: lessons from the evaluation of a theory. *Organizational Behavior and Human Decision Processes* 73, 105–115.

Tversky, A. and Kahneman, D. 1974. Judgment under uncertainty: heuristics and biases. *Science* 185, 1124–1131.

Tversky, A. and Kahneman, D. 1983. Extensional versus intuitive reasoning: the conjunction fallacy and probability judgment. *Psychological Review* 90, 293–315.

Upadhyaya, N. and Galizzi, M. M. 2023. In bot we trust? Personality traits and reciprocity in human-bot trust games. *Frontiers in Behavioral Economics* 2, 1164259.

Wildschut, T., Pinter, B., Vevea, J. L., Insko, C. A., and Schopler, J. 2003. Beyond the group mind: a quantitative review of the interindividual-intergroup discontinuity effect. Psychological Bulletin 129, 698–722.

Wischert-Zielke, M., Weigl, K., Steinhauser, M., and Riener, A. 2020. Age differences in the anticipated acceptance of egoistic versus altruistic crash-control-algorithms in automated vehicles. In: *Proceedings of the Conference on Mensch und Computer.* ACM, 467–471.

Whiting, T. et al. 2021. Confronting barriers to human-robot cooperation: balancing efficiency and risk in machine behavior. *iScience* 24, 101963.

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.