Re-Framing Transparency, Interpretability, and Explainability for Multi-Agent Systems

Jessica Woodgate[0000−0001−9039−846X], Daniel E. Collins[0000−0002−1075−4063], Yining Yuan[0009−0000−3093−5526], and Nirav Ajmeri[0000−0003−3627−097X]

University of Bristol, UK
{jessica.woodgate,daniel.collins,yining.yuan,nirav.ajmeri}@bristol.ac.uk

Abstract. Within a multi-agent system (MAS), multiple agents and humans interact to achieve goals. To respect autonomy and act effectively, it is crucial the participating entities of a MAS view and comprehend the workings of the system. This is supported by implementing transparency, interpretability, and explainability. Broadly, transparency conveys the ability of observers to freely observe information about the system; interpretability is the extent to which an observer can understand the system; and explainability involves logical reasons for decisions and behaviours of and within a system. Prior works largely focus on these concepts within individual or one-off interactions. However, in a MAS interactions are dynamic—as the system and relevant information evolves over time—and diverse—as entities may be technical or social, acting individually or in groups. Widening the perspective to incorporate heterogeneous interactions alters considerations relevant to engendering understanding and conveying information. We reflect on previous literature to re-frame transparency, interpretability, and explainability for MAS. We highlight important gaps and suggest research directions for making MAS transparent, interpretable, and explainable.

Keywords: Sociotechnical Systems · Human-Agent Interactions · Normative Frameworks.

1 Introduction

Multi-agent systems (MAS) are composed of multiple artificial agents and humans acting to achieve shared and individual goals [140]. Understanding MAS is challenging because system-level behaviours emerge from the interactions of these agents and cannot always be inferred from examining individual agents alone [123]. As MAS are increasingly integrated into everyday life, the ability of agents to make or support decisions raises both epistemological questions—about why certain decisions are made—and ethical questions—about whether good reasons can be given for those decisions [54].

Facilitating understanding of MAS is further complicated by the dynamic, heterogeneous, and often open nature of these systems [127, 137]. Entities in MAS may include purely technical agents, humans, or hybrid combinations, interacting in both one-off and temporally extended scenarios. The goals, roles,

and availability of entities can change over time, highlighting the need for an understanding that captures evolving information and supports cooperation, coordination, and accountability [11, 153].

Much of the prior research on transparency, interpretability, and explainability focuses on individual AI systems such as machine learning predictors or recommendation algorithms [113, 79]. These works typically examine one-off interactions between a single human and a single AI system, providing methods to elucidate how the AI system produces its outputs. While informative, this perspective is limited: AI systems in isolation differ fundamentally from MAS, where multiple agents, both technical and social, interact continuously within complex sociotechnical environments [70, 91].

Building upon insights from AI transparency and interpretability research [52, 56, 90], we extend these concepts to MAS settings. We consider interactions among multiple technical agents, combinations of technical and human agents, varying numbers of entities, and dynamic temporal contexts [21, 30]. By situating MAS within sociotechnical contexts, we highlight gaps in existing research and propose directions for enhancing transparency, interpretability, and explainability in multi-agent systems.

The rest of the paper is organised as follows: Section 2 examines previous literature to define each concept; Section 3 re-frames each concept to reflect MAS considerations; Section 4 highlights important gaps and suggests directions for future research; and Section 5 concludes the paper.

2 Examining Transparency, Interpretability, and Explainability

We first reflect upon prior AI and MAS literature to elucidate how transparency, interpretability, and explainability have been defined. Appendix A details the literature review methodology.

2.1 Transparency

Transparency concerns the visibility and availability of information, such that decisions or behaviours can, in principle, be discovered and examined [66]. Central to transparency is the truthfulness and reliability of information provided [67]. Transparency may apply to both events or processes. Events include inputs, outputs, actions, or outcomes, while processes include organisational rules, regulations, procedures, and agent decision-making mechanisms [60]. Such agent decision-making mechanisms range from simple heuristics to complex AI algorithms and ML predictors.

Transparency can be framed relative to an observer (agent, human, individual, or group), as the extent to which that observer possesses knowledge of an object or process [157]. It involves the intentional disclosure of information, enabling objects (events or processes) to be open to examination by relevant parties

[133, 53]. Transparency is also often defined in contrast to opacity, where epistemically relevant elements remain inaccessible to the observer [81, 157]. In this sense, transparency is a means of "seeing inside" an object, with the expectation that such visibility supports accountability and enables scrutiny [11].

A key function of transparency is to enable auditing, that is, the ability to trace, inspect, and verify decisions or behaviours [78, 127]. However, the mere availability of information does not guarantee that it is meaningful or usable by observers. In MAS, this challenge is amplified. Transparency must extend beyond individual agents to include interactions, coordination mechanisms, and system-level dynamics, which are often distributed, dynamic, and only partially observable.

2.2 Interpretability

Interpretability concerns the extent to which an observer can derive qualitative understanding from an object. In cognitive terms, understanding involves the ability to infer relationships or make predictions based on information represented in semantic memory [22]. Accordingly, interpretability is inherently relational: it is not solely a property of an object, but of the interaction between the object and an observer [54, 121]. For instance, the interpretability of an agent does not depend only on its internal structure, but on whether its users possess the relevant knowledge to form an appropriate understanding of its behaviour [76].

As the semantic memory of humans is a result of individual experiences, understandability incorporates some degree of subjectivity and variability [54]. What is understandable to one person may not be understandable to another. Relevant knowledge for interpreting an object is that which provides insight for a particular audience into a chosen problem [90]. Interpretation is thus the operation of binding objects to subjective meaning [33, 54] and making an object interpretable involves presenting it in understandable terms [43, 98].

In ML, interpretability is categorised as either intrinsic or post hoc [80, 90]. Intrinsic interpretability arises when models are designed to be inherently understandable, for example, through constrained complexity or transparent decision structures [123, 152]. In contrast, post hoc interpretability is achieved by applying methods after model development to extract insights from otherwise opaque systems, such as through explanations or visual analyses [43, 81]. These approaches differ fundamentally: intrinsic interpretability constrains the design of the system itself, whereas post hoc methods seek to recover understanding without altering the underlying model.

Interpretability serves a range of purposes, including supporting trust, enabling validation and debugging, uncovering causal relationships, and meeting legal or ethical requirements [81]. However, it is often a means to broader ends such as justification, accountability, or fairness [76]. In MAS, interpretability must extend beyond individual components to encompass how observers make sense of interactions, coordination patterns, and emergent behaviours arising from multiple agents acting over time.

2.3 Explainability

Explainability refers to the process of providing information that gives reasons for a particular phenomenon [86]. An explanation is a communicative activity between a source (explainer) and a receiver (explainee) describing relevant context or causes surrounding an event [3, 32].

The nature of what an explanation "is" can be conceptualised as an illocutionary act (something said with the effect of an action), typically expressing words in a context with an intention. Explanations should be distinguished from perlocutionary acts—the effect an explanation can have on others' thoughts and beliefs, such as helping someone to understand. Thus, explainability concerns the availability of reasons for a phenomenon, and an explanation is an interface between an explainer and explainee that communicates those reasons. An explanation should answer the explainee's how, what, where, or why questions.

Explainability can serve as a mechanism for achieving transparency by making underlying processes more accessible, and as a means of supporting interpretability by structuring information in a way that is meaningful to observers [9, 100, 157]. The aim of explanations is often to improve understanding, satisfaction, and acceptance of decisions or behaviours [86, 155].

The nature of explanations has been examined through causal [86] and normative [134] lenses. An explanation according to the causal theory traces the causal processes leading to and making up the event itself [112]. An explanation is a fact that is not known for certain but, if found to be true, would constitute a cause of the item to be explained (the explanandum) [57]. Some causes are better explanations than others; a question thus arises as to which causes to select [87]. Counterfactual explanations select relevant causality by comparing the case in which the event happened to a counterfactual case in which it did not, focusing on the factor differentiating the event and contrasting event [61, 155]. Normative explanations argue that some non-causal phenomena, such as rule-following, promising, and ethical behaviour, cannot be explained by causal explanations but are powered instead by normative forces, including commitment and prohibitions [134]. Norms can be used to guide causal explanations by highlighting normative deviations, where deviations explain causes by adopting the role of counterfactuals [61].

Explanations also play a central role in social interaction. Explainees may request explanations to understand decisions, while explainers may provide them to justify actions, persuade others, or build trust [84, 86]. Through this interaction, explanations support reasoning about behaviour, enable prediction, and facilitate coordination [41]. They also enable feedback loops, whereby explainees can challenge or refine decisions, and explainers can improve their behaviour accordingly [3, 56].

In MAS, explainability must address not only individual decisions but also joint actions, inter-agent dependencies, and temporally extended interactions. This introduces additional complexity in determining what constitutes a relevant explanation, who should provide it, and how it should be communicated across multiple agents and stakeholders.

3 Re-Framing Transparency, Interpretability, and Explainability for MAS

We conceptualise transparency, interpretability, and explainability in MAS through four dimensions: *epistemic*, *functional*, *temporal*, and *social*.

Epistemic dimension concerns the standpoint from which an observer examines the MAS: whether the observer focuses on the system as a whole, on interactions within the system, or on the system from an internal or external position.

Functional dimension concerns the intrinsic properties of the system and the decision-making mechanisms operating within it.

Social dimension concerns the entities involved, including whether observers and observed parties are individuals or groups, and whether understanding is required in one-off or interactive settings.

Temporal dimension concerns whether relevant phenomena are past, present, or future, and whether understanding must be updated over time as the MAS evolves.

Appendix B discusses ethical considerations for implementing transparency, interpretability, and explainability in MAS.

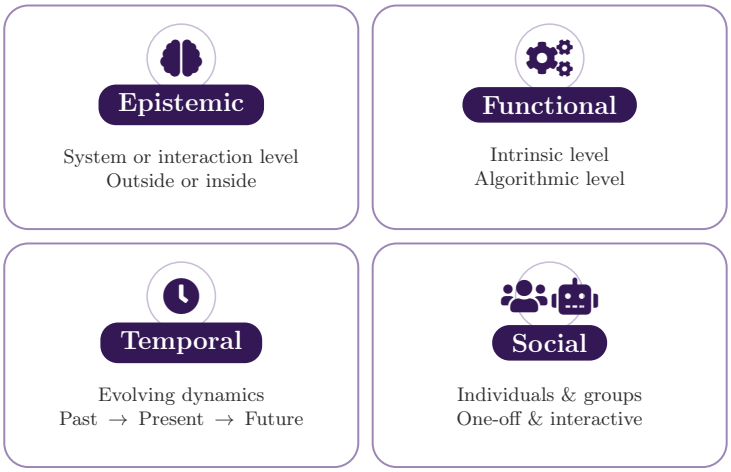

Fig. 1. Dimensions of transparency, interpretability, and explainability for MAS.

3.1 Re-Framing Transparency

A MAS has different transparency requirements for each stakeholder group (e.g., the general public, users, designers or developers, companies, and regulators [67])

and different transparency requirements arise regarding whether relevant entities are technical or social. Distinguishing between technical and social entities in making an object transparent is important as technical and social entities process and store information in significantly different ways. To make something transparent, relevant information to understand the behaviour of an agent includes its decision-making mechanism, parameters, and goals. Relevant information to understand human behaviour includes their cognition and values. With respect to receiving information, an agent may be able to process numerical information, for example, much faster and store much more of it than a human, whilst a human may be much better at processing and storing information about social dynamics. Auditing, facilitated by transparency, is especially important in MAS settings to determine responsible agents for specific outcomes [99].

Transparency is a property of an object Y (e.g., the behaviour of an agent A) relating to the availability of relevant information X about Y to an observer B. Y is transparent to B to the degree that B is able to observe X. Y can be made transparent to B through a process that disseminates X to B. For example, A may engender transparency of Y for B by disseminating X via an active process (e.g., communicating explanations) to B, or as inherent property (e.g., making goals public). B may then use X to update its decisions.

Agent-Agent Transparency. As the information available to agents influences how each agent updates its decisions, modifying transparency can alter the behaviour of the agents and system as a whole [44]. In the case that incentives are aligned between a group of agents, transparency plays an important role in facilitating cooperation. Opacity of individual agents within a MAS can harm the reputation of an agent as well as the trust of the overall system [54]. Conversely, if agents are unaligned, an agent may act to obscure its plans, reducing transparency, or act so as to encourage other agents to misinterpret their plans [106]. Transparency may not always be beneficial, as giving more agents more information can enable individual agents to optimise their own agendas more efficiently and lead to a worse global outcome, discrimination, or unfairness [141]. For example, when agents have information about which agents are in their tribe, agents are more likely to cooperate with others of the same tribe and defect against others [58]. This highlights the importance of considering the general equilibrium and fairness effects of information dissemination and transparency.

Human-Agent Transparency. When an agent acts on behalf of a human, transparency requirements of the agent arise, as when an agent has more information than a human, it is difficult for the human to check whether the agent is acting in the human's interest [141]. In mixed human-agent teams, agents that make information about decisions transparent influence human trust and group identification [132]. Transparency of past choices by group members plays

a key role in maintaining cooperation [50] and allows entities to indicate cooperative intentions, which may encourage cooperation from others [39]. Increased transparency has a positive effect on human perception of trust when agents use individualistic strategies, and a negative effect when agents unconditionally cooperate [132]. MAS can be made transparent to humans by incorporating explicit inspection points [111]; displaying reasoning paths and agent roles in a system [77]; revealing strategies agents adopt in a game [132]; identifying important states and expressing decision intention of agents [158]; and making clear current actions, plans, reasoning process, and outcome predictions [31]. Transparency requirements differ depending on the observer; for example, transparency is significantly important for developers and end users, but with different focuses and expectations [93].

3.2 Re-Framing Interpretability

As interpretability forms a key part of social interaction by helping interactees understand one another [54], it is a crucial component in MAS with multiple interacting agents and humans. Insights from interpretability are used to guide communication, actions, and discovery [90]. Interpretability entails that an observer A has the right sort of knowledge about an object Y that enables A to have a particular relationship R to Y. Interpretability is thus a property of R that A has with Y. In agent-agent interactions, R concerns agents being able to make correct inferences so that they are able to achieve their goals. In human-agent interactions, R involves humans being able to form cognitive understandings and preserve human autonomy. Interpretability can be imbued through the goals, plans, or rewards an observer can ascribe to an agent or group of agents based on observations of that agent or group.

> Interpretability is a relation between an observer A and X, a function of A's past or current observations, and Y, representing some object. The interpretability of Y given X is the extent to which X reduces A's uncertainty about, or improves A's predictions of, Y during inference. If X is interpretable to A, A can better infer which actions to take to achieve A's goals by estimating what X will be next.

Agent-Agent Interpretability. An agent in a MAS selects actions, conditioned on local observations, to steer outcomes towards goals [140]. When selecting actions, an agent makes inferences about environment dynamics—including the influence of other agents—from local observations [1]. Interpretability between agents thus concerns the inferences an agent makes about other agents given its observations. Within agent-agent interactions, interpretability requirements are contingent on agent goals, as to achieve shared goals, agents need to be able to effectively cooperate [140], and to effectively cooperate, agents need to be able to make inferences about one another [141]. Mutual interpretability of behaviour enables agents to coordinate action selection with the expected actions of others and thereby achieve shared goals.

In agent-agent interactions, A's model (internal representation of the current state [109]) includes other agents. A's model of another agent B can include B's beliefs, state information, goals, intentions, capabilities, or reward function [30]. To be interpretable to A, B's behaviour should in some way conform to, or inform, A's expectations reflected in the model that A has of B. A mismatch between B's behaviour and A's model entails that A's expectation will not be the same as B's behaviour, indicating a lack of interpretability. Interpretability also facilitates communication between A and B and enables cooperation and coordination to help achieve shared goals. If A's performance on some task is contingent on B's behaviour, the success of an attempt to make B interpretable to A could be measured by examining how A's performance improves after receiving information from B [141].

Human-Agent Interpretability. Interpretability in human-agent interactions differs from agent-agent interactions as the need for interpretability originates from a different place. Between agents, interpretability is needed to facilitate agents making correct inferences in order to cooperate, communicate, and achieve goals. For humans, interpretability requirements arise as they are intrinsic to human autonomy and dignity [46, 78]. An agent should be interpretable to a human so that the human remains autonomous over the agent's behaviour. Sometimes, this will be so that the human can achieve their goals; other times, interpretability will be an end in itself.

Interpretability is used to help humans understand the reasoning behind an agent's decision-making. Some works use post hoc methods to retroactively make behaviour interpretable to a human observer [107, 118]. Interpretability can be fostered by designing answerable systems, which are systems that provide reasons for an act or state why one act has been chosen over other alternatives [72]. Other works focus on making decision-making intrinsically interpretable to humans via policies that can be directly traced to human-meaningful variables [17], harnessing decision-making mechanisms that humans can immediately understand [23], communicating actions and intentions [142], and incorporating expert knowledge to regularise decision-making mechanisms and influence encoded information [156].

As well as it being important for humans to understand agents, it is also helpful for agents to be able to follow humans and human motivations [141]. Methods to support agents interpreting unclear instructions include leveraging normative frameworks, in which an agent seeks clarification or infers implied meaning if an instruction violates one of the norms of quantity, quality, relation, and manner [110]. Other approaches attempt to create interpretable models of humans for agents and other humans to later assess, to improve the accuracy of agent interventions and support human experts in understanding which interventions work for which individuals and why [96].

3.3 Re-Framing Explainability

A MAS encompasses the goals and constraints of multiple agents and stakeholders simultaneously, and different entities have different requirements for explanations [16]. Where the aim of an explanation is to increase satisfaction, understanding, and acceptance towards an object [86, 155], explanations should provide information that is meaningful to the observer [116]. What constitutes a valid explanation is therefore subject to the receiver and circumstances [55, 116].

Understanding the motivations and expectations behind the needs of different audiences, including whether audiences are social or technical, helps to produce appropriate explanations [25]. Humans and agents perceive, store, and interpret information in fundamentally different ways. Social explainees require presentation in human understandable terms, whilst technical explainees may have technical specifications or communication protocols. Technical entities have different requirements for explanations depending on the agent's varying properties, capabilities, constraints, knowledge bases, communication protocol, and observations [16]. Four distinct social audiences relevant to explanations of a MAS include: end users (apply decisions, desire explanations that build trust and confidence); affected users (impacted by decisions, desire explanations to understand if they were treated fairly and what factors could be changed to get a different result); regulatory bodies (such as government agencies, wanting to ensure decisions are made safely and efficiently); and system builders (including those who develop or deploy a system, wanting to know if the system is working as expected and how it can be diagnosed and improved) [62].

Explainability in MAS thus involves presenting a reasonable account of information that helps an observer to understand: which properties and constraints of an agent have influenced a decision and how; how the system balances or prioritises potentially conflicting preferences; and, considering the above information, why has this particular outcome been reached and not other possible outcomes? Some explanations are better than others, and can be evaluated by various objective and subjective measures such as interpretability, truthfulness, conciseness, and relevancy to the context of the communicative act [64, 65, 145]. The scope of an explanation is also important, which denotes whether an explanation applies only to a particular case, or whether it can be reused in other settings [131].

An explanation E is a communication about an object Y conveyed by an explainer A to an explainee B. E is an explanation if A utters E with the intention E will render Y understandable [2]. An explanation is useful if it is: (1) interpretable, insofar as E reduces B's uncertainty about, or improves B's predictions of, Y; and (2) transparent, insofar as the information in E is truthful. The usefulness of E can be measured by examining whether B's inferences regarding Y, given the combination of B's past observations and the explanation E, are better than inferences given only B's past observations. E is something that A can choose to give or be asked to provide (e.g., by B).

Agent-Agent Explainability. Explainability has been used in MAS for agents to justify decisions to other agents [92, 95], align agents with one another [12, 29], and negotiate shared understanding [21]. For example, explanations with blockchain enable an agent joining a MAS to increase its reputation by explaining its behaviour with tangible proofs [27]. Theory of mind—the attribution of mental states to oneself and others—has also been employed for agents to explain or make interpretable agent behaviour [45, 122].

The temporal dimension of explainability raises considerations of the time for conveying an explanation and time the explainee is allowed to spend on understanding an explanation. Agents need to be able to generate timely explanations and base decisions on evolving knowledge of the world within given temporal bounds [10]. For example, an explanation that is simple to understand is preferable when an observer needs to make a quick decision, whereas more exhaustive explanations may be preferred in settings with less time constraints [56].

Human-Agent Explainability. Explainability between an agent and a human involves the human forming an understanding that links the inputs, internal reasoning, and outputs of the agent in a way that fits the human's relevant goals, expertise, and context [105]. Interacting with an explainable agent empowers a human to adjust an agent to the human's goals through engendering understanding of whether the agent's behaviour is aligned with the human's goals. It also enables humans to adapt and coordinate their own behaviour to the agent [59]. To tailor explanations of agents for humans, methods have employed folk-psychological terms [143, 144, 151] (e.g., explaining an agent by citing its beliefs, goals, and intentions [8]), as it naturally aligns with how humans explain choices and actions [84]. Explanations for actions can be informed by the differences between an agent's own model and the human observer's mental model of the agent, in order to reconcile the human's expectation of the agent with the outcome of the agent's behaviour [126].

Explanations in MAS should address the needs of relevant parties. Explanations for humans should aim to increase the user's satisfaction by taking into account factors such as preferences, constraints, fairness, and privacy [18, 75]. Relevant considerations include the intentions of the system designer, previously implemented in a multi-actor explanation framework that considers the agent, user, and designer [26]. For end users, explanations have been found as more compelling if they explicitly state why one option was chosen over another (i.e., what the agent deems more important) [144].

4 Gaps

We now explore key gaps for implementing transparency, interpretability, and explainability in MAS and highlight possible directions for future research.

4.1 Longitudinal Effects

Existing tools largely focus on individual and static decisions [83, 102]. This fails to capture the dynamic nature of MAS where actions of one entity influence the behaviours of others, impact future states, and require balancing immediate rewards with long-term risks [120]. Widening the temporal window changes the requirements of transparency, interpretability, and explainability.

In interactions longer than one-off encounters, expectations or behaviour of interactees may evolve. For example, familiarity may increase as a human and an agent spend more time together, increasing alignment between expectations and behaviour. Closer alignment between expectations and behaviour increases interpretability, either because the human better predicts how the agent behaves, or because the agent better interprets the human's expectations and more accurately fulfils their goals. Interpretability can thus be framed as a dynamic process that evolves over time and across context [22]. Increased interpretability over time may change explainability requirements, such as reducing the need for explanations as interactees develop a better understanding of one another [30].

Longitudinal effects can be accounted for by characterising interpretability as a spectrum rather than a binary requirement, for example, by decomposing agent behaviour into step-by-step explanations for each important decision and accommodating for explanations changing over time [138]. Torta et al. [129] propose agents that create temporal explanations clarifying how historical actions may have affected later actions in a MAS, even those assigned to other agents. Towers et al. [130] explain how an agent made a decision via the temporal context of actions. Alelaimat et al. [8] use historical data to generate belief-based explanations of an agent's past actions.

In addition to interpretability and explainability, the appropriateness of transparency can change over time. Transparency is often beneficial, but it is not a universal good, and the level of information disclosure affects decision-making and outcomes in MAS [141]. A system with increased transparency but poor interpretability and explainability may reduce stakeholder trust [40]. For agent-agent interactions, as each agent in a MAS uses disseminated information to update its decisions, modifying transparency can alter the behaviour of agents and the system as a whole. If agents are solely self-interested, giving more agents more information can enable individual agents to optimise their own agendas more efficiently and lead to a worse global outcome, as individual agent objectives may not be aligned with each other or overall social welfare [35, 141]. In a traffic scenario, for example, Arnott et al. [15] find that if all queuing cars learn about an extra road and see an opportunity to reduce their delay by switching routes, it can lead to a greater delay for everyone. On the other hand, if some privileged subgroup is given additional information, it can result in a better outcome for everyone, but also some are more advantaged than others. Implementing transparency adaptively can improve system performance. Strategic information disclosure can balance revenue optimisation with allocation efficiency [154]. Levels of transparency in decision-making of self-interested agents and game outcomes can be adapted to promote optimal network performance with maximised wel-

fare and minimised free riding [14]. In negotiation scenarios, a negotiator who withholds information can obtain better outcomes for themselves, partially due to increased deception [68].

Future Directions. Widening the temporal window to consider varying timescales in the implementation of transparency, interpretability, and explainability. For example, instead of providing explanations for single decisions, methods could examine explanations over n-steps and incorporate how longitudinal effects alter the requirements for explanation.

4.2 From Individuals to Groups

Agents and humans may act individually or in groups in a MAS, raising considerations of: (1) making objects transparent, interpretable, and explainable for group observers; and (2) making the behaviour of groups transparent, interpretable, and explainable.

Decisions may be made by a group of humans sharing one observational interface [94], yet users may vary greatly, and therefore so may their understanding of phenomena [54]. Individuals who fall into the same stakeholder group can still have differing and dynamic roles and responsibilities with varying requirements. If an observer is a group, attempts to imbue transparency, interpretability, and explainability need to respond to group dynamics.

Explaining the individual contributions of agents that act in groups has been explored in reinforcement learning (RL) literature [12, 92]. Research has also examined how to aggregate multiple individual explanations [6, 37]. Whilst an individual agent can appeal to its policy or value estimate to provide an explanation, when agents act in groups, explanations are of joint behaviours or phenomena and encompass coordination protocol, norm emergence (where norms are adopted through interactions between agents [88]), and sanctions.

Groups vary in type and size. Explanations of group behaviour and explanations for groups should take into account the type of group in consideration, the group's circumstances, and its requirements. Groups can be broadly categorised into aggregate groups and jointly acting groups. An aggregate group is where each group member acts individually and no joint plan or coordination is implied, but members share a category label and are socially categorised together as a group under that label. For example, a classroom of students or welfare recipients. A jointly acting group is a set of individuals that form a joint intention and act as if they were a single agent, for example, councils or committees. In a jointly acting group, there is coordination, shared planning, and specific group-level decisions. The behaviours of aggregate groups can be explained using causal history explanations and the behaviours of jointly acting groups can be explained through intentions [97].

Future Directions. Expanding the scope from interactions between individuals to address interactions with multiple entities, making the behaviours of

groups of varying types and sizes transparent, interpretable, and explainable, as well as imbuing transparency, interpretability, and explainability for varying types and sizes of groups.

4.3 Personalisation

Particular observers require particular kinds of knowledge, making them likely to seek particular kinds of information through transparency, interpretability, and explainability [139, 157]. Relevant information can be determined by ascertaining the exact question the observer is interested in, what they already know, their level of expertise or knowledge within a context, long-term goals, short-term goals, and level of attention [20, 56, 138]. Different types of explanations serve different purposes. Personalised explanations tailor to the explainee's cognitive capabilities, epistemic state, and interests [28, 139]. The explanation should build bridges between presumed knowledge and information novel to the explainee [41]. Explanations should be understandable to all relevant observers, considering expertise and accessibility requirements, and fair in that they do not further marginalise particular groups or cause injustice.

One way to achieve personalisation for multiple users is by providing multiple explanations, with each explanation targeted to a specific user [103], altering the level of detail in the explanation according to the user [13], adjusting the warmth of language according to the task [71], or utilising adaptive transparency modules so that the level of system transparency can be tailored to individual preferences and specific users [93]. Situation awareness, which defines the informational needs for humans operating in a scenario, can also be used to assess explanation quality for a particular setting [114].

Future Directions. Considering personalised needs for transparency, interpretability, and explainability in MAS, taking into account the goals, needs, and backgrounds of relevant entities. Incorporating broad social contexts, such as whether interactees are technical or social entities, whilst addressing accessibility and fairness considerations.

4.4 Proactive Explanations

Explanations could be generated as a response to specific input from the explainee (reactive) or without any input from the explainee (proactive). Many previous works generate explanations when prompted, for example, when an explainee queries an agent for an explanation [75]. However, there may be some settings where it is desirable for an agent to proactively produce an explanation without being queried. A proactive explanation anticipates the need of the intended explainee and provides an explanation without a request.

To generate a proactive explanation, an agent needs to identify the conditions in which an explanation is required. Explanations may be required when something is not immediately interpretable, insofar as the observer cannot understand it, and transparent, insofar as truthful information is not available. For

humans, people tend to seek explanations either if they themselves wonder why a behaviour occurred, or if they expect that someone else wonders why a behaviour occurred [41]. An agent could provide an explanation alongside its decision if it predicts that a decision is likely to be misunderstood by the recipient.

Capabilities for generating proactive explanations may require modelling other agents or humans in the environment [131]. Interpretability conceptualised as the calibration of an agent's model of the state to an observer's model entails that a lack of alignment between the models indicates a lack of interpretability. If an agent detects that there is a lack of calibration between its model and its observer's model, it could proactively generate an explanation for the observer. Explanations can thus be used to pre-emptively improve calibration between models and thereby interpretability.

Future Directions. Methods to produce proactive explanations that are generated without input from the recipient, for example, by providing an explanation when interpretability is predicted to be low.

4.5 Interactive Explainability

In social settings, many decisions are not one-off instances but incorporate some form of back-and-forth interaction between relevant entities. To ensure explanations evolve along with the setting they are in, explanations can be re-framed as a process, requiring dynamic and iterative refinements of multiple explanations between two or more entities [22, 47, 139].

A static explanation does not change in response to feedback from the explainee, whilst an interactive explanation allows explainees to probe the explainer, such as by asking for different types of explanations [16, 33] or asking follow-up "what if?" questions [124]. A dialogical approach allows the explainee to guide the discussion in a way that satisfies their needs and preferences, and for entities to fill gaps in one another's knowledge [139]. Interactive explanations enable explainees who disagree with a behaviour to provide feedback, whilst allowing the explainer to offer additional information and explainees to delve deeper into the reasoning behind behaviours [25]. For some decisions, being clear about goals and preferences reduces the need for back-and-forth interaction [34].

MAS frameworks such as multi-agent argumentation have been harnessed to explain outputs of large and complex systems, emphasising the importance of viewing explanations as dialogues rather than static feature attributions. Kori et al. [74] present a method to generate explanations through debates between two agents. Rago et al. [101] propose interactive explanations, where agents use argumentative exchanges to resolve conflicts. Al Anaissy et al., Čyras et al. [7, 36] employ interactive explanations with argumentation. Sendi et al. [117] integrates multi-agent argumentation into ensemble learning, constructing arguments for and against a classifier to determine which are acceptable and provide reasons justifying the decision. Jentzsch et al. [69] foster explainability through conversational interfaces between humans and agents. Xu et al. [150] present

a formal framework for dialogues involving two entities to explain rule-based systems, considering the context of the human. Kekulluoglu et al. [72] propose a dialogue-based approach where an agent and human exchange information to reach a shared understanding so that the agent confirms its assumptions about the human, and the human gives feedback on and can challenge the reasons provided by the agent. Ciatto et al. [34] propose a general-purpose protocol for interactions between an explainee and explainer. A MAS with diverse social and technical entities may have a wide array of interactions happening.

Future Directions. Building on MAS interaction frameworks to facilitate explainability that incorporates feedback from the receiver, considering varying numbers of relevant entities and whether entities are technical or social.

4.6 Moral Explanations

The sociotechnical context of MAS, in which decisions made have varying effects on diverse stakeholders, necessitates careful considerations of moral implications, where morality concerns what is good or right [42, 51, 147]. Explaining and interpreting moral decisions is challenging because such decisions often have various valid (and invalid) explanations, with multiple reasoning steps that do not necessarily belong to the same moral paradigm [138, 146]. Morality is difficult to embed, as there can be multiple interpretations of any moral rule [19]. Representing morality using just one moral philosophy raises difficulties as peoples' morality and values (what is important to them in life [115]) may change depending on the situation they are faced with [82, 104, 119, 136]. Vijayaraghavan and Badea [138] suggest that interpretability can be harnessed as a debugging tool in moral decision-making so observers can provide feedback and improve behaviour. Mosca et al. [89] implements explainability through justifications for decisions that consider the promotion of moral values. Agiollo et al. [4] find large discrepancies in relevant concepts identified by post hoc explainer mechanisms for explaining moral decisions. Explaining moral decisions should reflect that diverse stakeholders may have different and potentially contrasting moral preferences, that moral rules can be interpreted in varying ways, and that there may be multiple moral rules that apply and potentially conflict [104, 148].

Future Directions. Methods to make moral decisions interpretable, explainable, and transparent. For example, moral theories could be utilised to explain decisions considering that multiple moral theories may be relevant for one decision. Alternatively, moral theories could be operationalised to help determine the appropriate level of transparency, interpretability, and explainability for a particular decision, given that requirements can change.

4.7 Normative Explanations

Social norms are standards of expected behaviour used to govern MAS and encourage coordination [24, 149]. A normative system—the collection of norms

in a MAS—is often dynamic with continuously evolving norms as participating entities, and what is important to them, change [48]. Normative requirements of explainability and related concepts may vary between different geographical and social domains [125]. To facilitate entities' understanding of the existing norms, explanations help illuminate normatively relevant features [49]. Agrawal et al. [5] utilise explanations to convey normative information to agents joining a MAS. Tzeng et al. [135] employ various levels of communication to enhance norm emergence.

Normative frameworks can also be applied to improve the transparency, interpretability, and explainability of a MAS. [85] harness normative information to make agent roles and interactions transparent. As agents in a MAS operate within explicit or implicit organisational and institutional structures, understanding both the agents within the system and the system itself involves collective behaviour, structural constraints, and dynamics, not just single-agent decisions. Normative concepts such as roles, obligations, and commitments can be harnessed to elucidate the behaviour of the MAS as a whole. Emerged norms can be used to help interpret the behaviour of groups and improve the transparency of group decision-making.

Future Directions. Exploring the bidirectional relationship of how transparency, interpretability, and explainability influence the norms in a MAS, as well as how the norms of a MAS alter requirements for transparency, interpretability, and explainability. Further research could also examine how system-level explanations, for example, of an institution or organisation, may be of a different nature to the explanation of an agent or interaction.

5 Conclusion

Ensuring that MAS are able to function effectively and respect human autonomy necessitates that participating entities are able to understand and have access to relevant information about the system. This gives rise to the need for MAS to be interpretable, explainable, and transparent. The sociotechnical nature of MAS necessitates that implementing these terms must incorporate the wider social setting. In this paper, we have reflected on definitions in literature and explored key considerations for employing these terms beyond one-off, single-user interactions. We have highlighted important gaps for research to address to ensure that systems reflect the needs and preferences of diverse users and incorporate the varying levels of interaction that take place within a MAS.

Acknowledgments. We thank Joseph Trevorrow for the discussions on reframing and for feedback on the earlier versions of this draft. JW thanks Google PhD Fellowship for support. DC was supported by the UKRI Centre for Doctoral Training in Interactive Artificial Intelligence Award Grant No. EP/S022937/1). YY thanks China Scholarship Council (CSC) and the University of Bristol under joint PhD Scholarship Award No. 202308060207. NA acknowledges support from the UKRI EPSRC Grant No. EP/Y028392/1: *AI for Collective Intelligence (AI4CI)*.

Bibliography

[1] David Abel, André Barreto, Michael Bowling, Will Dabney, Shi Dong, Steven Hansen, Anna Harutyunyan, Khimya Khetarpal, Clare Lyle, Razvan Pascanu, Georgios Piliouras, Doina Precup, Jonathan Richens, Mark Rowland, Tom Schaul, and Satinder Singh. Agency is frame-dependent, 2025. URL https://arxiv.org/abs/2502.04403.

[2] Peter Achinstein. What is an explanation? *American Philosophical Quarterly*, 14(1):1–15, 1977. ISSN 00030481. URL http://www.jstor.org/stable/20009644.

[3] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6(1): 52138–52160, 2018. https://doi.org/10.1109/ACCESS.2018.2870052.

[4] Andrea Agiollo, Luciano Cavalcante Siebert, Pradeep Kumar Murukannaiah, and Andrea Omicini. The quarrel of local post-hoc explainers for moral values classification in natural language processing. In *Explainable and Transparent AI and Multi-Agent Systems (EXTRAAMAS)*, pages 97–115, London, 2023. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-40878-6_6.

[5] Rishabh Agrawal, Nirav Ajmeri, and Munindar P. Singh. Socially intelligent genetic agents for the emergence of explicit norms. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI)*, pages 10–14, Vienna, July 2022. IJCAI.

[6] Tobias Ahlbrecht and Michael Winikoff. Explaining aggregate behaviour in cognitive agent simulations using explanation. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems (EXTRAAMAS)*, pages 129–146, Montréal, 2019. Springer International Publishing. https://doi.org/10.1007/978-3-030-30391-4_8.

[7] Caren Al Anaissy, Srdjan Vesic, and Nathalie Nevejans. Towards ethical argumentative persuasive chatbots. In *Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems XVI (COINE)*, pages 141–160, London, 2023. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-49133-7_8.

[8] Ahmad Alelaimat, Aditya Ghose, and Hoa Khanh Dam. Mining and validating belief-based agent explanations. In *Explainable and Transparent AI and Multi-Agent Systems (EXTRAAMAS)*, pages 3–17, London, 2023. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-40878-6_1.

[9] Abeer Alshehri, Hissah Alotaibi, Tim Miller, and Mor Vered. A hypothesis-driven approach to explainable goal recognition. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 105–114, Detroit, 2025. IFAAMAS. https://doi.org/10.5555/3709347.3743522.

[10] Francesco Alzetta, Paolo Giorgini, Amro Najjar, Michael I. Schumacher, and Davide Calvaresi. In-time explainability in multi-agent systems: Challenges, opportunities, and roadmap. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems (EXTRAAMAS)*, pages 39–53, Online, 2020. Springer International Publishing.

[11] Mike Ananny and Kate Crawford. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3):973–989, 2018. https://doi.org/10.1177/1461444816676645.

[12] Giorgio Angelotti and Natalia Díaz-Rodríguez. Towards a more efficient computation of individual attribute and policy contribution for post-hoc explanation of cooperative multi-agent systems using myerson values. *Knowledge-Based Systems*, 260:110189, 2023. https://doi.org/https://doi.org/10.1016/j.knosys.2022.110189.

[13] Sule Anjomshoae, Kary Främling, and Amro Najjar. Explanations of black-box model predictions by contextual importance and utility. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems (EXTRAAMAS)*, pages 95–109, Montréal, 2019. Springer International Publishing. https://doi.org/10.1007/978-3-030-30391-4_6.

[14] Sadegh Arefizadeh, Sadjaad Ozgoli, Sadegh Bolouki, and Tamer Başar. Compartmental observability approach for the optimal transparency problem in multi-agent systems. *Automatica*, 143:110398, 2022. https://doi.org/https://doi.org/10.1016/j.automatica.2022.110398.

[15] Richard Arnott, Andre de Palma, and Robin Lindsey. Does providing information to drivers reduce traffic congestion? *Transportation Research Part A: General*, 25(5):309–318, 1991. https://doi.org/10.1016/0191-2607(91)90146-H.

[16] Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilovi'c, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques, 2019. URL https://arxiv.org/abs/1909.03012.

[17] James Ault, Josiah P. Hanna, and Guni Sharon. Learning an interpretable traffic signal control policy. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 88–96, Virtual Event, New Zealand, 2020. IFAAMAS. https://doi.org/10.5555/3398761.3398777.

[18] Gönül Aycı, Arzucan Özgür, Murat Şensoy, and Pınar Yolum. Can we explain privacy? *IEEE Internet Computing*, 27(4):75–80, 2023. https://doi.org/10.1109/MIC.2023.3270768.

[19] Cosmin Badea and Gregory Artus. Morality, machines, and the interpretation problem: A value-based, Wittgensteinian approach to building moral agents. In *Artificial Intelligence XXXIX*, pages 124–137, Cham,

2022. Springer International Publishing. https://doi.org/10.1007/978-3-031-21441-7_9.

[20] Akhila Bairy, Martin Fränzle, and Maike Schwammberger. Optimised timing of multi-step explanations for multiple users through reactive games. In *Explainable, Trustworthy, and Responsible AI and Multi-Agent Systems (EXTRAAMAS)*, pages 104–123, Detroit, 2026. Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-01399-6_7.

[21] Katharine Beaumont, Elena Yan, Samuele Burattini, and Rem Collier. Engineering inter-agent explainability in BDI agents. In *Explainable, Trustworthy, and Responsible AI and Multi-Agent Systems (EXTRAAMAS)*, pages 147–168, Detroit, 2026. Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-01399-6_9.

[22] Leslie M. Blaha, Mitchell Abrams, Sarah A. Bibyk, Claire Bonial, Beth M. Hartzler, Christopher D. Hsu, Sangeet Khemlani, Jayde King, Robert St Amant, J. Gregory Trafton, and Rachel Wong. Understanding is a process. *Frontiers in systems neuroscience*, 16(800280):1–18, 2022. https://doi.org/10.3389/fnsys.2022.800280.

[23] Clément Blanco-Volle, Nicolas Verstaevel, Stéphanie Combettes, Marie-Pierre Gleizes, and Michel Povlovitsch Seixas. Explainability and interpretability of an ensemble multi-agent system for supervised learning. In *The 25th International Conference on Principles and Practice of Multi-Agent Systems (PRIMA 2024)*, pages 335–350, Kyoto, 2025. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-77367-9_26.

[24] Guido Boella, Leendert Van Der Torre, and Harko Verhagen. Introduction to normative multiagent systems. *Computational & Mathematical Organization Theory*, 12(2):71–79, 2006.

[25] Berk Buzcu, Emre Kuru, Davide Calvaresi, and Reyhan Aydoğan. Evaluation of the user-centric explanation strategies for interactive recommenders. In *Explainable and Transparent AI and Multi-Agent Systems (EXTRAAMAS)*, pages 21–38, Auckland, 2024. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-70074-3_2.

[26] Turgay Caglar, Sarath Sreedharan, and Mor Vered. Who am I dealing with? explaining the designer's hidden intentions. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 436–444, Detroit, 2025. IFAAMAS. https://doi.org/10.5555/3709347.3743558.

[27] Davide Calvaresi, Yazan Mualla, Amro Najjar, Stéphane Galland, and Michael Schumacher. Explainable multi-agent systems through blockchain technology. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems (EXTRAAAMAS)*, pages 41–58, Montréal, 2019. Springer International Publishing. https://doi.org/10.1007/978-3-030-30391-4_3.

[28] Davide Calvaresi, Giovanni Ciatto, Amro Najjar, Reyhan Aydoğan, Leon Van der Torre, Andrea Omicini, and Michael Schumacher. Expectation: Personalized explainable artificial intelligence for decentralized agents with heterogeneous knowledge. In *Explainable and Transparent AI and Multi-*

Agent Systems (EXTRAAMAS), pages 331–343, Online, 2021. Springer International Publishing. https://doi.org/10.1007/978-3-030-82017-6_20.

[29] Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. Why do multi-agent LLM systems fail?, 2025. URL https://arxiv.org/abs/2503.13657.

[30] Tathagata Chakraborti, Anagha Kulkarni, Sarath Sreedharan, David E. Smith, and Subbarao Kambhampati. Explicability? Legibility? Predictability? Transparency? Privacy? Security? The emerging landscape of interpretable agent behavior. *Proceedings of the International Conference on Automated Planning and Scheduling*, 29:86–96, 2019.

[31] Jessie Y. C. Chen, Shan G. Lakhmani, Kimberly Stowers, Anthony R. Selkowitz, Julia L. Wright, and Michael Barnes. Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical Issues in Ergonomics Science*, 19(3):259–282, 2018. https://doi.org/10.1080/1463922X.2017.1315750.

[32] Giovanni Ciatto, Roberta Calegari, Andrea Omicini, and Davide Calvaresi. Towards XMAS: eXplainability through multi-agent systems. In *Proceedings of the 1st workshop on Artificial Intelligence and Internet of things (AI & IoT 2019)*, pages 1–14, Rende, Italy, November 2019. CEUR Workshop Proceedings.

[33] Giovanni Ciatto, Michael I. Schumacher, Andrea Omicini, and Davide Calvaresi. Agent-based explanations in AI: Towards an abstract framework. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, pages 3–20, Auckland, 2020. Springer International Publishing. https://doi.org/10.1007/978-3-030-51924-7_1.

[34] Giovanni Ciatto, Matteo Magnini, Berk Buzcu, Reyhan Aydoğan, and Andrea Omicini. A general-purpose protocol for multi-agent based explanations. In *Explainable and Transparent AI and Multi-Agent Systems (EXTRAAMAS)*, pages 38–58, London, 2023. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-40878-6_3.

[35] Daniel E. Collins, Conor J. Houghton, and Nirav Ajmeri. Social value orientation and integral emotions in multi-agent systems. In *Proceedings of the International Workshop on Coordination, Organizations, Institutions, Norms and Ethics for Governance of Multi-Agent Systems (COINE)*, Lecture Notes in Computer Science, pages 118–138, London, May 2023. Springer. https://doi.org/10.1007/978-3-031-49133-7_7.

[36] Kristijonas Čyras, Myles Lee, and Dimitrios Letsios. Schedule explainer: An argumentation-supported tool for interactive explanations in makespan scheduling. In *Explainable and Transparent AI and Multi-Agent Systems (EXTRAAMAS)*, pages 243–259, Online, 2021. Springer International Publishing. https://doi.org/10.1007/978-3-030-82017-6_15.

[37] Alaa Daoud, Hiba Alqasir, Yazan Mualla, Amro Najjar, Gauthier Picard, and Flavien Balbo. Towards explainable recommendations of resource allocation mechanisms in on-demand transport fleets. In *Ex-*

plainable and Transparent AI and Multi-Agent Systems (EXTRAAMAS), pages 97–115, Online, 2021. Springer International Publishing. https://doi.org/10.1007/978-3-030-82017-6_7.

[38] Marian David. The correspondence theory of truth. In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Stanford, 2025. URL https://plato.stanford.edu/archives/sum2025/entries/truth-correspondence/.

[39] Douglas Davis, Oleg Korenok, and Robert Reilly. Cooperation without coordination: Signaling, types and tacit collusion in laboratory oligopolies. *Experimental Economics*, 13(1):45–65, March 2010. https://doi.org/10.1007/s10683-009-9228-6.

[40] Hans de Bruijn, Martijn Warnier, and Marijn Janssen. The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. *Government Information Quarterly*, 39(2):101666, April 2022. https://doi.org/10.1016/j.giq.2021.101666.

[41] Maartje M. A. de Graaf and F Malle, Bertra. How people explain action (and autonomous intelligent systems should too). In *Artificial Intelligence for Human-Robot Interaction AAAI fall symposia*, pages 19–26, Arlington, Virginia, 2017. AAAI.

[42] Virginia Dignum. *Responsible Artificial Intelligence*. Number 1 in Artificial Intelligence: Foundations, Theory, and Algorithms. Springer Cham, New York, November 2019. https://doi.org/10.1007/978-3-030-30371-6.

[43] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017. URL https://arxiv.org/abs/1702.08608.

[44] Kshama Dwarakanath, Svitlana Vyetrenko, Tucker Balch, and Toks Oyebode. Transparency as delayed observability in multi-agent systems. In *Proceedings of the Winter Simulation Conference (WSC)*, pages 279–290, San Antonio, Texas, 2024. IEEE Press. https://doi.org/10.5555/3643142.3643165.

[45] Emre Erdogan, Frank Dignum, Rineke Verbrugge, and Pınar Yolum. TOMA: Computational theory of mind with abstractions for hybrid intelligence. *JAIR*, 25:285–311, January 2025. https://doi.org/10.1613/jair.1.16402.

[46] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke, and Effy Vayena. AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4):689–707, December 2018. https://doi.org/10.1007/s11023-018-9482-5.

[47] Kary Främling. Social explainable AI: What is it and how to make it happen with CIU? In *Explainable, Trustworthy, and Responsible AI and Multi-Agent Systems (EXTRAAMAS)*, pages 58–63, Detroit, 2026. Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-01399-6_4.

[48] Thiago Freitas dos Santos, Stephen Cranefield, Bastin Tony Roy Savarimuthu, Nardine Osman, and Marco Schorlemmer. Cross-community

adapter learning (cal) to understand the evolving meanings of norm violation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI)*, pages 109–117. International Joint Conferences on Artificial Intelligence Organization, August 2023. https://doi.org/10.24963/ijcai.2023/13.

[49] Thiago Freitas dos Santos, Nardine Osman, and Marco Schorlemmer. Can interpretability layouts influence human perception of offensive sentences? In *Explainable and Transparent AI and Multi-Agent Systems (EXTRAAMAS)*, pages 39–57, Auckland, 2024. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-70074-3_3.

[50] Drew Fudenberg and Eric Maskin. The folk theorem in repeated games with discounting or with incomplete information. *Econometrica*, 54:533–554, 1986. https://doi.org/10.2307/1911307.

[51] H.J. Gensler. *Ethics: A Contemporary Introduction (3rd ed.)*. Routledge, New York, 2017. https://doi.org/https://doi.org/10.4324/9781351231831.

[52] Claire Glanois, Paul Weng, Matthieu Zimmer, Dong Li, Tianpei Yang, Jianye Hao, and Wulong Liu. A survey on interpretable reinforcement learning. *Machine Learning*, 113:1–44, 2024.

[53] GovUK. A pro-innovation approach to AI regulation. Technical Report E02886733 03/23, Department for Science, Innovation, and Technology, 2023. URL https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach. White paper.

[54] Mara Graziani, Lidia Dutkiewicz, Davide Calvaresi, José Pereira Amorim, Katerina Yordanova, Mor Vered, Rahul Nair, Pedro Henriques Abreu, Tobias Blanke, Valeria Pulignano, John O. Prior, Lode Lauwaert, Wessel Reijers, Adrien Depeursinge, Vincent Andrearczyk, and Henning Müller. A global taxonomy of interpretable AI: Unifying the terminology for the technical and social sciences. *Artificial Intelligence Review*, 56(4):3473–3504, September 2022. https://doi.org/10.1007/s10462-022-10256-8.

[55] Herbert P. Grice. *Logic and Conversation*, pages 41–58. Brill, Leiden, The Netherlands, 1 edition, 1975. https://doi.org/https://doi.org/10.1163/9789004368811_003.

[56] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *CSUR*, 51(5), August 2018. https://doi.org/10.1145/3236009.

[57] Joseph Y. Halpern and Judea Pearl. Causes and explanations: A structural-model approach. part ii: Explanations. *The British Journal for the Philosophy of Science*, 56:889–911, 2005. URL http://www.jstor.org/stable/3541871.

[58] Ross A. Hammond and Robert Axelrod. The evolution of ethnocentrism. *The Journal of Conflict Resolution*, 50(6):926–936, 2006. URL http://www.jstor.org/stable/27638531.

[59] Maaike Harbers, Jeffrey M. Bradshaw, Matthew Johnson, Paul Feltovich, Karel van den Bosch, and John-Jules Meyer. Explanation in human-agent teamwork. In *Coordination, Organizations, Institutions,*

and Norms in Agent System VII (COINE), pages 21–37, Lyon, 2012. https://doi.org/10.1007/978-3-642-35545-5_2.

[60] David Heald. Varieties of transparency. In *Transparency: The Key to Better Governance?* British Academy, New York, September 2006. https://doi.org/10.5871/bacad/9780197263839.003.0002.

[61] Denis J. Hilton. A conversational model of causal explanation. *European Review of Social Psychology*, 2(1):51–81, 1991. https://doi.org/10.1080/14792779143000024.

[62] Michael Hind, Dennis Wei, Murray Campbell, Noel C. F. Codella, Amit Dhurandhar, Aleksandra Mojsilović, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. TED: Teaching AI to explain its decisions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pages 123–129, Honolulu, 2019. ACM. https://doi.org/10.1145/3306618.3314273.

[63] Robert R. Hoffman, Shane T. Mueller, Gary Klein, Mohammadreza Jalaeian, and Connor Tate. Explainable AI: Roles and stakeholders, desirements and challenges. *Frontiers in Computer Science*, 5:1–18, 2023. https://doi.org/10.3389/fcomp.2023.1117848.

[64] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science*, 5, 2023. https://doi.org/10.3389/fcomp.2023.1096257.

[65] Joris Hulstijn, Igor Tchappi, Amro Najjar, and Reyhan Aydoğan. Metrics for evaluating explainable recommender systems. In *Explainable and Transparent AI and Multi-Agent Systems (EXTRAAMAS)*, pages 212–230, London, 2023. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-40878-6_12.

[66] IEEE. Ethically aligned design - a vision for prioritizing human wellbeing with autonomous and intelligent systems. *Ethically Aligned Design - A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, 1(1):1–294, 2019. URL https://ieeexplore-ieee-org.bris.idm.oclc.org/servlet/opac?punumber=9398611.

[67] IEEE. IEEE standard for transparency of autonomous systems. *IEEE Std 7001-2021*, 1(1):1–54, 2022. https://doi.org/10.1109/IEEESTD.2022.9726144.

[68] Nusrath Jahan and Johnathan Mell. Decoding negotiation dynamics: The impact of opponent identity and privacy on strategy, deception, and emotional transparency in human-agent interaction. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 2562–2564, Detroit, 2025. IFAAMAS. https://doi.org/10.5555/3709347.3743937.

[69] Sophie F. Jentzsch, Sviatlana Höhn, and Nico Hochgeschwender. Conversational interfaces for explainable AI: A human-centred approach. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems (EXTRAAMAS)*, pages 77–92, Montréal, 2019. Springer International Publishing. https://doi.org/10.1007/978-3-030-30391-4_5.

[70] Özgür Kafalı, Nirav Ajmeri, and Munindar P. Singh. Revani: Revision and verification of normative specifications for privacy. *IEEE Intelligent Systems*, 31(5):8–15, September 2016.

[71] Selim Karaoğlu, Marina Katoh, Titash Majumdar, Ethan Beaird, Feyza Merve Hafızoğlu, and Sandip Sen. Effect of agent explanations using warm and cold language on user adoption of recommendations for bandit problems. In *Explainable and Transparent AI and Multi-Agent Systems (EXTRAAMAS)*, pages 3–20, Auckland, 2024. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-70074-3_1.

[72] Dilara Kekulluoglu, Michael Rovatsos, and Nadin Kökciyan. Answerable sociotechnical systems. In *Proceedings of the 27th European Conference on Artificial Intelligence (ECAI)s*, volume 392, pages 4451–4454, Santiago de Compostela, Spain, 2024. Frontiers in Artificial Intelligence and Applications. https://doi.org/10.3233/FAIA241027.

[73] Barbara Kitchenham and Stuart Charters. Guidelines for performing systematic literature reviews in software engineering. Technical report, Keele University and Durham University Joint Report, January 2007.

[74] Avinash Kori, Antonio Rago, and Francesca Toni. Free argumentative exchanges for explaining image classifiers. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1172–1180, Detroit, 2025. IFAAMAS. https://doi.org/10.5555/3709347.3743638.

[75] Sarit Kraus, Amos Azaria, Jelena Fiosina, Maike Greve, Noam Hazon, Lutz M. Kolbe, Tim-Benjamin Lembcke, Jörg P. Müller, Sören Schleibaum, and Mark Vollrath. AI for explaining decisions in multiagent environments. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, volume 34, pages 13534–13538, New York, 2020. https://doi.org/10.1609/aaai.v34i09.7077.

[76] Maya Krishnan. Against interpretability: a critical examination of the interpretability problem in machine learning. *Philosophy & Technology*, 33 (3):487–502, September 2020. https://doi.org/10.1007/s13347-019-00372-9.

[77] Haoran Li, Xusen Cheng, and Xiaoping Zhang. Accurate insights, trustworthy interactions: Designing a collaborative AI-human multi-agent system with knowledge graph for diagnosis prediction. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, Yokohama, Japan, 2025. ACM. https://doi.org/10.1145/3706598.3713526.

[78] Zhaoxing Li, Vahid Yazdanpanah, Stefan Sarkadi, Yulan He, Elnaz Shafipour, and Sebastian Stein. Towards citizen-centric multiagent systems based on large language models. In *Proceedings of the 2024 International Conference on Information Technology for Social Good*, GoodIT '24, pages 26–31, Bremen, Germany, 2024. ACM. https://doi.org/10.1145/3677525.3678636.

[79] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris B. Kotsiantis. Explainable AI: A review of machine learn-

ing interpretability methods. *Entropy*, 23, 2020. URL https://api.semanticscholar.org/CorpusID:229722844.

[80] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pages 3122–3130, Stockholm, July 2018. PMLR. URL https://proceedings.mlr.press/v80/lipton18a.html.

[81] Zachary C. Lipton. The mythos of model interpretability. *Commun. ACM*, 61(10):36–43, September 2018. https://doi.org/10.1145/3233231.

[82] Enrico Liscio, Roger Lera-Leri, Filippo Bistaffa, Roel I.J. Dobbe, Catholijn M. Jonker, Maite López-Sánchez, Juan A. Rodríguez-Aguilar, and Pradeep K. Murukannaiah. Value inference in sociotechnical systems. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pages 1774–1780, London, 2023. IFAAMAS.

[83] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, pages 4768–4777, Long Beach, California, USA, 2017. Curran Associates Inc. https://doi.org/10.5555/3295222.3295230.

[84] Bertram F. Malle. *How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction*. The MIT Press, September 2004. https://doi.org/10.7551/mitpress/3586.001.0001.

[85] Leila Methnani, Andreas Antoniades, and Andreas Theodorou. Embracing AWKWARD! Real-time adjustment of reactive plans using social norms. In *Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems XV (COINE)*, pages 54–72, Online, 2022. Springer International Publishing. https://doi.org/10.1007/978-3-031-20845-4_4.

[86] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019. https://doi.org/https://doi.org/10.1016/j.artint.2018.07.007.

[87] Tim Miller, Piers Howe, and Liz Sonenberg. Explainable AI: Beware of inmates running the asylum. In *IJCAI-17 Workshop on Explainable AI (XAI)*, Melbourne, 2017.

[88] Andreasa Morris-Martin, Marina De Vos, and Julian Padget. Norm emergence in multiagent systems: A viewpoint paper. *Autonomous Agents and Multi-Agent Systems (JAAMAS)*, 33(6):706–749, 2019.

[89] Francesca Mosca, Ştefan Sarkadi, Jose M. Such, and Peter McBurney. Agent EXPRI: Licence to explain. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems (EXTRAAMAS)*, pages 21–38, Auckland, 2020. Springer International Publishing. https://doi.org/10.1007/978-3-030-51924-7_2.

[90] W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Definitions, methods, and applications in interpretable machine

learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, 2019. https://doi.org/10.1073/pnas.1900654116.

[91] Pradeep K. Murukannaiah and Munindar P. Singh. From machine ethics to Internet ethics: Broadening the horizon. *IEEE Internet Computing*, 24 (3):51–57, May 2020. https://doi.org/10.1109/MIC.2020.2989935.

[92] Kartik Nagpal, Dayi Dong, and Negar Mehr. Leveraging large language models for effective and explainable multi-agent credit assignment. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1501–1510, Detroit, 2025. IFAAMAS. https://doi.org/10.5555/3709347.3743784.

[93] Suchismita Naik, Austin L. Toombs, Ph.D. Snellinger, Amanda, Scott Saponas, and Amanda K Hall. Designing with multi-agent generative AI: Insights from industry early adopters. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference (DIS)*, pages 1961–1972, Funchal, Madeira, 2025. ACM. https://doi.org/10.1145/3715336.3735823.

[94] Mohammad Naiseh, Catherine Webb, Tim Underwood, Gopal Ramchurn, Zoe Walters, Navamayooran Thavanesan, and Ganesh Vigneswaran. XAI for group-AI interaction: Towards collaborative and inclusive explanations. In *Joint Proceedings of the xAI 2024 Late-breaking Work, Demos and Doctoral Consortium*, volume 3793, pages 249–256, Valletta, Malta, July 2024. CEUR.

[95] Dundefinedng Nguyen, Ariel Vetzler, Sarit Kraus, and Anil Vullikanti. Contrastive explainable clustering with differential privacy. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1548–1556, Detroit, 2025. IFAAMAS. https://doi.org/10.5555/3709347.3743789.

[96] Eura Nofshin. Leveraging interpretable human models to personalize AI interventions for behavior change. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 2761–2763, Auckland, New Zealand, 2024. IFAAMAS. https://doi.org/10.5555/3635637.3663279.

[97] Matthew J. O'Laughlin and Bertram F. Malle. How people explain actions performed by groups and individuals. *Journal of Personality and Social Psychology*, 82(1):33–48, 2002. https://doi.org/10.1037/0022-3514.82.1.33.

[98] Andrés Páez. The pragmatic turn in explainable artificial intelligence (XAI). *Minds and Machines*, 29(3):441–459, September 2019. https://doi.org/10.1007/s11023-019-09502-w.

[99] Charles Chimwemwe Phiri. Creating characteristically auditable agentic AI systems. In *Proceedings of the Intelligent Robotics FAIR 2025*, pages 1–14, Budapest, 2025. ACM. https://doi.org/10.1145/3759355.3759356.

[100] Emilee Rader, Kelley Cotter, and Janghee Cho. Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 36th CHI Conference on Human Factors in Computing Systems*, pages 1–13, Montréal, 2018. ACM. https://doi.org/10.1145/3173574.3173677.

[101] Antonio Rago, Hengzhi Li, and Francesca Toni. Interactive explanations by conflict resolution via argumentative exchanges. In *Proceedings of the In-*

ternational Conference on Knowledge Representation and Reasoning (KR), pages 582–592, Rhodes, January 2023. AAAI.

[102] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, San Francisco, 2016. ACM. https://doi.org/10.1145/2939672.2939778.

[103] Mireia Ribera and Agata Lapedriza. Can we do better explanations? A proposal of user-centered explainable AI. In *Joint Proceedings of the ACM IUI 2019 Workshops*, Los Angeles, 2019. ACM.

[104] Pamela Robinson. Moral disagreement and artificial intelligence. *AI and Society*, 38(3):1–14, June 2023. ISSN 1435-5655. https://doi.org/10.1007/s00146-023-01697-y.

[105] Avi Rosenfeld and Ariella Richardson. Explainability in human–agent systems. *Autonomous Agents and Multi-Agent Systems*, 33(6):673–705, November 2019. ISSN 1573-7454. https://doi.org/10.1007/s10458-019-09408-y.

[106] Silvia Rossi, Alessandra Rossi, and Kerstin Dautenhahn. The secret life of robots: Perspectives and challenges for robot's behaviours during non-interactive tasks. *International Journal of Social Robotics*, 12:1265–1278, 2020. URL https://api.semanticscholar.org/CorpusID:218994603.

[107] James Rudd-Jones, Mirco Musolesi, and María Pérez-Ortiz. Multi-agent reinforcement learning simulation for environmental policy synthesis. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 2890–2895, Detroit, 2025. IFAAMAS. https://doi.org/10.5555/3709347.3744041.

[108] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, May 2019. https://doi.org/10.1038/s42256-019-0048-x.

[109] Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Education Limited, 2016.

[110] Fardin Saad, Pradeep K. Murukannaiah, and Munindar P. Singh. Gricean norms as a basis for effective collaboration. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1812–1820, Detroit, 2025. IFAAMAS. https://doi.org/10.5555/3709347.3743817.

[111] Albert Sadowski and Jaroslaw A. Chudziak. On verifiable legal reasoning: A multi-agent framework with formalized knowledge representations. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 2535–2545, Seoul, 2025. ACM. https://doi.org/10.1145/3746252.3761057.

[112] Wesley C. Salmon. *Scientific Explanation and the Causal Structure of the World*. Princeton University Press, Princeton, 1985. https://doi.org/10.1515/9780691221489.

[113] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *ITU Journal: ICT Discoveries*, 1(5):1–10, 2017.

[114] Lindsay Sanneman and Julie A. Shah. A situation awareness-based framework for design and evaluation of explainable AI. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems (EXTRAAMAS)*, pages 94–110, Auckland, 2020. Springer International Publishing. https://doi.org/10.1007/978-3-030-51924-7_6.

[115] Shalom H Schwartz. An overview of the schwartz theory of basic values. *Online readings in Psychology and Culture*, 2(1):2307–0919, 2012.

[116] Andrew D. Selbst and Julia Powles. Meaningful information and the right to explanation. *International Data Privacy Law*, 7(4): 233–242, December 2017. https://doi.org/10.1093/idpl/ipx022. URL https://doi.org/10.1093/idpl/ipx022.

[117] Naziha Sendi, Nadia Abchiche-Mimouni, and Farida Zehraoui. Towards a transparent deep ensemble method based on multiagent argumentation. In Davide Calvaresi, Amro Najjar, Michael Schumacher, and Kary Främling, editors, *Explainable, Transparent Autonomous Agents and Multi-Agent Systems (EXTRAAMAS)*, pages 3–21, Montréal, 2019. Springer International Publishing. https://doi.org/10.1007/978-3-030-30391-4_1.

[118] Pedro Sequeira and Melinda Gervasio. Interestingness elements for explainable reinforcement learning: Understanding agents' capabilities and limitations. *Artificial Intelligence*, 288:103367, 2020. ISSN 0004-3702. https://doi.org/10.1016/j.artint.2020.103367.

[119] Marc Serramia, Manel Rodriguez-Soto, Maite López-Sánchez, Juan A. Rodríguez-Aguilar, Filippo Bistaffa, Paula Boddington, Michael Wooldridge, and Carlos Ansotegui. Encoding Ethics to Compute Value-Aligned Norms. *Minds and Machines*, 33:1–30, November 2023. ISSN 1572-8641. https://doi.org/10.1007/s11023-023-09649-7.

[120] Risal Shahriar Shefin, Md Asifur Rahman, Thai Le, and Sarra Alqahtani. xSRL: Safety-aware explainable reinforcement learning - safety as a product of explainability. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1932–1940, Detroit, 2025. IFAAMAS. https://doi.org/10.5555/3709347.3743830.

[121] Youyu Sheng, Yaoqin Gu, Jianqin Cao, Yuhan Liu, Xiaoyu Wang, Jiani Chen, Xianghong Sun, and Jingyu Zhang. Measuring understandability of intelligent systems: Scale development and validation across three domains. *International Journal of Human-Computer Studies*, 203:103592, 2025. https://doi.org/https://doi.org/10.1016/j.ijhcs.2025.103592.

[122] Maayan Shvo, Toryn Q. Klassen, and Sheila A. McIlraith. Towards the role of theory of mind in explanation. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems (EXTRAAMAS)*, pages 75–93, Auckland, 2020. Springer International Publishing. https://doi.org/10.1007/978-3-030-51924-7_5.

[123] Andrew Smart and Atoosa Kasirzadeh. Beyond model interpretability: socio-structural explanations in machine learning. *AI & SOCIETY*, 40(4): 2045–2053, April 2025. https://doi.org/10.1007/s00146-024-02056-1.

[124] Kacper Sokol and Peter Flach. One explanation does not fit all. *KI - Künstliche Intelligenz*, 34(2):235–250, June 2020. https://doi.org/10.1007/s13218-020-00637-y.

[125] Timo Speith and Jing Xu. Explainability and transparency in practice: A comparison between corporate and national AI ethics guidelines in Germany and China. In *Explainable and Transparent AI and Multi-Agent Systems (EXTRAAMAS)*, pages 205–223, Auckland, 2024. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-70074-3_12.

[126] Sarath Sreedharan, Anagha Kulkarni, and Subbarao Kambhampati. *Explainable Human-AI Interaction: A Planning Perspective*. Springer Cham, Switzerland, 2022. https://doi.org/10.1007/978-3-031-03767-2.

[127] Sebastian Stein and Vahid Yazdanpanah. Citizen-centric multiagent systems. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pages 1802–1807, London, 2023. IFAAMAS. https://doi.org/10.5555/3545946.3598843.

[128] Mayesha Tasnim, Paul Verhagen, Tobias Blanke, Erman Acar, and Sennay Ghebreab. Modeling strategic risk in school choices: A case for transparent design. *Proceedings of the 8th AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 8(3), 2025. https://doi.org/10.1609/aies.v8i3.36731.

[129] Gianluca Torta, Roberto Micalizio, and Samuele Sormano. Temporal multiagent plan execution: Explaining what happened. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems (EXTRAAMAS)*, pages 167–185, Montréal, 2019. Springer International Publishing. https://doi.org/10.1007/978-3-030-30391-4_10.

[130] Mark Towers, Yali Du, Christopher Freeman, and Tim Norman. Temporal explanations of deep reinforcement learning agents. In *Explainable and Transparent AI and Multi-Agent Systems (EXTRAAMAS)*, pages 99–115, Auckland, 2024. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-70074-3_6.

[131] Niko Tsakalakis, Sophie Stalla-Bourdillon, Dong Huynh, and Luc Moreau. A typology of explanations to support explainability-by-design. *ACM J. Responsib. Comput.*, 2(1), February 2025. https://doi.org/10.1145/3708504.

[132] Silvia Tulli, Filipa Correia, Samuel Mascarenhas, Samuel Gomes, Francisco S. Melo, and Ana Paiva. Effects of agents' transparency on teamwork. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems (EXTRAAMAS)*, pages 22–37, Montréal, 2019. Springer International Publishing. https://doi.org/10.1007/978-3-030-30391-4_2.

[133] Matteo Turilli and Luciano Floridi. The ethics of information transparency. *Ethics and Information Technology*, 11(2):105–112, June 2009. https://doi.org/10.1007/s10676-009-9187-9.

[134] Stephen Turner and Peter Olen. Normativity and social explanation. *Oxford Bibliographies in Philosophy*, 2022. https://doi.org/10.1093/obo/9780195396577-0235.

[135] Sz-Ting Tzeng, Nirav Ajmeri, and Munindar P. Singh. Norm enforcement with a soft touch: Faster emergence, happier agents. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1837–1846, Auckland, May 2024. IFAAMAS.

[136] Candace L. Upton. Virtue ethics and moral psychology: The situationism debate. *The Journal of Ethics*, 13(2):103–115, September 2009. https://doi.org/10.1007/s10892-009-9054-2.

[137] Harko Verhagen, Corinna Elsenbroich, and Nanda Wijermans. Agent decision-making heterogeneity—agent (meta)frameworks for agent-based modelling. In *Advances in Social Simulation*, pages 621–630, Cham, 2024. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-57785-7_48.

[138] Avish Vijayaraghavan and Cosmin Badea. Minimum levels of interpretability for artificial moral agents. *AI and Ethics*, 5(3):2071–2087, June 2025. https://doi.org/10.1007/s43681-024-00536-0.

[139] David S. Watson. Conceptual challenges for interpretable machine learning. *Synthese*, 200(2):65, March 2022. https://doi.org/10.1007/s11229-022-03485-5.

[140] Gerhard Weiss. *Multiagent systems*. Intelligent robotics and autonomous agents. The MIT Press, Cambridge, Massachusetts, second edition, 2013. URL http://site.ebrary.com/id/10674446.

[141] Adrian Weller. Transparency: Motivations and challenges. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 23–40. Springer International Publishing, Cham, 2019. https://doi.org/10.1007/978-3-030-28954-6_2.

[142] Licheng Wen, Pinlong Cai, Daocheng Fu, Song Mao, and Yikang Li. Bringing diversity to autonomous vehicles: An interpretable multi-vehicle decision-making and planning framework. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 2571–2573, London, 2023. IFAAMAS. https://doi.org/10.5555/3545946.3599005.

[143] Marcus Westberg, Amber Zelvelder, and Amro Najjar. A historical perspective on cognitive science and its influence on XAI research. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems (EXTRAAMAS)*, pages 205–219, Montréal, 2019. Springer International Publishing. https://doi.org/10.1007/978-3-030-30391-4_12.

[144] Michael Winikoff and Galina Sidorenko. Evaluating a mechanism for explaining bdi agent behaviour. In *Explainable and Transparent AI and Multi-Agent Systems (EXTRAAMAS)*, pages 18–37, Cham, 2023. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-40878-6_2.

[145] Michael Winikoff, John Thangarajah, and Sebastian Rodriguez. A scoresheet for explainable AI. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent*

Systems (AAMAS), pages 2171–2180, Detroit, 2025. IFAAMAS. https://doi.org/10.5555/3709347.3743856.

[146] Jessica Woodgate and Nirav Ajmeri. Macro ethics for governing equitable sociotechnical systems. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1824–1828, Online, May 2022. IFAAMAS. https://doi.org/10.5555/3535850.3536118. Blue Sky Ideas Track.

[147] Jessica Woodgate and Nirav Ajmeri. Macro ethics principles for responsible AI systems: Taxonomy and directions. *CSUR*, 56(289):1–37, July 2024. ISSN 0360-0300. https://doi.org/10.1145/3672394.

[148] Jessica Woodgate and Nirav Ajmeri. Combining normative ethics principles to learn prosocial behaviour. In *Proceedings of the 24th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, Detroit, 2025. IFAAMAS.

[149] Wright. *Norm and Action: A Logical Enquiry*. Humanities, New York, 1963.

[150] Yifan Xu, Joe Collenette, Louise Dennis, and Clare Dixon. Dialogue explanations for rule-based AI systems. In *Explainable and Transparent AI and Multi-Agent Systems (EXTRAAMAS)*, pages 59–77, London, 2023. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-40878-6_4.

[151] Elena Yan, Samuele Burattini, Jomi Fred Hübner, and Alessandro Ricci. A multi-level explainability framework for engineering and understanding bdi agents. *Autonomous Agents and Multi-Agent Systems*, 39(1):9, January 2025. ISSN 1573-7454. https://doi.org/10.1007/s10458-025-09689-6.

[152] Fan Yang, Kai He, Linxiao Yang, Hongxia Du, Jingbang Yang, Bo Yang, and Liang Sun. Learning interpretable decision rule sets: A submodular optimization approach. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 27890–27902, Online, 2021. Curran Associates, Inc.

[153] Vahid Yazdanpanah, Enrico Gerding, Sebastian Stein, Mehdi Dastani, Catholijn M. Jonker, and Timothy Norman. Responsibility research for trustworthy autonomous systems. In *Proceedings of the 20th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pages 57–62, Virtual Event, London, May 2021. IFAAMAS. URL https://eprints.soton.ac.uk/447511/.

[154] Yue Yin. Too much information? Investigating information disclosure in auction systems with LLM simulations. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, Yokohama, Japan, 2025. ACM. https://doi.org/10.1145/3706599.3720022.

[155] Petri Ylikoski and Jaakko Kuorikoski. Dissecting explanatory power. *Philosophical Studies*, 148(2):201–219, March 2010. ISSN 1573-0883. https://doi.org/10.1007/s11098-008-9324-z.

[156] Renos Zabounidis, Joseph Campbell, Simon Stepputtis, Dana Hughes, and Katia P. Sycara. Concept learning for interpretable multi-agent reinforcement learning. In *Proceedings of The 6th Confer-*

ence on Robot Learning, volume 205 of *Proceedings of Machine Learning Research*, pages 1828–1837, Auckland, 2023. PMLR. URL https://proceedings.mlr.press/v205/zabounidis23a.html.

[157] Carlos Zednik. Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & Technology*, 34(2):265–288, June 2021. https://doi.org/10.1007/s13347-019-00382-7.

[158] Zeren Zhang, Zhiwei Xu, Guangchong Zhou, Dapeng Li, Bin Zhang, and Guoliang Fan. Unveiling decision intention for cooperative multi-agent reinforcement learning. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 2345–2354, Detroit, 2025. IFAAMAS. https://doi.org/10.5555/3709347.3743875.

A Methodology

The literature search resulting in the body of work surveyed in this paper was conducted on 05/08/2025. We conducted an initial search string from preliminary research. Using a forwards and backwards snowballing technique [73], we searched selected resources (the University of Bristol Library, ACM Digital Library, and Google Scholar) using our search string. We applied inclusion and exclusion criteria to identify primary studies and followed relevant citations to expand the search. For further details of the methodology, see Appendix A

Our search string contains two main components. The first component relates to autonomous agents and multi-agent systems. The second component relates to transparency, interpretability, and explainability. The search string used was ('Agent' OR 'Multiagent' OR 'Multi-agent') AND ('Interpret*' OR 'Expla*' OR 'Transpar*').

We searched the selected resources (the University of Bristol Library, ACM Digital Library, and Google Scholar) with our search string. We used a forwards and backwards snowballing technique, applying the inclusion and exclusion criteria to the titles of the first five pages of each resource. This narrowed our search to a smaller selection of works. We applied the inclusion and exclusion criteria to the abstracts of these works and identified our primary studies. From our primary studies, relevant citations were followed to expand the search. The inclusion and exclusion criteria can be found in Table 1.

B Ethical Considerations

Appropriate implementation of transparency, interpretability, and explainability can change according to the situation. Ethical concerns arise, as incorrect application of each term can create opportunity for harm. For example, too much transparency can compromise privacy, whilst too much focus on interpretability can prioritise understandability above truthfulness. Other times, too little focus can reduce human autonomy by withholding important information. Careful consideration is needed to appropriately balance competing concerns.

Table 1. Inclusion and Exclusion Criteria

Inclusion Criteria	**Exclusion Criteria**
Peer reviewed and published works from ACM CSUR, AIES, FAccT, AAAI, IJCAI, (J)AAMAS, TAAS, TIST, JAIR, AIJ, Nature, Science	Works that have not been peer reviewed, gray literature, duplicate studies, or older versions of the same research
Autonomous agents or multi-agent systems works which explicitly concern interpretability, explainability, or transparency	Research that involves interpretability, explainability, or transparency solely for AI or ML and not autonomous agents or multi-agent systems

Interpretability. Techniques to imbue interpretability come with inherent risks of information loss and meaning mistranslation through too much complexity or simplicity. A key aim of interpretability is to help an observer create the right mental model of a system. Yet, many agent decision-making mechanisms involve approximations of data entailing that some information is inherently lost in the decision-making mechanism itself. Information is again lost when that decision-making mechanism is translated into an interpretation, either through post hoc techniques or by an observer forming an understanding of an intrinsically interpretable decision-making mechanism. Post hoc interpretations do not elucidate precisely how a decision-making mechanism works, which means that they may be misleading depending on what they are focusing on [81]. As interpretation is necessarily situated between object and observer, there is an inevitable sacrifice of meaning giving rise to a risk of omission during interpretation that could be due to either oversimplification or overcomplexity. Multiple levels of information loss entails several differing interpretations are possible for the same observation [54].

Explainability. Explanations must convey a suitable amount of information given the setting, considering some detail must be sacrificed in translating a decision-making mechanism into an understandable communication, yet losing too much risks misleading observers. Humans often do not need complete causal chains of explanation, however, this opens up ethical issues regarding intentional concealing of information [54]. Information may be concealed simply by providing explanations that require prior expert knowledge which thereby limits understandability [11]. Explanations are often approximations which do not have perfect fidelity with respect to the original decision-making mechanism, as if the explanation was completely faithful to what the original decision-making mechanism computes, the explanation would equal the decision-making mechanism. By virtue of being approximations, explanations will therefore be sometimes wrong. Even if the explanation is correct, it is often unavoidable to leave out some details of how the decision was reached [108]. To avoid installing a false sense of confidence in the explanation and object being explained, it is thus im-

portant to consider the likelihood that the explanation is accurate as well as the scope of what the explanation is attempting to address, being clear about the limitations of an object and the context the object exists within [63].

Transparency. A lack of transparency may diminish the trust of users [91]. Goal obfuscation primarily focusses on not revealing true intentions, but not necessarily actively misleading. Deception may involve obfuscation as well as active misleading [30]. Alternatively, too much transparency could result in systems being gamed if the underlying logic is fully available [54, 128]. Actors with misaligned interests can abuse transparency by using it to manipulate or else inappropriately apply the information gained. Observations of agent or human behaviour could be employed to infer information that would not otherwise be shared, which may then be used maliciously and create real world risks if shared.

Difficulties arise depending on whether truth is understood as derived from reason (positivist maxim), or relationally, derived from human practice (pragmatic maxim) [11]. The positivist maxim perceives transparency as a conveyance of truth and rests on the epistemological assumption that truth is correspondence to a fact [38]. This conception of truth entails that the more facts that are revealed, the more truth that can be known. The more that is known about a system's workings, the more defensibly it can be governed and held accountable. The pragmatic maxim, on the other hand, sees truth as meaning achieved through relations, emphasising the importance of understanding a system's connections to the environment as well as its internal workings [11].

The relational perspective of truth illuminates how transparency alone cannot create accountability. Ten important limitations of transparency to consider include: disconnection from power, where making information transparent has no meaningful effect; harmfulness, by threatening privacy; inadvertent or strategic opacity by making too much information visible; creating false binaries between complete secrecy and total openness; placing too much burden on the individual to seek out information; transparency does not necessarily build trust; transparency can be performative; seeing inside an object does not necessitate understanding; there can be technical limitations related to the scale and speed of decision-making mechanisms; there are temporal limitations as systems change over time and different moments in time may require different kinds of accountability [11].