Responsibility of AI Systems

Mehdi Dastani[1] and Vahid Yazdanpanah[*2]

[1] *Utrecht University*
[2] *University of Southampton*

March 28, 2022

Abstract

To support the trustworthiness of AI systems, it is essential to have precise methods to determine what or who is to account for the behaviour, or the outcome, of AI systems. The assignment of responsibility to an AI system is closely related to the identification of individuals or elements that have caused the outcome of the AI system. In this work, we present an overview of approaches that aim at modelling responsibility of AI systems, discuss their advantages and shortcomings to deal with various aspects of the notion of responsibility, and present research gaps and ways forward.

keywords: Responsibility Modelling; Actual Responsibility; Responsibility of AI Systems; Trustworthy Autonomous Systems.

1 Introduction

The rapid development of Artificial Intelligence (AI) systems and their increasing impact on our daily life are unprecedented. Autonomous systems such as drones and self-driving cars, (semi)automatic decision support systems being used in hospitals, financial markets, and courtrooms, and the use of machine learning techniques to model social and natural phenomena (e.g., climate change extremes or global pandemics) are nowadays among our daily practices. The impact of AI raises various questions related to the trustworthiness and accountability of artificial intelligence systems (Ramchurn et al., 2021; Chopra and Singh, 2021). Who or which part of the system is responsible if something goes wrong? Why does a decision support system propose certain decisions or predict a certain outcome? And, how can we ensure that the automatically generated decisions and predictions are taken in a human-centred responsible manner, in the sense that they are not based on accidental correlations or any undesirable or unknown bias. We argue that answering such questions is necessary for ensuring the trustworthiness of AI systems and, to that end, (formal) approaches to model and reason about responsibility can contribute.

[*] Corresponding Author: v.yazdanpanah@soton.ac.uk

In the AI literature, the notion of responsibility, and related concepts such as accountability and blameworthiness, is used in various ways with different meanings. For example, the question of who or which part of an AI system is responsible for a certain outcome is different from the question of how to use AI technologies responsibly. The former relates to causal chains and abilities of the involved components, including human and artificial systems, while the latter focuses on how various stakeholders, e.g., the designers or the the final users of AI systems, take into account social, legal, and ethical issues such as privacy laws, social fairness, and bias. The former notion of responsibility (van de Poel, 2011) involves a much fundamental concept of causality (Pearl, 2009)—the type of which may explain various uses of the notion of responsibility in the AI literature. In some cases, one may be interested in questions like which particular AI systems, or which parts of them, have caused a particular outcome. While in other cases, one may be interested in questions like whether certain data-driven decisions or predictions are caused by accidental correlations in the data and being free of undesirable or unknown biases in the data. In the first case, we may be interested in particular events resulting in a particular outcome, while in the second case we may be interested in understanding the causal relations between events in populations.

As noted by Halpern (2016), causal theories distinguish two forms of causality: type and actual causality. Type causality concerns general statements such as "smoking causes cancer" and the actual causality concerns specific statements such as "John have caused the accident". In general, type causality is assumed to be concerned with populations, while actual causality is assumed to be concerned with particular materialised events and the individuals behind them. Moreover, type causality is often used in statistical machine learning that aims at building a causal model from data in order to make predictions. This notion of causality is also essential for some notion of responsible use of AI systems, e.g., to avoid making conclusions based on accidental correlations or bias (Pearl, 2009; Benjamins, 2021; Pearl and Mackenzie, 2018). In contrast, actual causality is often used to trace back the cause of a specific outcome or event, and to find out who or what has contributed to that cause, hence are responsible for the outcome and have to account for it. As explained by Halpern (2016), these two types of causality are intertwined in the sense that actual causality can help us to understand the reasons for a certain outcome, which can in turn be used to prevent those particular outcomes in future.

The notion of responsible use of AI in relation to *type* causality, which is central for the current data-driven AI systems, has been studied extensively (Smith, 2020; Dignum, 2019). In this paper, we ignore the notion of responsibility that is based on *type* causality and survey existing approaches that aim at modelling the notion of AI responsibility based on *actual* causality. We will use the term *actual responsibility* to refer to this notion of responsibility. Actual responsibility allows us to trace back the behaviour of AI systems and to assign responsibility to those AI systems, or their parts, that have contributed to the causation of a particular outcome. In other words, actual responsibility allows us to determine who or what is, and to what extent, responsible for the so-called algorithmic harm.[1] In this context, we understand algorithmic harm as a potential or already materialised harm resulted from applying algorithmic artefacts such as AI systems. For instance, decisions made by an algorithm that "drives" an autonomous vehicle (without or in collaboration with a human user) may result in harmful outcomes for the user herself or for others in the vicinity. We also ignore the question as to why a particular outcome is considered as harmful. The notion of harm caused by AI systems (e.g., classifiers, recommendation systems, decision support tools) due to for example bias, incomplete or imperfect data, or the condition under which AI systems are used, is extensively studied in the literature, see e.g., O'neil (2016); Safransky (2020); Cugurullo (2021); Macrorie et al. (2020).

Throughout this work, we use the example of autonomous vehicles to elaborate on various aspects of the problem but our overview applies to AI systems in general. We understand the behaviour of AI systems as a contextual phenomenon which requires extensive studies with appropriate degrees of granularity with respect to the domain of application, see e.g., Stilgoe (2018, 2020). However, our overview abstracts from contextual subtleties and highlights how the modelling perspectives affect modelling responsibility of AI systems. That is why we generally refer to concepts such as *action* and *event* in an abstract sense, and avoid articulating how they should be interpreted in a given context and semantics behind the occurrence of an action/event. Using responsibility models, one can reason about eventualities of interest and determine responsible agents in prospect or look at already materialised situations, retrospectively, and determine the responsibility of a component (e.g., an agent) in an AI system or of various components (e.g., agent groups) for a harmful outcome.

Against this background, this work provides an overview of various approaches to solve the actual responsibility ascription problem, discusses strengths and shortcomings of main approaches to responsibility modelling, and highlights new research directions. In this work, we will not delve into a technical comparison among various approaches but focus on how they aim at modelling responsibility of AI systems, e.g., whether they have an agent-oriented point of view or focus on event-based modelling. We discuss how such conceptual differences in perspective led to different results and flavours of responsibility. To that end, we use the philosophical literature on moral responsibility to elicit conditions for being responsible and link these to epistemic, motivational, and normative aspects of AI decision-making. This approach allows us to identify similarities among the existing approaches that aim at modelling responsibility of AI systems, and to highlight aspects that need further investigations.

[1] In particular, understanding the extent of artificial agents' responsibility and the ability to quantify them as responsibility degrees are key to bridging responsibility gaps and addressing the problem of many hands. In Section 6, we further elaborate on this aspect.

2 Actual Responsibility

In the AI literature on modelling actual responsibility, the following methodological perspectives are distinguished.

Event-Oriented Responsibility: This perspective, rooted in Halpern (2016) and Chockler and Halpern (2004), uses causal models to relate the chain of materialised events in an environment as the base for understanding which particular events have caused a specific outcome. This perspective considers responsibility as a measure of causality in the sense that some particular events are responsible for an outcome (or other events) to the extent that the outcome counterfactually depends on the events, i.e., the outcome would not have been realised if the event had not occurred. This perspective on responsibility can be applied to AI systems by considering the decisions/actions of the AI systems as events.

Agent-Oriented Responsibility: This perspective, rooted in Bulling and Dastani (2013) and Yazdanpanah and Dastani (2016), considers coalitional abilities in strategic settings as a base for seeing groups of agents responsible for an outcome in a multiagent environment. This view builds on Bratman's philosophical account (Bratman, 2013) that groups can intentionally act towards collective goals, hence are able to be considered responsible and accordingly account for their collective behaviour. This notion of responsibility captures the interaction among AI systems by considering their collective abilities, shared knowledge, and communication.

In the rest of this paper, we elaborate on these perspectives and their possible relations. In particular, we argue that the agent-oriented perspective on responsibility can be interpreted as an extension of the event-oriented responsibility by considering events as actions decided by the agents. The agentification of an event is modelled by capturing the reasons behind agents' actions, which can be explained in terms of the following concepts that are fundamental in decision-making.

- *Epistemic* capacities of agents: The knowledge of the agents, including their knowledge of their environment, abilities and strategies, is essential for their decision-making behaviours. Agents are assumed to be responsible for a certain outcome if they have knowingly contributed to the realisation of the outcome (Houlgate, 1968). We would like to emphasise that the notion of epistemic as used in Chockler and Halpern (2004) concerns the knowledge of a *reasoner* who is in charge of assigning responsibility to the acting agents. Although this notion of epistemic is also important for reasoning about responsibility assignment, we believe it is less general than the notion of epistemic in multiagent settings where the knowledge of agents or agent groups (e.g., common knowledge or distributed knowledge) are essential for responsibility assignment to agent groups.

- *Motivational* attitude of the agents: The motivation of agents is key in determining whether an agent has acted intentionally, which is in turn essential for the responsibility assignment problem. In the legal literature (where responsibility assignment is understood as liability ascription), an agent is liable only if the intentional connections can be established, meaning that the agent have acted knowingly and preferably (Petersen, 2013).
- *Normative* stance of the agents: The agents operating in an environment may be expected to respect norms of various kinds, e.g., legal, social, moral or rational norms. The assignment of responsibility to agents in an environment depends on the norms being in place, the mechanisms that enforce the norms, and the awareness of the agents about the norms and the enforcement mechanisms (Vargas, 2013).

For instance, imagine the scenario depicted in Figure 1. We see two autonomous vehicles $blue$ and red reaching at an intersection.[2] Their most preferred path of travel appear as solid arrows. Should both vehicles avoid going forward, as $Alice$ is blocking their most preferred path? Who is responsible if they both follow their preferences, and one hits $Alice$ first? Does the presence of the building that blocks the observability of $blue$ changes its responsibility for such an undesired event? Arguably, the highlighted aspects of decision-making—i.e., agents' epistemic capacity, their motivational attitude and preferences, as well as norms they adhere to—play a key role in the process of responsibility ascription. In the following, we look at these aspects and sociotechnical dynamics of responsibility of AI systems, discuss responsibility conditions and how modelling approaches to actual responsibility deal with them, and highlight advantages, shortcomings, and ways forward.

3 Responsibility Conditions

In the literature on moral philosophy, Braham and van Hees (2012) develop an agent-oriented account of actual responsibility. They present three conditions for assigning responsibility for an outcome to an agent. An agent is seen as responsible for an outcome if and only if (1) the agent is autonomous, intentional, and capable of distinguishing right and wrong and good and bad, (2) there exists a causal relation between the action of the agent and the outcome in question, and (3) the agent has had a reasonable opportunity to have done otherwise. While the second condition corresponds to causal dependencies in Halpern (2016) and Pearl (2009), the first condition not only determines that agency is necessary for being responsible but also demands that the epistemic capacities, motivations, and normative stance of the agent should be

[2] We consider autonomous vehicles as the AI software that controls the vehicles, not the physical car consisting of the engine, breaks, etc. In such a situation, an autonomous vehicle can decide to break even if the physical braking system is broken.

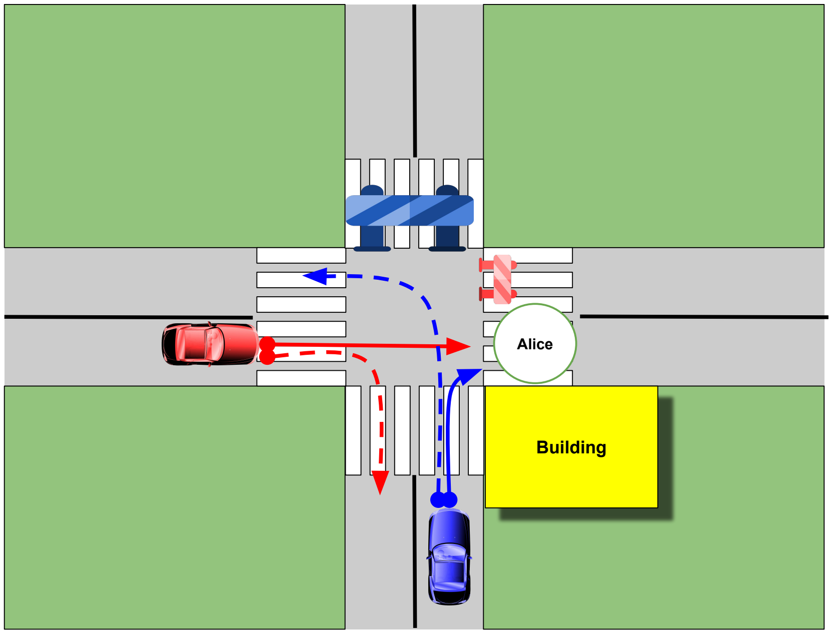

Figure 1: The Intersection Scenario: solid arrows represent the most preferred travel paths of the vehicles (*blue* and *red*). Dashed lines are their alternative (less preferred) options. Note that the two roadblocks are robust-enough to damage a vehicle that hits them and cause disastrous consequences for their riders. Here, *Alice* is a pedestrian, and the yellow building is an object that blocks the visibility of $blue$ on the road section it is intending to enter.

taken into account as well. The third condition (also known as the *avoidance potential* criterion or the principle of *alternative possibilities*) requires that causing an outcome is not sufficient to see one responsible if she had no other act available to her. For example, consider the scenario in Figure 1 where the self-driving vehicle $blue$ causes an accident with pedestrian $Alice$. In order to assign the responsibility of this outcome to the self-driving vehicle, the principle of alternative possibilities (Braham and van Hees, 2012) suggests that the vehicle can be held responsible if it had alternative decisions to prevent the outcome, e.g., to turn left (dashed blue line). In a hypothetical situation where the vehicle had no option to prevent the accident with the pedestrian, because for example the vehicle's physical break was broken and there were railroads on both sides of the road (i.e., turning left is not an option), then the principle of alternative possibilities suggests that it is unjustified to hold the autonomous vehicle (i.e., the AI software that controls the vehicle) responsible for the accident.

In the case of agentive entities with autonomy (Dastani et al., 2003), the notion of avoidance potential does not only depend on the agent's abilities, but also on its epistemic, motivational, and normative stance. For example, consider a revised scenario where the railroad did not exist on the sides of the road such that our self-driving vehicle $blue$ could decide to turn left, to go straight and hit the blue roadblock, or even run off the road, to prevent the accident with $Alice$. Suppose further investigations suggest that the car's decision was taken because it had not enough information to properly evaluate its current situation and to determine the feasibility or possible outcomes of alternative decisions, such as running off the road. This would be the case, for example, when the building is blocking its observability on $Alice$ (i.e., the vehicle had no information that turning right would cause an accident with $Alice$) or when the car could not conclude if running off the road is feasible because it did not have enough information about the existence of road rails on the sides of the road. In such a situation, the agent's (lack of) knowledge may eliminate all the alternative possibilities for avoiding the accident with the pedestrians and could therefore be considered as an acceptable excuse for the car to be relieved from responsibility of causing the accident with the pedestrians. Of course, one may argue that the responsibility can still be assigned to the car if the car had the opportunity to gain such knowledge. For instance, $blue$ possibly had the chance to communicate with other agents in the scene and gather information. In a future where self-driving vehicles are on the road, one can expect that $blue$ communicates with red, with the smart intersection coordination platform at the intersection, or with sensors/cameras installed on the building (that is blocking its view but has observability on $Alice$). Indeed, the assignment of responsibility requires a further step in analysis in order to determine if the car had the possibility to gain sufficient information to make an informed decision. Such an analysis may involve epistemic actions such as sense and communication actions, epistemic reasoning, and learning from experiences.

The assignment of responsibility may also depend on the preferences of the involved stakeholders on the consequence of alternative possibilities. In our running example, suppose that the self-driving vehicle's manufacturer has guaranteed the safety of passengers by designing the self-driving cars to prefer the safety of passengers above having them injured. The decision to run off the road or to hit the roadblock can therefore be considered by the car as a sub-optimal decision—in some cases the least preferred outcome—and thus not an alternative possibility. In a more complex multiagent scenario, the preferences of other stakeholders such as the traffic authority or other cars may be aligned or in conflict with the preference of our self-driving car. Of course, the fact that our car is designed to guarantee the safety of its own passengers may not be an ultimate reason to discharge its responsibility for the caused accident. The general question we like to pose here is whether the fact that the decisions of AI systems are driven by their preferences can influence the assignment of responsibility to them. More specifically, would an AI system be excused from responsibility if it is designed to behave according to the principle of economic rationality (i.e., maximising utility/preference)? In our running example, would the self-driving car be excused from responsibility of causing an accident with $Alice$ because its alternative decision to run off the road or hit the roadblock has been eliminated (or not considered) due to the fact that it was not aligned with its preference, i.e., the decision that caused an accident with $Alice$ has been the only/most economically rational decision for the self-driving car? In Figure 1, if $blue$ is not aware of $Alice$ and turning right in the intersection is a part of a shortest path to its destination, turning left would be an irrational choice as it may cause crashing into red. This leads to our next point on the importance of capturing norms, such as traffic rules and regulations, that are in place.

Imagine now that the self-driving vehicle $blue$ had enough information to know and conclude that running off the road is feasible and can prevent the accident. Moreover, assume that the decision to run off the road is also aligned with the preference of our car because the manufacturer has now decided to design the preference of the car conditional, i.e., safety of passengers should be secured under the condition that pedestrians are secured. Now suppose that our new self-driving car decides not to run off the road because it is aware of a traffic law that forbids running off the road at that location (for instance, because it increases the chance of hitting others on the sidewalk). Would the existence of this traffic law be an excuse to relieve the car from responsibility? What if the traffic laws foresee possible violations, but dictate legal consequences? For example, running off the road may incur severe sanctions, such as a very high payment by the owner or the car's manufacturer. This shows the importance of capturing norms that are in place in the environment as well as the conformance degree of the agents to such norms and also cases of norms conflict when complying with one norm may lead to the violation of another norm Broersen et al. (2001); Vasconcelos et al. (2009).

Of course, we do not aim at resolving the delicate problem of responsibility assignment once and for all. We believe that the assignment of responsibility in practice should be resolved by legal procedures and through the court. However, as automating transportation may lead to such complex scenarios, the court necessarily needs to be supported by automated responsibility reasoning tools and such tools need to capture various aspects key to the notion of responsibility. The other point we want to make here is, to investigate concepts that are relevant for analysing and assigning responsibility, such that AI systems can be designed based on such concepts. This would allow us to provide detailed analysis on the behaviour of AI systems from a responsibility perspective and to consider aspects that are relevant for the responsibility assignment problem. In other words, in order to assign responsibility to AI systems, we would like to be able to reason whether an AI system had sufficient information, acted based on its design motivation, and has been aware of possible legal, social, moral and rational norms.

4 Event-Oriented Responsibility

Following the event-oriented perspective, championed by Chockler and Halpern (2004), responsibility is defined in terms of causality, i.e., as the *extent* that a specific set of events caused an outcome. In this view, the set of events is *the* cause of an outcome only if the outcome counterfactually depends on the set of events. In particular, a specific set of events causes an outcome if and only if (1) the set of events and the outcome have taken place, (2) the outcome depends counterfactually on the set of events, and (3) the set of events is minimal in the sense that no subset of those events could be the cause. This notion is generalised to any set of events by introducing the notion of *responsibility degree*. A particular set of events is seen as responsible for a specific outcome to the degree equal to $1/(k+1)$ if the occurrence or avoidance of k number of those events could make it the cause for the outcome. In other words, the set of events is seen responsible to such a degree if k number of events need to be changed in order to make the outcome counterfactually dependent on the set of events (Chockler and Halpern, 2004).

In this view, agentive relations are implicit, as events are not explicitly linked to agents. For instance, Chockler and Halpern (2004) discusses a scenario in which Billy and Sue already thrown a stone, one after another, towards a bottle and shattered it. Then, the event "*Billy (and respectively Sue) thrown a stone*" is represented by $t_B = 1$ (and $t_S = 1$). Neither of the events is the cause, as the shattering would have happened even if one of the two events did not occur. (This is because if one fails to shatter the bottle, the other one succeeds.) Then, responsibility of the event $t_S = 1$ for the materialised shattering of the bottle is $1/2$ as one change, i.e., that $t_B = 0$ (representing that the event of *Billy thrown a stone* did not occur), could make the shattering counterfactually dependent on $t_S = 1$. Analogously, $t_B = 1$ is $1/2$ responsible for the shattered bottle.

In principle, this approach abstracts from subtleties on how events are linked to agents' autonomous actions. Note that saying that event $t_S = 1$ is to some degree responsible for the shattering of the bottle considers only Sue's actions, but ignores Sue's motivational attitude, knowledge, and normative stance based on which the action is decided. While the original modelling of Chockler and Halpern (2004) abstracts from explicitly linking events to actions and to agents, one way to apply their approach to reason about agents and their responsibility is to assume a mapping between a set of agents and events, represented by variables in their causal model. Then, each agent has full control over a (set of) variables and different possible values for a variable can be translated into the agent's repository of actions. Even in such an agentive interpretation of Chockler and Halpern (2004), i.e., if we assume agents are in control of variables, causal models are not expressive for reasoning about epistemic, motivational and normative subtleties that are crucial for assigning responsibility to agents.

For instance, to use this approach for reasoning about responsibilities in our intersection scenario (Figure 1), we need to map the event of $blue$ and red going straight, turning right, turning left, or stopping to corresponding actions available to them and then use the event-oriented model to reason about responsibilities of $blue$ and red. For instance, if we know that red and $blue$ both went towards $Alice$ and $blue$ hit her first before red reaching the scene, the event-oriented approach sees both the two vehicles $1/2$ responsible for the undesirable event of $Alice$ being hit. This way, whether $blue$ had limited knowledge in comparison to red, the preferences of the vehicles, or norms that are in place, are not playing a role in ascribing responsibility. In other words, this approach disregards the normative, motivational, and to a certain extent, epistemic aspects of the problem as it implicitly links events to agents' decisions without considering the reasons why such decisions have been made. We would like to emphasise that the notion of blameworthiness, which is defined in Chockler and Halpern (2004) as expected responsibility, is an epistemic form of responsibility. However, the introduced notion of blameworthiness takes into account the knowledge that is available to the reasoner who aims at analysing and assigning responsibility to agents, and abstracts from modelling agents' knowledge with which they do reason to decide their actions.

5 Agent-Oriented Responsibility

In multiagent scenarios where a set of AI systems interact with each other and with their shared environment (Singh, 1994), it is common that groups of agents collaboratively cooperate towards the deliberation of joint goals or in a non-cooperative fashion compete to achieve individual objectives (Dastani et al., 2004; Dastani and van der Torre, 2004). In either of the two cases, an outcome may be realised not just as an event in the environment but as a result of individual agents' actions or the outcome of collective actions. For instance, if some autonomous vehicles—e.g., red and $blue$ in the intersection scenario—crash into each other or hit a pedestrian, such an undesirable outcome is (in most cases) neither a result of any individual vehicle's desire to crash nor can be avoided unless a subset of vehicles could find means to communicate and collectively coordinate to avoid the accident. In such settings, individual agents or agent groups can be seen as being responsible for an outcome based on the set of actions available to them, their knowledge, motivation, and normative stance.

The agent-oriented perspective of Bulling and Dastani (2013) works on the idea of multiagent responsibility and coalitional abilities by assuming the capacity of agents in a coalition to communicate and form collective intentionality as a joint motivational attitude (Bratman, 2013). They ascribe responsibility to a group of agents for an outcome if and only if (1) the outcome is realised, (2) the agent group has collective actions in possession to avoid the outcome, regardless of what other agents outside the group do, and (3) that they are a minimal group with such an ability meaning that no subgroup can satisfy condition 2. The rationale is that the occurrence of the outcome was *allowed* by the group, hence they can be seen responsible as they have had the potential to avoid it. In this approach, the second condition corresponds to the avoidance potential criterion in (Braham and van Hees, 2012) while the third condition relates to the minimality condition for event-oriented responsibility (Chockler and Halpern, 2004).

In this approach, autonomy and agency of responsible agents are explicitly linked to how they manifested their power in the environment in terms of their actions. In Yazdanpanah and Dastani (2016), modelling of action-based responsibility of Bulling and Dastani (2013) is extended to multistep strategies that agents possess. They also linked this form of reasoning about already materialised chains of actions to backward-looking responsibility (van de Poel, 2011). However, as they model responsibility in the temporal logics, their notion of responsibility is also applicable for forward-looking forms of responsibility.[3] For instance, their notion of forward-looking responsibility can be applied for planning and task coordination as "*Alec and Bob can be responsible for providing masks to the hospital if they can avoid any shortages*" and for ensuring the ethical behaviour of AI systems as "*this AI system is responsible for handling HR decisions in our institute if it can avoid all the biased decisions*".

To incorporate epistemic aspects of actual responsibility, Yazdanpanah et al. (2019) and Naumov and Tao (2020) integrate agent's lack of information into multiagent responsibility models. While they both use indistinguishability relations, their approaches to the problem are different. In Yazdanpanah et al. (2019), the existence of strategies under imperfect information (towards avoiding a potential outcome) is taken as a condition for ascribing responsibility to agent groups. And to that end, they use an epistemic notion of strategy as a chain of actions that ensures a state of affairs even under agents' lack of information. On the other hand, Naumov and Tao (2020) model agents' epistemic states to reason about the knowledge of agents about their strategies (towards an outcome). They argue that knowingly causing an outcome is a base to take their collective action as an intentional one and ascribing blameworthiness to them.

Capturing such epistemic aspects allows distinguishing those who caused harm from those who caused it knowingly by considering their knowledge about the environment, their abilities, and the consequences of their actions. For instance, in the intersection scenario, applying the responsibility model of Yazdanpanah et al. (2019) allows reasoning about what agents could do to avoid crashing into $Alice$ given the knowledge they had at each point in time. As discussed earlier and depicted in Figure 2, the two vehicles have different knowledge about the presence of $Alice$. As $blue$'s visibility was blocked by the building, it had no understanding of the disastrous consequences of turning right in the intersection. However, red had a clear view on the intersection and could inform $blue$. Such a communication action could update $blue$'s knowledge and, according to Yazdanpanah et al. (2019), key to determining if, and to what extent, $blue$ is responsible for the harm it caused. As presented in Figure 2, the same intersection case with the identical history of events (under which the two vehicles proceed with their most preferred option) leads to epistemically distinguishable scenarios if we consider only one communication action, namely that red informs $blue$ (about $Alice$). Only under scenario 1 and 2 in Figure 2, in which both vehicles knowingly went towards $Alice$, they are both $1/2$ responsible for the caused harm. Here, they could decide to go for alternative options they had

[3] As discussed by van de Poel (2011), forward-looking responsibility refers to responsibility for eventualities while backward-looking responsibility is concerned with already materialised state of affairs.

and avoid crashing into *Alice*. However, without communication, when *blue* was ignorant about *Alice*, it is not reasonable to assign responsibility to it as there is no strategy in its possession to avoid the harm. In such cases, i.e., when no individual could have avoided a state of affairs (but some agent groups could do so collectively), Bulling and Dastani (2013) and Yazdanpanah et al. (2019) allow ascribing responsibility to agent groups. We later (in Section 6) discuss how this form of collective responsibility relates to the so-called responsibility gaps (Braham and VanHees, 2011) and the problem of many hands in moral philosophy (van de Poel et al., 2015). Finally, in scenario 4 of Figure 2 the solely responsible agent is *red* as it had means to avoid the crash. Note that Figure 2 provides a pruned decision tree[4] just to show how communicative actions and epistemic aspects play a role in the process of agent-oriented responsibility. One can see that if blue knew about *Alice* before turning right, it could avoid the crash by choosing its less preferred path and turning left, or by stopping for a while and seeing if *Alice* passes the intersection and clears the way.

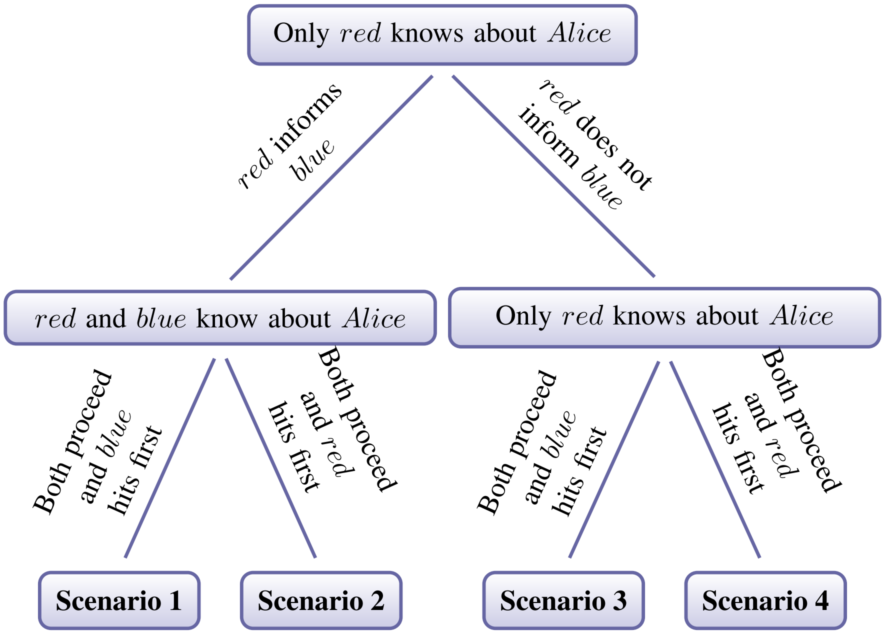
Only red knows about Alice
red informs blue
red does not inform blue
red and blue know about Alice
Only red knows about Alice
Both proceed and blue hits first
Both proceed and red hits first
Both proceed and blue hits first
Both proceed and red hits first
Scenario 1
Scenario 2
Scenario 3
Scenario 4

Figure 2: Four Scenarios: Epistemic subtleties in the intersection scenario given that both vehicles proceeded with their most preferred option (i.e., *red* went straightforward and *blue* turned right).

We highlight that established norms in the environment, in this case the traffic law, could make it explicit that after seeing a pedestrian, vehicles are obliged

[4] For instance, we did not go through all actions available to vehicles, e.g., to run into roadblocks or to go off the road, and also disregarded other means of communication, e.g., that vehicles could seek information from sensors on the building.

to stop unless the road is clear, and that this safety-oriented norm has priority over performance, e.g., that vehicles are supposed to reach their destination as fast as possible. So, although the presented approaches take into account the epistemic aspect, further work is required to capture and integrate the normative and motivational stance of agents. To capture the normative aspect, we envisage the integration of such methods into norm-aware multiagent reasoning techniques Alechina et al. (2014); Dastani et al. (2017). Such an integration allows reasoning about responsibility in presence of norms and under different norm enforcement schemes. For instance, in the intersection scenario, we can determine $blue$'s responsibility under normatively distinguishable scenarios. Its responsibility will differ if, according to traffic rules, it was prohibited to go out of the road in that specific part of the city, allowed to do so, or obliged to do so if the vehicle foresees crashing into $Alice$. Another way forward is to integrate agents' motivational stance and capturing the intent of agents. To that end, we ideate using agents' list of available options and ranking them with respect to their desirability for the agent (to capture the preference of agents over actions). This allows distinguishing a materialised action from an intentional one, following the idea that intention can be defined and reasoned about in terms of desires (and actions to fulfil them) and agents' commitment to deliberate an action or a chain of actions over time (Cohen and Levesque, 1990).

6 The Problem of Many Hands in AI Systems

Ascribing responsibility to collectives in multiagent settings raises the question on how to link the collective-level responsibility to individuals within the collective. After all, what came out of the collective was a result of the aggregation of individual actions and decisions. This problem is known in political and moral philosophy (Thompson, 1980; Braham and VanHees, 2011) as *the problem of many hands*. When many hands (i.e., multiple agents) have been involved to materialise an outcome, the identification of causal or strategic contribution is not straightforward.[5]

As AI tools are gradually embedding in society, their decisions become ethically and normatively loaded because of their potential to lead to (un)desirable consequences. Hence, it is key to respond to the problem of many hands in AI systems in a systematic and verifiable fashion. A key to ensuring the trustworthiness of such AI systems is to determine how and to what extent every individual agent who contributed to a collective decision-making process is responsible for the outcome. Braham and VanHees (2011) highlight that a method for addressing the problem of many hands, and ascribing responsibility to individuals,

[5] Note that the problem of many hands in moral philosophy (van de Poel et al., 2015) is not concerned with coordinating group actions and collective planing towards particular outcomes. The main focus of this problem, in the sense understood in the philosophical literature, is on ascribing responsibility to group members given an already materialised outcome. See van de Poel (2011) and van de Poel et al. (2015) for a detailed analysis on the relations and distinctions between the (forward-looking) planning-oriented approach and the (backward-looking) problem of many hands.

needs to capture the three main dimensions of the notion: (1) capturing how agents within a collective *causally* contributed to the outcome (e.g., via executing or failing to execute individual actions), (2) taking into account norms that agents adhered to or violated, the *normative* nature of the outcome in question, and established rules in the environment (e.g., whether in a given context, an action or an outcome resulted from collective actions is known to be undesirable), and (3) considering how *epistemic* aspects affects the collective ability to cause or avoid an outcome, and in turn their partial responsibility for it (e.g., whether agents are fully knowledgeable about their own abilities, know about the presence of others in the environment, or are capable of communicating with one another to share knowledge).

To address the problem of many hands in modelling responsibility of AI systems, Yazdanpanah and Dastani (2015) propose quantitative degrees of responsibility. They focus on settings in which decisions are weighted or mainly depend on the size of groups, e.g., partisan voting in parliaments. In such settings, power of groups is quantified as a measure for ascribing responsibility to individuals, e.g., based on the number of votes they possess or can control to ensure or avoid a collective decision. The idea to ascribe quantified degrees of responsibility to individuals (within a responsible group) is further explored in Friedenberg and Halpern (2019) and Yazdanpanah et al. (2019). They both apply fair division notions from microeconomics but in different modelling settings. In Friedenberg and Halpern (2019), causal models of Chockler and Halpern (2004) are the base while Yazdanpanah et al. (2019) has a logic-based setting rotted in Bulling and Dastani (2013).

7 Concluding Remarks

As AI systems are increasingly embedded into our society, we argue that as a means to support ethical and human-centred AI, it is necessary to develop precise models for reasoning about responsibility of AI systems. Such formal responsibility models can be embedded into the reasoning engines of the AI systems. This way, AI systems can reason about how (in a mixed society of humans and artificial systems) one should behave in view of her potential responsibilities for consequences. Then, given a set of social values, ethical concerns, and technical reliability conditions (i.e., sociotechnical requirements that are expected to be fulfilled), AI systems can call their responsibility reasoning component to find out if they will be accounted for certain outcomes and held responsible for being in compliance or in violation of some values and concerns.

Enriching AI systems with responsibility reasoning models becomes gradually possible as AI systems, in particular autonomous systems, are increasingly often designed and developed based on high-level concepts such as knowledge, sense data, preferences, and norms (including rational, economic rational, legal, social, or moral norms). The behaviour of such AI systems can therefore be analysed and explained as being based on, or caused by, reasoning with specific knowledge, sensed data, preferences, and specific norms. The assignment of responsibility and related concepts such as blameworthiness to AI systems can then be analysed and motivated by such concepts based on which the decisions of AI agents have been taken. Such analyses would allow us to verify whether responsibility can be assigned to one or a group of AI systems, or to verify the conditions under which AI systems can be excused from responsibility.

As discussed, models that allow reasoning about actual responsibility are key to autonomous AI systems and can be used to ensure reliable and trustworthy embedding of AI systems in society. In addition, as future directions, we envisage the applicability of actual responsibility models in related domains such as AI planning by giving a task to responsible groups able to deliver the task. Such responsibility models can also be used to ensure the legality of AI systems by ascribing liability to agents who could avoid harm in view of epistemic, normative, and motivational considerations. Finally, models of actual responsibility can be used to design and develop adaptive AI systems, e.g., by using the degree of blameworthiness as a base for regret/award values in reinforcement learning.

Data access statement

Data sharing is not applicable to this article because no new data were created or analysed in this study.

Acknowledgements

We thank the anonymous reviewers for their incisive comments that were most useful in revising this paper. This work is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) through the “Trustworthy Autonomous Systems Hub” (EP/V00784X/1). For the purpose of open access, the author has applied a creative commons attribution (CC BY) licence to any author accepted manuscript version arising.

References

Alechina, N., Dastani, M., and Logan, B. (2014). Norm approximation for imperfect monitors. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 117–124.

Benjamins, R. (2021). A choices framework for the responsible use of ai. *AI and Ethics*, 1(1):49–53.

Braham, M. and van Hees, M. (2012). An anatomy of moral responsibility. *Mind*, 121(483):601–634.

Braham, M. and VanHees, M. (2011). Responsibility voids. *The Philosophical Quarterly*, 61(242):6–15.

Bratman, M. E. (2013). *Shared agency: A planning theory of acting together*. Oxford University Press.

Broersen, J. M., Dastani, M., Hulstijn, J., Huang, Z., and van der Torre, L. W. N. (2001). The BOID architecture: conflicts between beliefs, obligations, intentions and desires. In André, E., Sen, S., Frasson, C., and Müller, J. P., editors, *Proceedings of the Fifth International Conference on Autonomous Agents*, pages 9–16. ACM.

Bulling, N. and Dastani, M. (2013). Coalitional responsibility in strategic settings. In *Proceedings of the International Workshop on Computational Logic in Multi-Agent Systems*, pages 172–189.

Chockler, H. and Halpern, J. Y. (2004). Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research*, 22:93–115.

Chopra, A. K. and Singh, M. P. (2021). Accountability as a foundation for requirements in sociotechnical systems. *IEEE Internet Computing*, 25(6):33–41.

Cohen, P. R. and Levesque, H. J. (1990). Intention is choice with commitment. *Artificial intelligence*, 42(2-3):213–261.

Cugurullo, F. (2021). Urban artificial intelligence: From automation to autonomy in the smart city. *Frontiers in Sustainable Cities*, 2(38).

Dastani, M., Dignum, F., and Meyer, J. C. (2003). Autonomy and agent deliberation. In Nickles, M., Rovatsos, M., and Weiß, G., editors, *First International Workshop on Computational Autonomy - Potential, Risks, Solutions*, volume 2969 of *Lecture Notes in Computer Science*, pages 114–127. Springer.

Dastani, M., Dignum, F., and Meyer, J.-J. (2004). Autonomy and agent deliberation. In *Agents and Computational Autonomy*.

Dastani, M., Sardina, S., and Yazdanpanah, V. (2017). Norm enforcement as supervisory control. In *International Conference on Principles and Practice of Multi-Agent Systems*, pages 330–348.

Dastani, M. and van der Torre, L. W. N. (2004). Programming boid-plan agents: Deliberating about conflicts among defeasible mental attitudes and plans. In *3rd International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 706–713. IEEE Computer Society.

Dignum, V. (2019). *Responsible artificial intelligence: how to develop and use AI in a responsible way*. Springer Nature.

Friedenberg, M. and Halpern, J. Y. (2019). Blameworthiness in multi-agent settings. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, pages 525–532.

Halpern, J. Y. (2016). *Actual Causality*. MIT Press.

Houlgate, L. D. (1968). Knowledge and responsibility. *American Philosophical Quarterly*, 5(2):109–116.

Macrorie, R., Marvin, S., and While, A. (2020). Robotics and automation in the city: a research agenda. *Urban Geography*, 42(2).

Naumov, P. and Tao, J. (2020). An epistemic logic of blameworthiness. *Artif. Intell.*, 283:103269.

O'neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown Books.

Pearl, J. (2009). *Causality*. Cambridge university press.

Pearl, J. and Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic books.

Petersen, S. (2013). Utilitarian epistemology. *Synthese*, 190(6):1173–1184.

Ramchurn, S. D., Stein, S., and Jennings, N. R. (2021). Trustworthy human-ai partnerships. *Iscience*, 24(8):102891.

Safransky, S. (2020). Geographies of algorithmic violence: Redlining the smart city. *International Journal of Urban and Regional Research*, 44(2):200–2018.

Singh, M. P. (1994). *Multiagent Systems - A Theoretical Framework for Intentions, Know-How, and Communications*, volume 799 of *Lecture Notes in Computer Science*. Springer.

Smith, H. (2020). Clinical ai: opacity, accountability, responsibility and liability. *AI & SOCIETY*, pages 1–11.

Stilgoe, J. (2018). Machine learning, social learning and the governance of self-driving cars. *Social studies of science*, 48(1):25–56.

Stilgoe, J. (2020). *Who's Driving Innovation. New Technologies and the Collaborative State*. Palgrave Macmillan.

Thompson, D. F. (1980). Moral responsibility of public officials: The problem of many hands. *The American Political Science Review*, pages 905–916.

van de Poel, I. (2011). The relation between forward-looking and backward-looking responsibility. In *Moral responsibility*, pages 37–52. Springer.

van de Poel, I., Royakkers, L. M., Zwart, S. D., and De Lima, T. (2015). *Moral responsibility and the problem of many hands*. Routledge New York.

Vargas, M. (2013). *Building better beings: A theory of moral responsibility*. Oxford University Press.

Vasconcelos, W. W., Kollingbaum, M. J., and Norman, T. J. (2009). Normative conflict resolution in multi-agent systems. *Autonomous Agents and Multi-Agent Systems*, 19(2):124–152.

Yazdanpanah, V. and Dastani, M. (2015). Quantified degrees of group responsibility. In *Proceedings of the International Workshop on Coordination, Organizations, Institutions, and Normes in Agent Systems*, pages 418–436.

Yazdanpanah, V. and Dastani, M. (2016). Distant group responsibility in multi-agent systems. In *Proceedings of the 19th International Conference on Principles and Practice of Multi-Agent Systems*, pages 261–278.

Yazdanpanah, V., Dastani, M., Jamroga, W., Alechina, N., and Logan, B. (2019). Strategic responsibility under imperfect information. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 592–600.