Reasoning about responsibility in autonomous systems challenges and opportunities

Yazdanpanah, Vahid; Gerding, Enrico H.; Stein, Sebastian; Dastani, Mehdi; Jonker, Catholijn M.; Norman, Timothy J.; Ramchurn, Sarvapali D.

DOI
10.1007/s00146-022-01607-8

Publication date
2022

Document Version
Final published version

Published in
AI and Society

AI & SOCIETY
https://doi.org/10.1007/s00146-022-01607-8

OPEN FORUM

Reasoning about responsibility in autonomous systems: challenges and opportunities

Vahid Yazdanpanah[1] · Enrico H. Gerding[1] · Sebastian Stein[1] · Mehdi Dastani[2] · Catholijn M. Jonker[3] · Timothy J. Norman[1] · Sarvapali D. Ramchurn[1]

Received: 14 June 2021 / Accepted: 22 November 2022

Abstract
Ensuring the trustworthiness of autonomous systems and artificial intelligence is an important interdisciplinary endeavour. In this position paper, we argue that this endeavour will benefit from technical advancements in capturing various forms of responsibility, and we present a comprehensive research agenda to achieve this. In particular, we argue that ensuring the reliability of autonomous system can take advantage of technical approaches for quantifying degrees of responsibility and for coordinating tasks based on that. Moreover, we deem that, in certifying the legality of an AI system, formal and computationally implementable notions of *responsibility*, *blame*, *accountability*, and *liability* are applicable for addressing potential responsibility gaps (i.e. situations in which a group is responsible, but individuals' responsibility may be unclear). This is a call to enable AI systems themselves, as well as those involved in the design, monitoring, and governance of AI systems, to represent and reason about who can be seen as responsible in prospect (e.g. for completing a task in future) and who can be seen as responsible retrospectively (e.g. for a failure that has already occurred). To that end, in this work, we show that across all stages of the design, development, and deployment of trustworthy autonomous systems (TAS), responsibility reasoning should play a key role. This position paper is the first step towards establishing a road map and research agenda on how the notion of responsibility can provide novel solution concepts for ensuring the *reliability* and *legality* of TAS and, as a result, enables an effective embedding of AI technologies into society.

Keywords Trustworthy autonomous systems · Human–agent collectives · Multi-agent responsibility reasoning · Human-centred AI

1 Introduction

To develop and effectively deploy trustworthy autonomous systems (TAS), regulatory bodies are clear in suggesting that the behaviour of such systems should be monitored and controlled, such that potential violations of legal norms and societal values are avoided (Office for Artificial Intelligence 2020; European Commission: The High-Level Expert Group on AI 2019). Confirming the need to focus on human-centred artificial intelligence (AI), the academic community argues that it is crucial to coordinate the behaviour of AI systems (Bryson and Winfield 2017; Rahwan et al. 2019), to ensure their compatibility with our social values (Russell 2019; Ramchurn et al. 2021), and to design verifiably safe and reliable human–agent collectives (Jennings et al. 2014; Abeywickrama et al. 2019). Despite such a general agreement, we still face sociotechnical challenges for embedding autonomous systems in society and, in particular, for ensuring their reliability and legality. This work highlights these challenges, motivates their importance for safe and *trustworthy* practice of AI, and argues that addressing challenges in ensuring the reliability and legality of autonomous systems benefits from formal responsibility reasoning methods.[1]

The need for ensuring trustworthiness of autonomous systems is known and well-argued in the literature (Murukannaiah et al. 2020; Dignum 2019). However, as long as

✉ Vahid Yazdanpanah
v.yazdanpanah@soton.ac.uk

1 University of Southampton, Southampton, UK

2 Utrecht University, Utrecht, Netherlands

3 Delft University of Technology, Delft, Netherlands

[1] See Yazdanpanah et al. (2021a) for a conceptual analysis of different forms of responsibility and their sociotechnical characteristics.

Published online: 02 December 2022

we remain at an abstract level and merely discuss how TAS ought to behave (i.e. without clear instructions on potential ways to ensure trustworthiness), the gap will not be bridged and challenges for the practice of AI, and its embedding in society, will remain unsolved. Following Ramchurn et al. (2021), we argue that, to ensure trustworthiness in the design and development of TAS, we require novel operational tools to represent and reason about *reliability* and *legality* as two facets of trustworthiness in autonomous systems.[2] In contrast to purely *technical* engineering-oriented perspectives on reliability as coherence of the system behaviour with its design goals, e.g. by Birolini (2013) and O'Connor and Kleyner (2012), we understand reliability of autonomous systems in relation to society as their context of application and, in turn, as a sociotechnical notion. This calls for methods that are, on one hand, expressive enough to capture the sociotechnical nature of TAS and, on the other hand, computationally implementable. To address this gap, we argue that responsibility reasoning enables addressing open problems in assuring the reliability and legality of autonomous systems.[3] The notion of responsibility and the use of formal methods to represent and reason about responsibility can play a key role, as they connect the social requirements and technical capacities of TAS. Responsibility models can act as a formal apparatus to model and reason about legality of behaviours and ethical consequences (social requirement of TAS) (Constantinescu et al. 2021; Yeung 2018), and as computational tools for reliable coordination of tasks in multi-agent settings (technical capacities of TAS) (Dignum 2019; Ramchurn et al. 2021; Yazdanpanah et al. 2020)

The rest of the paper is structured as follows: we first elaborate on conceptual connections between responsibility and autonomy, and discuss how responsibility reasoning supports the design and development of trustworthy autonomous systems (TAS). Then, in Sects. 3 and 4, we discuss how technical advancements in computational responsibility can address open challenges in ensuring the *reliability* and *legality* of autonomous systems. To that end, for each challenge, we discuss its importance, show the gap in existing work, and motivate the relevance and potential of responsibility reasoning methods for bridging the gap. Building on this foundation, Sect. 5 focuses on the development of concrete research themes in an underdeveloped research area on *responsibility research in autonomous systems*. In Sects. 6 and 7, we conclude this position paper by showing how the proposed line of research relates to neighbouring domains and summarising its potentials for a safe and trustworthy embedding of autonomous systems in society.

2 Conceptual analysis: responsibility in and for TAS

There are various links between the notion of responsibility and the concept of autonomy (in autonomous systems). In principle, the relation between autonomy and responsibility (for a particular outcome) is as follows.[4] Responsibility necessitates autonomy, as this is defined only for an agent with a level of autonomy (Braham and van Hees 2012; Champlin 1994). From the other side, autonomy is about the capacity of an entity to manifest its agency via performing actions, either communicative or physical (Searle 1989, 1995), and thereby causing change in the environment to maximise its utility or reach its goals in an environment (Rao and Wooldridge 1999; Bratman 2007; Georgeff et al. 1998; Dastani et al. 2003). Then focussing on the causal account of the notion of responsibility, agent A causing change and reaching outcome O in the environment indicates *"A's responsibility for O"*. The concept is also related to blameworthiness, but this is distinct from responsibility. In particular, *"A is blameworthy for O"* if A acted knowingly (Chockler and Halpern 2004). For instance, using the example by Chockler and Halpern (2004), an underaged person can be responsible for killing her dad because of playing with a pistol but not necessarily blameworthy as she might not be knowledgeable about the harm that a pistol can cause.

Complementary to this, in multi-agent settings, the line of research on strategic responsibility and action-state semantics (Bulling and Dastani 2013; Yazdanpanah and Dastani 2016) focuses on the strategic capacities of agents or groups of agents with respect to eventualities in prospect. Here, the

[2] We follow Ramchurn et al. (2021); Jennings et al. (2014) and see trustworthiness in autonomous systems as a multifaceted concept. This work focusses on reliability and legality as two facets of trustworthiness in autonomous system and discusses how responsibility reasoning can contribute to these two aspects in Sects. 3 and 4, respectively.

[3] As presented in Ramchurn et al. (2021), establishing trustworthy autonomous systems requires focussing on other aspects such as verifiability, security, safety and functionality but, in this work, we mainly focus on how responsibility reasoning methods can support *reliability* and *legality* of TAS as they directly relate to two forms of forward- and backward-looking responsibility, respectively (van de Poel, 2011).

[4] Among various form of responsibility—first distinguished by Hart (1968) and further discussed by van de Poel (2011)—we are neither focussing on a particular form nor introducing a novel notion of responsibility. Our aim is to show how different notions of responsibility can support addressing challenges in the design and development of trustworthy autonomous systems. This is a call to the AI community to explore the capacities of the rich literature on responsibility reasoning for addressing issues that are key to the widespread and socially acceptable deployment of autonomous systems.

agents' responsibility is formulated in terms of pre-conditions as an ex ante notion.

On the other hand, Santoni de Sio and van den Hoven (2018) argue that, ultimately, it should be humans that are to remain in control of, and thus responsible for, relevant decisions. However, it is important to realise that humans are not always in a position to reason about, and understand, what part in a system they are expected to 'take over control of' and at which appropriate moment. This is where responsibility reasoning research is necessary to decide who, and to what extent, is responsible for the long-term behaviour of, or a specific decision made by, an AI system. By enabling AI-based autonomous systems to reason about potential responsibility-related issues (in prospect), they can minimise harmful consequences and ensure ethical and trustworthy behaviours. Such an expectation is unlikely from agents that are purely focussed on maximising efficiency-oriented indicators and that ignore different types of responsibility (from being accountable to legally liable) in their automated decision-making process.

Against this background, this position paper focuses on elaborating how different dimensions and notions of responsibility, as a sociotechnical concept (Yazdanpanah et al. 2021a), relate to, and can address, interdisciplinary challenges in the design, development and deployment of trustworthy autonomous systems.[5] Our aim is not to develop a comprehensive philosophical account of *responsibility challenges in autonomous systems* but to focus on mapping some known challenges to sociotechnical approaches and to provide a research agenda on how sociotechnical notions of responsibility, responsibility quantification and the computational account of blameworthiness, accountability and liability contribute to addressing these challenges. To that end, and due to the interdisciplinary nature of the open challenges presented here, we avoid going into the philosophical details and controversies. Neither do we delve into the specific technical requirements behind computational solution concepts. Instead, we redirect interested readers to the relevant literature throughout the discussion. Indeed, our main aim is to establish the research agenda on responsibility research for TAS by articulating the challenges to which responsibility reasoning methods have the opportunity to contribute.

To do so, the paper (1) identifies challenges in TAS reliability and legality; (2) presents the type of responsibility models, theoretical frameworks and hypotheses that could lead to significant advances (theoretical, operational, etc.) and steps towards solutions for identified challenges; (3) explains related questions about responsibility in the field of AI; and (4) discusses how this research agenda relates to other neighbouring scientific domains.

3 Responsibility research for reliability of TAS

In principle, the *forward-looking* perspective on the notion of responsibility—in contrast to the *backward-looking* view (van de Poel 2011)—is focussed on eventualities as potential situations that may be materialised in future and analyses how individual agents or agent collectives can or ought to affect such state of affairs in future (van de Poel 2011). For instance, consider how we might use responsibility to determine roles while planning a picnic. We say Alice is responsible for transportation and Bob is for preparing food. Hart (1968) refers to this form as task/role responsibility. Such a notion of responsibility is also applicable for ensuring the reliability of TAS. To that end, ascription of responsibilities needs to take into account the abilities of agents involved and their potential to complete tasks we allocate to them. If Alice is the name of an autonomous vehicle that is going to take care of transportation, we need to make sure that it is capable of completing the task in view of circumstances in the environment and in the presence of other agents. Roughly speaking, for a reliable TAS, we require effective responsibility ascription methods that are able to reason about the requirements and potentials of human and artificial agents as well as barriers in their environment. In this section, we elaborate on TAS challenges that call for novel responsibility reasoning research and discuss desirable requirements to be met.

3.1 Responsibility degrees as a base for resilience reasoning

Moving from AI systems in the lab towards real-life autonomous systems, e.g. in transportation and healthcare, reliability of the systems and their ability to handle potential failures are key for social acceptance. Society will not accept the integration of autonomous vehicles unless they show the capacity to perform reliably and in a fault-tolerant manner—e.g. see the EU Commission's proposal to establish harmonised regulations on artificial intelligence, the "AI Act" (European Commission 2021; European Parliament, 2021). Although system designers and manufactures ought to aim for optimal performance, they should also take into account how their systems handle failures. Are we putting in place resilience-ensuring mechanisms, e.g. by integrating some backup measures, such that a failure in one part of the system does not lead to a significant damage in the performance of the whole system?

[5] The initial ideas developed in this article were presented at the International Conference on Autonomous Agents and Multiagent Systems, AAMAS'21 (Yazdanpanah et al., 2021b).

One should never expect that all the components in an autonomous system behave as expected, and so one has to put in place overarching methods to ensure reliability and resilience. For this, we can rely on formally verifiable responsibility reasoning methods (Chockler and Halpern 2004; Yazdanpanah and Dastani 2016; Naumov and Tao 2020). Following Chockler and Halpern (2004), we deem that the notion of responsibility can be a base for conceptualising resilience and agree with Vardi's call on the need for methods capable of analysing the tradeoff between efficiency and resilience in sociotechnical systems and for developing comprehensive models of resilient human–AI partnerships (Vardi 2020; Ramchurn et al. 2021).

In brief, resilience of a system increases if the agents' degree of responsibility (for their task) is *partial* and no individual agent has full responsibility meaning that they share responsibility to complete a task with some other agents. For instance, imagine a three-member multi-agent software system in which only agent *A* has the full responsibility with respect to updating a block/value (task responsibility). It means that, if *A* fails, no one is able to correct the problem. If the system was designed such that at least two (coordinated) agents were responsible for updating the block/value, we would have some level of inefficiency, but responsibilities would be distributed. In such a coordinated system, one can guarantee a certain level of resilience against potential failures. We propose further investigation on how different formalisations of the notion of responsibility—e.g. the causal notion of Chockler and Halpern (2004) or the strategic notion of Bulling and Dastani (2013)—can be of use in different domains to ensure the resilience of autonomous systems.[6] Then the main idea is to use *responsibility degrees* as a measure of resilience in autonomous systems[7]. And enabling resilience, in turn, promotes fault tolerance and reliability of autonomous systems and their trustworthiness from the users' point of view.

In some contexts, we face a more complex situation where control is shared: two or more agents may simultaneously be exerting control of the system. This is what Flemisch et al. (2016) call *shared control* in human–machine teams. One should realise that, in future, it is likely that various active agents in a system may be designed/owned by different teams, for different objectives. This form of heterogeneity makes it more challenging to create a functional but resilient design. We formulate this challenge as *the need for practical and provably sound degrees of responsibility to ensure system reliability and fault tolerance*.

3.2 Accountability reasoning for task coordination

As discussed earlier, ensuring reliability and resilience, especially in heterogeneous teams, requires some form of coordination. Accountability reasoning, as a task-related form of responsibility (Yazdanpanah et al. 2021a) for failing to deliver an allocated task, can be applied to addressing the task coordination challenge in TAS. The challenge and open problem are due to *the need for operational accountability ascription and task coordination methods in the organisational context of TAS*. Here, the main idea is to allocate tasks to agents that are able to deliver them and are, in addition, accountable for doing so. This ensures a more reliable task coordination process in autonomous systems and promotes users' trust in how a TAS handle complex tasks.

Acting in a coordinated manner will be particularly challenging in human–agent teaming. In human–agent collectives, where human and artificial agents collaborate towards ensuring goals, it is crucial to put in place mechanisms for balancing the two decision-making types in what Jennings et al. call *flexible autonomy* (Jennings et al., 2014). In essence, flexibly autonomous systems allow "agents to sometimes take actions in a completely autonomous way without reference to humans [type 1], while at other times being guided by much closer human involvement [type 2]". In specific cases, e.g. in the aviation industry, where we have mature autopilot systems, the resilience of the human–agent system may be improved by allowing (some) agents to take over control from humans in the loop. In particular, we can argue that, in cases where an agent is more knowledgeable (i.e. has a higher level of observability), it is reasonable to allow the agent to lead the operation. Then the main problem is to understand.

This issue also relates to the notion of "*interdependency*" in co-active design (Johnson et al. 2014). In principle, Johnson et al. (2014) argue that in collaborative AI systems where humans and artificial agents form hybrid intelligence and act as a team, activities of participating actors are interdependent. Here, interdependence refers to the set of relationships used to manage dependencies. By engaging in such relationships, the context of the activity now encompasses

[6] Discussing technical differences of these formalisations is beyond the scope of this position paper but, in brief, Chockler and Halpern (2004) model responsibility for an already materialised outcome using causal networks and how a set of events contributed to causing the outcome in question. The formalisation that Bulling and Dastani (2013) put forward is rooted in strategic agent-oriented logics (distinguishable from event-oriented perspective of Chockler and Halpern (2004)) and models responsibility in terms of the agents' capacity to avoid an outcome. Both of these approaches allow reasoning about responsibility of agents for materialised outcomes (retrospectively) while the latter is also expressive for modelling responsibility of agents for possible outcomes (prospectively). See Dastani and Yazdanpanah (2022) for a detailed analysis of different computational perspectives on responsibility modelling in AI systems

[7] See Yazdanpanah and Dastani (2015) for methods to quantify group responsibility in multiagent settings. who is, and to what extent they are, accountable for the outcome of such decisions that are made under flexible autonomy.

all parties involved as a single joint system and these relationships then define what is pertinent for common ground. Such forms of interdependency complicate assigning tasks and then accountability to individual actors. As a remedy, the *co-active design method* is proposed to enhance a reliable way of designing systems for such collaborations with features that enable (1) additional monitoring (to enhance mutual observability) functionalities, (2) agents taking over tasks from other team members (to improve resilience), (3) team members informing and directing other actors (to support mutual directability) based on insights in upcoming complications and (4) actors knowing how the collaborating actors work (to establish mutual predictability). Using this approach, it will be clear who is accountable for what task, at what stage of delivery, and give assurances to the end users that: "*in a trustworthy autonomous system, if a failure occurs, accountability will not be voided and who to account for it can be determined at every point of operation*".

Another suggested way forward is to employ multi-agent organisation (MAO) models (Ferber et al. 2003; Horling and Lesser 2004; Santoni de Sio and van den Hoven 2018; van der Waa et al. 2020) and develop accountability ascription methods for human–agent autonomous systems. Such methods are expected to be expressive to reason about task coordination, delegation, and shared control in TAS (Norman and Reed 2000; Flemisch et al. 2016; Yazdanpanah et al. 2020).

3.3 Responsibility research for legality of TAS

In the generic sense of the term, responsibility is commonly understood as a backward looking notion with focus on reasoning about who should account for or be seen blameworthy for a situation. For instance, imagine a multi-agent system with three autonomous vehicles, two pedestrians, and one human-driven vehicle. After the occurrence of a crash in which some of these agents are involved, backward-looking responsibility reasoning is concerned with individuals or groups of agents who caused the crash, knew about ways to avoid it, or intentionally orchestrated the situation (e.g. by disrupting the communication system of the vehicles). While each of these groups are somehow responsible, they are responsible in different ways. An agent may be responsible as its actions caused a situation but not blameworthy if it did the act with no knowledge of the consequence (Hart 1968). The other form of responsibility is responsibility as liability. In essence, one is liable for a consequence if they found to be responsible for violating a regulative norm, e.g. for going over the speed limit where an established norm regulates the speed limit and determines a sanction for violating the norm. Note that, in this work, we abstract from contextual differences in various legal systems and how liability reasoning differs in the criminal and tort law, e.g. with respect to intentionality.

Here, backward-looking responsibility reasoning methods can be used as decision support tools for automated liability determination in TAS. To be specific, we focus on their applicability for (1) addressing the so-called responsibility voids (Santoni de Sio and Mecacci 2021), defined as situations in which a collective is known to be responsible, but determining individuals' degree of responsibility is not straightforward, and (2) developing sanctioning mechanisms that are applicable as a means to ensure the reliability of new forms of artificial autonomy.

3.4 Quantified degrees to address responsibility voids

Responsibility voids are well-studied situations in moral philosophy (Braham and van Hees 2011). If we allow agent groups to take an intentional stance, e.g. following Bratman (1993), then we face situations where a group is found to be responsible. How to distribute this responsibility and attribute it to individuals in the group (partially) can be a challenge in cases where the causal links between agents' actions and the outcome are unclear, or when the distribution of knowledge among the agents is not fully known to the reasoner. For instance, imagine a scenario, adapted from McLaughlin (1925), where a traveller's water canteen is poisoned by one enemy and then emptied by another one. We refer to both fellow travellers as enemies to clarify that their actions were intentionally aimed at harming the traveller in question. The traveller dies of thirst in the middle of the desert. For a judge who is reasoning about the case, it is clear that the two enemies are responsible as a collective but the extent and the degree of responsibility of each is not clear. In this case, the traveller would die even if one enemy avoided doing his respective action. Considering counterfactual dependence as a necessity for building causal relation, and, in turn, considering causal relation as a necessity for seeing one responsible for an outcome, neither of the enemies is responsible for the death. This is a stranded case of the so-called *responsibility void* (Braham and van Hees 2011), where linking collective to individual responsibility is a challenge.

Handling responsibility voids is even more challenging in mixed human–agent collectives (Jennings et al. 2014) with *flexible autonomy* in place. These are teams in which artificial agents sometimes make decisions with complete autonomy and sometimes operate under more control from humans. For instance, imagine a healthcare scenario where human surgeons are performing an operation in collaboration with semiautonomous robots. For some tasks, as a part of the complex task of performing the whole surgery, robots

act with full autonomy, while for other parts of the process, humans are in control. Then, if the operation results in a failure, who is, and to what extent are they responsible for it? Does the answer differ from cases where a single surgeon handles the whole procedure? We formulate this challenge as *the need for effective tools to distribute collective-level responsibilities into quantitative individual-level degrees of responsibility.*

This motivates developing tools for ascribing responsibility to interconnected human–agent teams and then to assign degrees of responsibility to team members, considering that they may have acted in an asynchronous and uncoordinated manner. As the performance of such teams may involve convoluted processes, an act may be safe at a point in time but eventually forces the system to face an inevitable failure in a later stage. This necessitates developing tools to monitor and manage the history of events and keeping track of responsibility trails at each stage, e.g. using provenance tracking methods (Ramchurn et al. 2016).

As we are faced with dynamic degrees of autonomy in TAS, we require contextualised methods that are able to ascribe responsibility dynamically. A way forward is to capture resource and cost dynamics (Alechina and Logan 2020). Using such cost-aware methods, one can formulate degrees of responsibility based on agents' control over the resources. In other words, responsibilities differ as the abilities of agents (to cause or avoid harm) differs with respect to the control they had over different resources in different time periods.

3.5 Liability reasoning in view of new forms of autonomy

Over centuries, we established various measures to avoid, or nudge people to reduce, the violation of regulative norms and social values in societies. This includes mechanisms to impose sanctions on those liable for violating a norm. For instance, in the context of traffic law, if a vehicle is found to be the cause of an accident, the driver who is controlling the vehicle will be liable for the damage and face some sanctions. However, by giving more autonomy to artificial systems (such as autonomous vehicles), one cannot see them as object-like tools that merely follow instructions. To effectively reason about liabilities in view of these new forms of autonomy, we *need context-aware blameworthiness reasoning tools as a basis for effective liability measures to ensure the legality of TAS.*

An autonomous vehicle is not receiving direct instructions. Thus, when collisions occur, a judge cannot simply apply "Qui facit per alium, facit per se", *who acts through another does the act himself* (Conard 1948; Norman and Reed 2010) to see the owner as the only liable agent. Note that we are not motivating the idea to see an artefact as liable but to articulate the challenge of reasoning about liabilities when the only de facto agent who is in full control of the vehicle is not the driver. The control is shared; thus, it is reasonable that any involved agent with a degree of autonomy takes a degree of liability if a failure occurs. This makes the process of liability reasoning complex as, in each and every case, the judge should take into account not only the motives and abilities of the drivers, but also those of the manufacturers of some key elements of the involved vehicles, designers of decision-making components and infrastructure-related entities.

It is clear that introducing new forms of autonomy, the level of automation and the involvement of numerous (semi) autonomous entities in each and every case result in cumbersome legal procedures. Note that declaring that the use of a particular system is legal is different from determining if a particular (legal-to-use) system caused an illegal behaviour. In principle, deriving whether using a product is legal depends on (and can be derived from) what a governing jurisdiction has chosen to declare legal. However, determining whether a particular autonomous system or an autonomous component of a human–AI system caused an *illegal behaviour* is not a straightforward procedure given new forms of non-human autonomous agents (Chesterman 2021). A government could legislate that all autonomous systems are now legal (or illegal), but that is not sufficient without providing scalable methods to monitor their behaviour and tools able to distinguish sanctionable components for a liable behaviour. Capturing all actions and communications among components of autonomous systems will be an inefficient and resource-intensive task. To avoid that, formal methods for responsibility reasoning can be an important tool, as they allow identifying minimal features that are necessary to be recorded for reasoning about different forms of responsibility (Yazdanpanah et al. 2021a). For example, if a crash among a couple of autonomous vehicles occurs, the main source for reasoning about liable bodies are log files and records stored in those vehicles and on the cloud. At that point, we require techniques to analyse such a large dataset, to use a formal responsibility reasoning method (rooted in the normative theory that the legal authorities are using), and to decide who is to what extent sanctionable for the crash in an automated manner. We argue that, for effective deployment of autonomous systems, it is neither effective nor efficient to rely on non-automated resource-consuming judiciary processes. Otherwise, we will automate transportation and manufacturing, but require much more capacities, in human labour, time and judiciary expertise, to judge each and every incident of failure. This is not an attempt for full automation of the judiciary system, but, in contrast, a proposal to capture the capacities of non-human agents, integrate them with social values and develop human-centred legal decision support tools for TAS.

As a way forward, we call for the integration of human-dependent behaviour enforcement methods (e.g. imposing limitations on resources) with mechanisms and coordination measures that are applicable to artificial agents. To that end, the literature on normative multi-agent systems (Boella et al. 2006) offers methods for incentive engineering and norm-aware mechanism design (Castelfranchi 1998; Bulling and Dastani 2016), techniques for sanction-based enforcement (Dell'Anna et al. 2020) and models for integrating social norms and ethical values into the governance of socio-technical systems (Singh 2013). Such techniques provide a base for effective liability measures in view of new forms of autonomy in TAS. This perspective defends the idea of imposing sanctions, not to merely punish the agents, but with the overarching goal to nudge the behaviour of autonomous agents, and in turn the behaviour of the collective, towards contextual human-centred values.

4 Concrete research directions

In this section, we elaborate on three concrete domains of responsibility research with the potential to contribute to addressing the highlighted challenges. In each domain, we highlight open problems and solution concepts that relate to challenges we already discussed, and depict a research approach to support TAS.

4.1 Developing responsibility-aware agents

In human societies, being responsible is conditioned on the capacity to reason and judge the consequences, as well as the awareness and knowledge of forward-looking responsibilities that an individual ought to fulfil over time, e.g., as tasks and roles ascribed to her (Hart 1968). Such an awareness is also necessary to justifiably ascribe backward-looking blame, liability, and consequently punishment to agents. If we are expecting artificial agents to behave in a responsible way, it is natural to ensure that they are able to reason about different forms of responsibility.

We see the ability to reason about responsibilities as a meta-reasoning capacity. Here, meta-reasoning refers to the capacity of agents to reflect on their own reasoning (Cox and Raja 2011). While being able to analyse inputs and flexibly choose an optimal action with respect to the agent's goals defines it to be intelligent (Wooldridge and Jennings 1995), we see responsibility reasoning as a meta-level capacity that requires the agent to be self-aware and possess (partial) situation awareness (Dennis and Fisher 2020; Stanton et al. 2017). This enables the agent to be aware of and reason about its own responsibilities and the responsibilities of other human/artificial agents in the environment. For instance, imagine an autonomous vehicle with the goal to reach to its destination as early as possible but also in view of responsibilities that may be assigned for its actions if harm is caused. The basic idea is that artificial agents would take into consideration the potential costs of being treated as accountable, e.g. to capture the risk that the performance of autonomous vehicles will not be evaluated only based on reaching to their destination as early as possible but may be discounted if harms for which they are accountable will emerge. This, in turn, will make the artificial agents more prudent, i.e. to prefer less risky conducts or to invest in strategies designed to reduce uncertainties.

In this way, a responsibility-aware agent would be able to reason about the consequences of its available actions not only in view of its own goals but also with respect to its degree of responsibility for potential consequences. Note that, similar to human agents, such reasoning will be based on the agent's limited knowledge and observability. Here, explainability of AI systems and opening the black boxes (Dubljevic´ and Racine 2014) is key for reliable responsibility reasoning. In other words, the behaviour of AI agents should be interpretable—albeit to an extent that supports the privacy of entities they represent. Following Dignums' proposal for the so-called *social agents* (Dignum and Dignum 2020), we envisage responsibility-aware agents operational in a social context to weigh their available actions according to responsibilities. This way, in addition to a traditional decision-making unit for evaluating the optimality of actions—merely with respect to the agent's goals—artificial agents require a meta-level unit to represent and reason about their degree of responsibility under different eventualities. By enriching agents with such a responsibility reasoning unit, the procedure of responsibility verification and evaluation of consequences will be integrated into agents' decision making and ascription of utility to available actions. Then, a responsibility-aware agent can update the utility attached to each action in view of responsibilities (e.g. by reasoning about the extent to which it will be seen as responsible for the violation of an established norm and the amount of sanction attached to such a violation). Further calculation of utility of the consequences can be the basis for the agent's decision.

4.2 Developing responsibility reasoning tools operational under norm conflict

As discussed by Yazdanpanah et al. (2021a), ascribing liability in autonomous systems is conditioned on the violation of norms and socially established values. This raises the challenge of how to determine liability when adhering to one norm results in the violation of another norm. When an agent violates a norm, e.g. by performing a prohibited act, they will be seen as a responsible agent for the violation and consequently liable for the caused harm only if they

had another option available. This is known in the literature on moral philosophy as the *avoidance potential* condition (Braham and van Hees 2012). In other words, if one had no option other than doing *X*, they can be seen as the cause of *X* but not morally responsible and liable for *X* and what doing *X* implies (in a social context). For instance, an autonomous vehicle with the option to swerve to a side and avoid a crash with a pedestrian can be seen as responsible if the crash occurs. Here, the vehicle is violating the norm that colliding with pedestrians is prohibited by the traffic law. The challenge arises when the agent's potential to avoid violating a norm results in the violation of another norm. What if swerving aside results in hitting another pedestrian? In principle, the vehicle has the potential to avoid hitting only one of the pedestrians by hitting the other one. Reasoning about the extent of responsibility of such an artificial agent is an open problem that requires computational methods for evaluating the importance and priority of norms and developing responsibility reasoning tools that are operational under norm conflicts.

Norm conflicts are dilemmatic situations where an agent's compliance with one norm results in the violation of another (Michael and Anderson 1987). Such situations are not limited to conflicts between similar norms with explicit regulations (like our pedestrians' case). The conflict can be between norms with different natures, e.g. the moral norm to deliver your tasks and the norm to comply with traffic regulations. While, in human societies, we expect individuals to be able to reason about such tradeoffs and make decisions to the best of their ability, AI-based agents require tools to reason about such aspects. Without such tools, they may simply prioritise the delivery of an insignificant task over the compliance with a social norm and cause harm. Such conflicts may also occur in relation to preserving the privacy of users. For instance, an autonomous vehicle may rightfully follow its owner's instruction and opt to keep some information, e.g. about its internal states and plans, private. This way, it complies with the norm to preserve the privacy of its user. However, such a conservative behaviour may avoid others from reaching the information required for avoiding a collision. Such forms of norm conflicts are well studied in the legal context (Vranes 2006). However, how moral and legal principles for handling norm conflicts can be tailored to incorporate new forms of autonomy is still an open problem.

For instance, imagine an autonomous vehicle with a passenger on board who urgently requires medical attention. Through the journey to the hospital, the vehicle is forced to choose between (in this case) two options: to keep its speed below the safe limit (which increases the chance of arriving late and causing harm to its passengers) or going above the speed limit (which violates safety norms). In this case, both options are normatively undesirable as they violate established norms that expect the vehicle to avoid causing harm to the best of its ability. As discussed in Bonnefon et al. (2016), re-solving such situations and understanding how to ascribe responsibilities to the agents involved are crucial for ensuring the reliability and safety of AI systems, and accordingly their embedding in society.

To address norm conflicting situations as a base for a justifiable responsibility ascription, we aim to develop norm ranking tools, rooted in argumentation theory (Modgil and Luck 2008) and value-aware norm selection methods (Serramia et al. 2018).[8] This way, we can formulate responsibility quantification techniques that capture not only one norm but a ranked set of norms. Note that the aim is not to establish a unique ranking but to consider various normative theories and provide a set of rankings. Indeed, the intention is not to work on developing novel normative ethics but to allow formalising them in a computer-interpretable language and enable AI systems to reason about and decide about responsibilities.

Thinking of a future in which AI technology is embedded in our society, establishing agents' avoidance potential is not only about the physical actions available to them, but also concerns what they knew at what time and what sorts of communicative actions were available to them. One can imagine that the knowledge of the predicament and norms is distributed, and any agent (partially) aware of the situation had the chance to contribute to avoiding the harm, and thus deserves a degree of liability. This calls for further investigations on how distributed situation awareness (Stanton 2016) relates to responsibility reasoning under norm conflict. To that end, formal behaviour verification techniques in computer science can be used to evaluate if and why a rule (or a set of rules) was violated by an autonomous AI system, and whether (given the limitations of the system and uncertainties in the environment) they could avoid the violation. We argue that, to make TAS legally align with rules and regulations, such a verification step needs to precede the responsibility ascription phase.

4.3 Developing hybrid responsibility learning-reasoning tools

Moving from theoretical responsibility reasoning tools towards real-life applications necessitates capturing various forms of uncertainty within the responsibility ascription

[8] For related computational methods to support representing and reasoning about norm conflict in multi-agent settings, see computational techniques rooted in deep learning and contract theory for identifying norm conflicts (Aires et al., 2017; Aires and Meneguzzi, 2017), agent-based computational methods for resolving norm conflicts (Kasenberg and Scheutz, 2018; Kollingbaum and Norman, 2004) and their application for handling inconsistencies in norm-regulated virtual organisations (Vasconcelos et al., 2007).

process. Such uncertainties are not only on the side of agents, to whom we are aiming to ascribe responsibility, but also on the side of the "judging" agent who aims to ascribe responsibilities. Note that following the idea to integrate a responsibility reasoning unit into AI agents, as discussed earlier, an agent may play both roles: of being the actor who takes responsibility and also the judge who reasons about her own responsibilities as well as responsibilities of others.

The presence of uncertainties motivates the development of responsibility ascription tools operational under imperfect information. In other words, AI agents need to be able to reason about responsibilities given their own uncertain understanding of the world. In dynamic multi-agent settings, the knowledge agents have about their environment, their own abilities and abilities of others is in most cases imperfect. This includes not only their knowledge about the consequence of the actions they perform, but also their understanding of established norms and sanctions attached to violating norms. Note that an agent's knowledge affects different forms of responsibility differently. For instance, knowledgeably causing harm is crucial for liability ascription but not necessary for causal responsibility (Hart 1968).

In dynamic settings, human agents have the capacity to learn about norms, and norm changes (Castelfranchi 2015), as the multi-agent system evolves and accordingly ought to reason about their responsibilities (e.g. for normatively undesirable situations). To have a smooth and effective embedding of AI into society, we need to enable AI agents, as well, to integrate their dynamic understanding and learning about the world into their responsibility reasoning process. To capture such dynamics and model hybrid[9] notions of responsibility, we propose the integration of norm-learning methods, e.g. Dell'Anna et al. (2020), with frameworks that allow combining symbolic and sub-symbolic features of the environment (Zhang et al. 2020). Such an integration allows learning and reasoning about the world in a dynamic fashion and formulating hybrid learned-reasoned notions of responsibility.

5 Positioning: complementary research avenues

In this section, we position our suggested research agenda and elaborate on relations to proposals focussed on neighbouring domains. Our research agenda relates to recent proposals focussed on social agents (Dignum and Dignum 2020), ethical multi-agent systems (Murukannaiah et al. 2020) and the application of formal verification for ethical autonomous systems (Dennis et al. 2016). In the following, we elaborate on relations, differences and points where these neighbouring domains complement our approach to address challenges for establishing trustworthy autonomous systems.

In principle, Dignum and Dignum (2020) argue that the agent technology needs to incorporate social aspects, as an intrinsic component, to remain relevant for solving real-life problems. They argue for the importance of novel agent architectures that are aware of social values and have the capacity to reason about agents' goals in view of the agents' social relations. This follows the theory that sees intelligence as a social phenomenon defined, understood, and exhibited by an agent in relation to its society (Epstein and Axtell 1996). Practically, what goals an agent selects to commit to follows its preferences and, in turn, such preferences reflect norms and values that the agent adopted from its surrounding social context. Our proposal to enable agents to reason about responsibilities and to use responsibility reasoning as a means for ensuring the reliability and legality of TAS focuses on a specific aspect of Dignum and Dignum's social agents as being *responsibility-aware* agents. We argue that being aware of responsibilities of the agent itself as well as responsibilities of others (as discussed in Sect. 5) is a key step towards developing social agents. For instance, task and role relations are key in forming and maintaining agent societies where forward-looking responsibility notions, in terms of what an agent is able to deliver strategically, is a reliable notion for allocating tasks and organisational roles to agents (Yazdanpanah et al. 2020)

In a related research agenda on ethical multi-agent systems, Murukannaiah et al. (2020) argue that addressing ethical concerns related to the behaviour of AI systems requires multi-agent modelling of what ethicality means for a society of agents, methods to analyse such a notion in the multi-agent context and finally tools to elicit it. Their focus on the need for methods to determine what is ethical (e.g. in terms of the behaviour of a sociotechnical system or a situation that may occur as a result of collective decisions in a multi-agent system) provides an input and is necessary for ascribing responsibilities in TAS. As discussed, ascribing responsibility to an agent is always in relation to a state of affairs. For instance, agent A or a group of agents G may be responsible for the occurrence of situation S or behaviour B. Then understanding whether S or B are ethically undesirable is crucial to determine whether the responsible agent A or agents involved in the responsible group G are to be sanctioned or seen as liable. To that end, the line of research suggested by Murukannaiah et al. (2020) is key for what we called *liability reasoning in view of new forms of autonomy* (in Sect. 4) and contributes to ensuring the legality of TAS.

[9] Here, hybrid refers to cases in which learned norms are generated by a combination of observations and formal methods that are given a-priori. The hypothesis that justifies this research is that it is necessary to build the cognitive capacity of an artificial agent to evaluate its own degrees of responsibility and plan actions accordingly.

Finally, from a methodological point of view, reasoning about responsibilities in TAS requires verifiable techniques rooted in formal methods and system verification. The approach suggested in Dennis et al. (2016) uses a cognitive model of agents and provides a method for task planning such that the AI system preserves some given ethical principles. Despite computational complexity issues (that are common for cognitive agent models), their approach is applicable for determining the ethicality and reliability of safety-critical systems, such as aircraft fleet or connected autonomous vehicles. It complements our suggestion to apply responsibility reasoning for ensuring the reliability of TAS. They consider the fact that an agent may not be able to avoid an ethically undesirable action and use a ranking of ethical principles to resolve this issue. This is an interesting approach that abstracts from the norm-level or action-level rankings, and focuses on a ranking on the high-level principles. We argue that ensuring the reliable and legal behaviour of TAS requires a more granular ranking (on the norm-level as suggested earlier) mainly because even within a principle, agents may need to prioritise among their personal norms (e.g. as values to preserve), organisational norms (e.g. as task to deliver) and social norms (e.g. as regulations to follow). Another point of commonality is with our concern to ascribe responsibilities based on the concept of *avoidance potential*. We propose that Dennis et al.'s verification tools to reason about the ethicality of AI systems' behaviour can be a base for reasoning about liabilities, as they can be integrated with logic-based methods, e.g. in Naumov and Tao (2020); Alechina et al. (2017), for reasoning about responsibilities.

6 Conclusion

The presented work highlights open challenges in reasoning about responsibility in autonomous systems and discusses how various notions of responsibility relate to reliability and legality of such systems. We presented three research themes focussed on the development of (1) *responsibility-aware agents*, (2) *tools for responsibility reasoning under norm conflict* and (3) *hybrid responsibility learning-reasoning methods*.

Developing responsibility-aware agents supports the idea that agents need to integrate potential responsibilities into their decision-making process. This promotes more prudent action choices, and supports trustworthiness of autonomous systems from the users' point of view. Furthermore, to determine liabilities in real-life situations where norms and social values may conflict with one another, developing responsibility reasoning tools—operational under norm conflict—allows effective ascription of sanction and penalties to the involved agents. This ensures that, even under such norm conflicts, wrong-doing can be addressed in TAS and will be penalised proportionally. Finally, to capture inherent uncertainties in different application domains, developing hybrid responsibility modelling tools allows combining data-driven models with logic-based techniques and, in turn, ensures that responsibilities will not be voided in TAS even in the case of high uncertainties.

In addition, we elaborated on methods and technical approaches that are applicable for investigating these open lines of research and linked them to related research avenues. Crucially, we argued that responsibility research has the potential to contribute to the interdisciplinary endeavour on ensuring trustworthy autonomous systems and, in turn, to support an effective embedding of artificial intelligence technologies into society.

Acknowledgements This work is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) through the Trustworthy Autonomous Systems Hub (EP/V00784X/1), the platform grant entitled "AutoTrust: Designing a Human-Centred Trusted, Secure, Intelligent and Usable Internet of Vehicles" (EP/R029563/1), and the Turing AI Fellowship on Citizen-Centric AI Systems (EP/V022067/1). This research was also (partly) funded by the Hybrid Intelligence Center, a 10 year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, grant number 024.004.022 and by EU H2020 ICT48 project "Humane AI Net" under contract #952026. The authors also thank the anonymous reviewers of the special issue on "Embedding AI in Society" (Journal of AI & Society) and participants, organisers, and referees of the Rabb Symposium at North Carolina State University on *Embedding AI in Society* (18-19 February 2021). Their constructive comments significantly improved the work and are much appreciated.

Data availability Not applicable.

References

Abeywickrama DB, Cˆırstea C, Ramchurn SD (2019) Model checking human-agent collectives for responsible AI. In: *28th IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2019, New Delhi, India.* 1–8. New York, NY. IEEE.

Aire, JP, Meneguzzi F (2017) Norm conflict identification using deep learning. In: *International Conference on Autonomous Agents and Multiagent Systems*. 194–207. Springer.

Aires JP, Pinheiro D, Lima VSD, Meneguzzi F (2017) Norm conflict identification in contracts. Artific Intell Law. 25(4):397–428

Alechina N, Halpern JY, Logan B (2017). Causality, responsibility and blame in team plans. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2017, Sa˜o Paulo, Brazil.* 1091–1099. Richland, SC. IFAAMAS.

Alechina N, Logan B (2020) State of the art in logics for verification of resource bounded multi-agent systems. Fields of Logic and Computation III—Essays Dedicated to Yuri Gurevich on the Occasion of His 80th Birthday. Springer, Cham, pp 9–29

Birolini A (2013) *Reliability engineering: theory and practice*. Springer Science & Business Media. https://doi.org/10.1007/978-3-662-05409-3

Boella G, van der Torre LWN, Verhagen H (2006) Introduction to normative multiagent systems. Comput Math Organ Theory 12(2–3):71–79

Bonnefon J-F, Shariff A, Rahwan I (2016) The social dilemma of autonomous vehicles. Science 352(6293):1573–1576

Braham M, van Hees M (2011) Responsibility Voids. The Philosophical Quar Terly 61(242):6–15

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

Bratman ME (1993) Shared intention. Ethics 104(1):97–113

Bratman ME (2007) Structures of agency: essays. Oxford University Press, Oxford

Bryson J, Winfield AFT (2017) Standardizing ethical design for artificial intelligence and autonomous systems. Computer 50(5):116–119

Bulling N, Dastani M (2013) Coalitional responsibility in strategic settings. In: *Proceedings of the 14th International Workshop on Computational Logic in Multi Agent Systems, CLIMA XIV, Corunna, Spain*. 172–189. Berlin, Heidelberg. Springer.

Bulling N, Dastani M (2016) Norm-based mechanism design. Artif Intell 239:97–142

Castelfranchi C (1998) Modelling social action for AI agents. Artif Intell 103(1–2):157–182

Castelfranchi, C. (2015). A cognitive framing for norm change. In: *Proceedings of the 11th International Workshop on Coordination, Organizations, Institutions, and Normes in Agent Systems—COIN 2015, Istanbul, Turkey*. 22–41. Cham, Springer.

Champlin, T. S. (1994). Responsibility. *Philosophy*, 69(268):254–255. Chesterman, S. (2021). *We, the robots?* Cambridge University Press.

Chockler H, Halpern JY (2004) Responsibility and blame: a structural-model approach. J Artific Intell Res 22:93–115

Conard A (1948) What's wrong with agency. J Leg Educ 1:540

Constantinescu M, Voinea C, Uszkai R, Vica˘ C (2021) Understanding responsibility in responsible AI. dianoetic virtues and the hard problem of context. Ethics Inform Technol 23(4):803–814

Cox MT, Raja A (2011) Metareasoning: thinking about thinking. MIT Press, Cambridge, MA

Dastani M, Dignum F, Meyer JC (2003) Autonomy and agent deliberation. In: *Proceedings of the 1st International Workshop on Computational Agents and Computational Autonomy—Potential, Risks, and Solutions*. 114–127.

Dastani M, Yazdanpanah V (2022) Responsibility of ai systems. AI Soc. https://doi.org/10.1007/s00146-022-01481-4

Dell'Anna D, Dastani M, Dalpiaz F (2020) Runtime revision of sanctions in normative multiagent systems. Auto Agents Multi Agent Syst 34(2):43

Dennis LA, Fisher M (2020) Verifiable self-aware agent-based autonomous systems. Proc IEEE 108(7):1011–1026

Dennis LA, Fisher M, Slavkovik M, Webster M (2016) Formal verification of ethical choices in autonomous systems. Robot Auton Syst 77:1–14

Dignum, V. (2019). *Responsible Artificial Intelligence—How to Develop and Use AI in a Responsible Way*. Artificial Intelligence: Foundations, Theory, and Algorithms. Springer, Cham.

Dignum V, Dignum F (2020) Agents are dead. long live agents! In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '20, Auckland, New Zealand*. 1701–1705. Richland, SC. IFAAMAS.

Dubljevic´ V, Racine E (2014) The adc of moral judgment: opening the black box of moral intuitions with heuristics about agents, deeds, and consequences. AJOB Neurosci 5(4):3–20

Epstein JM, Axtell R (1996) *Growing artificial societies: social science from the bottom up*. Brookings Institution Press.

European Commission (2021). Europe fit for the digital age: Commission proposes new rules and actions for excellence and trust in artificial intelligence. https://ec. europa.eu/commission/presscorner/detail/en/ip 21 1682. Accessed: 2021–06–09.

European Commission: The High-Level Expert Group on AI (2019). Ethics guidelines for trustworthy AI. https://ec.europa.eu/digital-single-market/en/news/ ethics-guidelines-trustworthy-ai. Accessed: 2021–02–15.

European Parliament (2021). Meps debate new "ai act" with ec vp margrethe vestager. https://www.europarl.europa.eu/news/en/press-room/20210527IPR04915/ meps-debate-new-ai-act-with-ec-vp-margrethe-vestager. Accessed: 2021–06–09.

Ferber J, Gutknecht O, Michel F (2003) From agents to organizations: An organizational view of multi-agent systems. In: *Proceedings of the 4th International Workshop on Agent-Oriented Software Engineering, AOSE 2003, Melbourne, Australia,* 214–230. Berlin. Heidelberg, Springer.

Flemisch F, Abbink DA, Itoh M, Pacaux-Lemoine M-P, Weßel G (2016) Shared control is the sharp end of cooperation: towards a common framework of joint action, shared control and human machine cooperation. IFAC-PapersOnLine 49(19):72–77

Georgeff MP, Pell B, Pollack ME, Tambe M, Wooldridge MJ (1998) The belief-desire-intention model of agency. In: *Intelligent Agents V, Agent Theories, Architectures, and Languages, 5th International Workshop, ATAL '98, Paris, France, Proceedings*. 1–10. Berlin, Heidelberg, Springer.

Hart H (1968) Punishment and responsibility. Philosophy 45(172):210–237

Horling B, Lesser VR (2004) A survey of multi-agent organizational paradigms. Knowl Eng Rev 19(4):281–316

Jennings NR, Moreau L, Nicholson D, Ramchurn SD, Roberts SJ, Rodden T, Rogers A (2014) Human-agent collectives. Commun ACM 57(12):80–88

Johnson M, Bradshaw JM, Feltovich PJ, Jonker CM, van Riemsdijk MB, Sierhuis M (2014) Coactive design: designing support for interdependence in joint activity. J Human-Robot Inter 3(1):43–69

Kasenberg D, Scheutz M (2018) Norm conflict resolution in stochastic domains. Proceed AAAI Conf Artific Intell. https://doi.org/10.1609/aaai.v32i1.11295

Kollingbaum M, Norman T (2004) Strategies for resolving norm conflict in practical reasoning. In: ECAI workshop coordination in emergent agent societies. 2004, pp 1–10

McLaughlin JA (1925) Proximate cause. Harv Law Rev 39(2):149–199

Michael DN, Anderson WT (1987) Norms in conflict and confusion: six stories in search of an author. Technol Forecast Soc Chang 31(2):107–115

Modgil S, Luck M (2008) Argumentation based resolution of conflicts between desires and normative goals. In *Argumentation in Multi-Agent Systems, Fifth International Workshop, ArgMAS, Estoril, Portugal. Revised Selected and Invited Papers*. 5384. 19–36. Berlin, Heidelberg. Springer.

Murukannaiah PK, Ajmeri N, Jonker CM, Singh MP (2020) New foundations of ethical multiagent systems. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '20, Auckland, New Zealand*. 1706–1710. Richland, SC. IFAAMAS.

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

Norman TJ, Reed C (2000) Delegation and responsibility. In *Intelligent Agents VII. Agent Theories Architectures and Languages, 7th International Workshop, ATAL 2000, Boston, MA, USA, Proceedings*. 136–149. Berlin, Heidelberg. Springer.

Norman TJ, Reed C (2010) A logic of delegation. Artif Intell 174(1):51–71

O'Connor P, Kleyner A (2012) Practical reliability engineering. John Wiley & Sons

Office for Artificial Intelligence (2020). A guide to using artificial intelligence in the public sector. https://www.gov.uk/government/publications/ a-guide-to-using-artificial-intelligence-in-the-public-sector. Accessed: 2021–02-

Rahwan I, Cebrian M, Obradovich N, Bongard J, Bonnefon J-F, Breazeal C, Crandall JW, Christakis NA, Couzin ID, Jackson MO et al (2019) Ma- chine behaviour. Nature 568(7753):477–486

Ramchurn SD, Huynh TD, Wu F, Ikuno Y, Flann J, Moreau L, Fischer JE, Jiang W, Rodden T, Simpson E, Reece S, Roberts SJ, Jennings NR (2016) A disaster response system based on human-agent collectives. J Artif Intell Res 57:661–708

Ramchurn SD, Stein S, Jennings NR (2021) Trustworthy human-AI Part- nerships. Iscience 24(8):102891

Rao AS, Wooldridge M (1999) *Foundations of Rational Agency*. 1–10. Springer, Dordrecht.

Russell S (2019) Human compatible: Artificial intelligence and the problem of con- trol. Viking, New York, NY

Santoni de Sio F, Mecacci G (2021) Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology*. 1–28.

Santoni de Sio F, van den Hoven J (2018) Meaningful human control over autonomous systems: a philosophical account. Front Robot AI 5:15

Searle JR (1989) How performatives work. Linguist Philos 12(5):535–558

Searle JR (1995) The construction of social reality. Free Press, New York, NY

Serramia M, Lo´pez-Sa´nchez M, Rodr´ıguez-Aguilar JA, Rodr´ıguez M, Wooldridge M J, Morales J, Anso´tegui C (2018) Moral values in norm de-cision making. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2018, Stockholm, Sweden,* 1294–1302, Richland, SC. IFAAMAS.

Singh MP (2013) Norms as a basis for governing sociotechnical systems. ACM Trans Intell Syst Technol. 5(1):21

Stanton NA (2016) Distributed situation awareness. Theoretic Issues Er Gonomics Sci 17(1):1–7

Stanton NA, Salmon PM, Walker GH, Salas E, Hancock PA (2017) State-of-science: situation awareness in individuals, teams and systems. Er Gonomics 60(4):449–466

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

van der Waa J, van Diggelen J, Siebert LC, Neerincx M, Jonker CM (2020) Allocation of moral decision-making in human-agent teams: A pattern approach. International Conference on Human-Computer Interaction. Springer, Cham, pp 203–220

Vardi MY (2020) Efficiency vs resilience: what COVID-19 teaches computing. Communicat ACM 63(5):9

Vasconcelos W, Kollingbaum MJ, Norman TJ (2007) Resolving conflict and inconsistency in norm-regulated virtual organizations. In *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*. 1–8.

Vranes E (2006) The definition of 'norm conflict' in international law and legal theory. Europ J Intern Law 17(2):395–418

Wooldridge MJ, Jennings NR (1995) Intelligent agents: theory and practice. Knowled Eng Rev 10(2):115–152

Yazdanpanah V, Dastani M (2015) Quantified degrees of group responsibil- ity. Coordination, Organizations, Institutions, and Normes in Agent Systems XI- COIN 2015 International Workshops, COIN@ AAMAS, Istanbul, Turkey. Springer, Cham, pp 418–436

Yazdanpanah V, Dastani M (2016) Distant group responsibility in multi-agent systems. In *PRIMA 2016: Princiles and Practice of Multi-Agent Systems—19th International Conference, Phuket, Thailand, Proceedings.* 261–278. Cham, Springer.

Yazdanpanah V, Dastani M, Fatima S, Jennings NR, Yazan DM, Zijm WHM (2020) Multiagent task coordination as task allocation plus task responsibility. In *Multi-Agent Systems and Agreement Technologies—17th European Conference, EUMAS 2020, Thessaloniki, Greece, Revised Selected Papers*. 571–588. Cham, Springer.

Yazdanpanah V, Gerding EH, Stein S, Cirstea C, Schraefel MC, Norman TJ, Jennings NR (2021a) Different forms of responsibility in multiagent systems: sociotechnical characteristics and requirements. IEEE Internet Comput 25(6):15–22

Yazdanpanah V, Gerding EH, Stein S, Dastani M, Jonker CM, Norman TJ (2021b) Responsibility research for trustworthy autonomous systems. In *Proceedings of the 20th International Conference on Autonomous Agents and Multi Agent Systems*. 57–62.

Yeung K (2018) A study of the implications of advanced digital technologies (including ai systems) for the concept of responsibility within a human rights framework. MSI-AUT 2018:5

Zhang Y, Radulescu R, Mannion P, Roijers DM, Nowe´ A (2020) Opponent modelling for reinforcement learning in multi-objective normal form games. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '20, Auckland, New Zealand.* 2080–2082, Richland, SC. IFAAMAS.