Sociotechnical Perspectives on AI Ethics and Accountability
Editors: N. Kokciyan, B. Srivastava, M.N. Huhns, and M.P. Singh
Email: ic6-2021@computer.org

Different Forms of Responsibility in Multiagent Systems: Sociotechnical Characteristics and Requirements

Vahid Yazdanpanah, Enrico H. Gerding, Sebastian Stein,
Corina Cirstea, m.c. schraefel, Timothy J. Norman
University of Southampton

Nicholas R. Jennings
Imperial College London

***Abstract*—Ensuring trustworthy performance of autonomous agents and Multiagent Systems (MAS) requires computational methods and formal tools to support reasoning about different forms of responsibility. In particular, such tools are needed to support identifying agents or agent groups that are *responsible*, *blameworthy*, *accountable*, or *sanctionable* for outcomes of collective decisions, for fulfilling tasks, or for adhering to norms and social values. As a step towards developing computational frameworks to represent, reason about, and distinguish these forms of responsibility in MAS, for the first time, we present sociotechnical characteristics of these notions of responsibility, identify their requirements, and discuss their applicability for coordinating MAS and ensuring their trustworthiness. This is a step towards establishing a research agenda on how computational techniques for reasoning about and distinguishing different forms of responsibility contribute to the transformation towards ethical and trustworthy autonomous systems.**

THE DEPLOYMENT OF AUTONOMOUS AGENTS AND MULTIAGENT SYSTEMS offers a promising prospect for more reliable and efficient performance in various domains. However, at the current stage, developed Multiagent Systems (MAS), e.g., in logistics [1] and governance [2],

mainly aim for reliability and efficiency but are incapable of reasoning about and distinguishing *different notions of responsibility* in such sociotechnical systems. We understand sociotechnical systems as "multistakeholder cyber-physical systems" [2]. Ensuring ethical and trustworthy behaviour of such systems requires satisfying sociotechnical requirements (that are on one hand technical and on the other hand relate to social concern) [3]. For instance, ensuring that an autonomous vehicle behaves in a responsible and ethical way requires capturing the technical requirements (e.g., the ability to detect barriers) as well as the contextual social norms and values (e.g., the ability to learn and reason about driving norms in a city). In particular, *responsibility*—in its various forms—is a notion with sociotechni-cal characteristics as it relates to sociotechnical concepts such as ability, knowledge, task, and norm. Thus, to enable agents in a MAS to reason about responsibility requires distinguishing different forms of responsibility and articulating how each form conceptually relates to *strategic ability* (and distribution of power); *epistemic ability* (and distribution of knowledge); *tasks* (and distribution of obligations); and norms and values (and distribution of preferences).

Developing such a conceptual basis enables reasoning about how and to what extent each agent is *responsible* for an outcome or poten-tially undesirable situation that resulted from the system's behaviour, and whether it should be *blamed* for it, can be seen *accountable*, or be *sanctioned*. Thus, if an undesirable situation, such as a collision among autonomous vehicles occurs, it will be clear which system component should be assigned with which form of responsibility. Autonomous systems and Artificial Intelligence (AI) technologies need to be embedded in society and in such a social context, in which humans and AI systems are expected to effectively interact, different forms of responsibility have distinguishable sociotechnical characteristics and attributes. To foster such an embedding, we follow Jennings' argument, in the foundational work "*On Being Responsible*" [4], that the meta-level notion of responsibility is applicable for coordinating AI systems. Building on this idea, we argue that *different forms* of this notion—in particular, the notions of *responsibility*, *blameworthiness*, *accountability*, and *sanctionability*—are key to developing trustworthy AI systems and ensuring their sociotechnically desirable behaviour.

In principle, giving more autonomy to var-ious components of a multiagent system, one cannot see them as objects that merely follow instructions. For instance, a driverless vehicle is not receiving direct instructions. Thus, when collisions occur, the judge cannot simply apply the well-established judicial axiom that "Qui facit per alium, facit per se" (those who act through another do the act themselves) [5] to see the owner as the only responsible agent. It is reasonable that any involved agent with a degree of autonomy takes a degree of responsibility. In response, and supporting the guidelines published by UK's Office for Artificial Intelligence on the need for responsible and accountable AI systems [6], we acknowledge that autonomous agents in a MAS and the behaviour of MAS as a whole need to be in line with social values and ethical concerns. To that end, computational tools for reasoning about different forms of responsibility can play a key role. Such tools can be embedded into agents to enable them to reason about their responsibility, to identify those to account for an outcome [7], and to ensure the alignment with constitutive norms that represent the implicit values of society, as well as explicit rules and regulations [8] for governing autonomous systems.

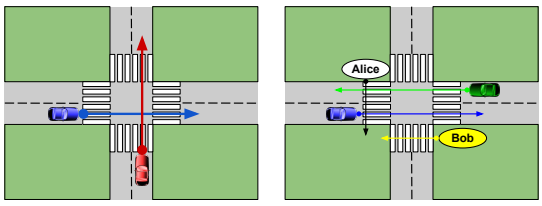
While interest in ensuring that AI systems behave responsibly exists in both scientific work [9], [10] and in regulatory bodies [6], [8], there is a gap in distinguishing various forms of responsibility and relating their characteristics to sociotechnical concepts like strategic ability, knowledge, norms, and tasks. In this work, the emphasis is on how to support this endeavour by presenting the conceptual foundations for developing computational tools to reason about and distinguish *different notions of responsibility*. To that end, for the first time, this paper (i) articulates the strategic, epistemic, normative, and task-oriented conditions behind different notions of responsibility, blameworthiness, accountability, and sanctionability in MAS; (ii) presents challenges and requirements for developing frameworks to represent, reason about, and distinguish these forms of responsibility in the context of

This article has been accepted for publication in IEEE Internet Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MIC.2021.3107334

MAS; and (iii) discusses their applicability for coordinating and ensuring the trustworthiness of MAS.

CONCEPTUAL FOUNDATION

In this section, we present a conceptual account of the notions of *responsibility*, *blameworthiness*, *accountability*, and *sanctionability*, based on the literature on moral philosophy [11], [12]. Responsibility is a relational concept defined between at least two entities A and φ with the generic structure "A *is responsible for* φ" [11]. In this structure, A is an autonomous agent or agent group able to perform actions in an environment and φ is a state of affairs, outcome, task, norm, or in general a potential situation in the environment in which A is active. Note that responsibility reasoning mostly takes place in a given environment, i.e., a fixed setting or a specific contextualised scenario in which agents are active. This is why we simply say "A is responsible for φ" instead of "A is responsible for φ in environment E". Then the nature of φ (e.g., whether it is a norm that agents are expected to adhere to) and the level of influence/activeness of A on φ (e.g., whether A can prevent/provide φ in prospect or could have done so in retrospect) define different forms and degrees of responsibility.



(a) (b)

Figure 1. Intersection Scenarios: (a) the two-vehicle case and (b) vehicles in presence of pedestrians *Alice* and *Bob*.

To discuss different forms of responsibility in MAS, imagine the two scenarios in Figure 1. In the two-vehicle case, Figure 1-a with vehicles red and $blue$, one may be interested in reasoning about ensuring safety in prospect, e.g., to verify if there exist an individual autonomous vehicle or a group (A) able to ensure that both the vehicles pass the intersection safely with no crashing (φ). Neither of the two vehicles can guarantee this individually but they can do so together (one goes first and the second one goes afterwards). We say "$\{red, blue\}$ *is forward-looking responsible for* φ". The temporally dual form of *forward-looking* responsibility is *backward-looking* responsibility. This is when φ already took place (e.g., that a crash occurred) and one would like to determine whether a vehicle could have avoided it, hence can be now seen responsible for φ. In this case, either red or $blue$ could avoid entering the intersection and preclude any possible crash. We say "red *and* $blue$ *are both backward-looking responsible for* φ". Note that although the two notions are related, they are distinguishable as they require the vehicles to posses a different form of ability with respect to φ. The backward-looking form requires the ability to *prevent* an outcome while the forward-looking form requires the group to be able to *provide* it.

Another relevant concept is blame and blameworthiness reasoning. Blameworthiness is inherently backward-looking, meaning that we are mainly reasoning about who to blame for an already materialised φ. As presented, responsibility can be formulated in terms of the ability of vehicles to provide an outcome φ (in prospect) or to prevent it (in retrospect). However, to see them blameworthy, their knowledge about the consequences of their actions is crucial [13]. In Figure 1(a), red may not be able to evaluate $blue$'s speed accurately, hence lack the knowledge that going forward results in a crash. In this case, red is responsible but not necessarily blameworthy. The blameworthy group who could avoid the crash is $\{red, blue\}$ only if they had sufficient knowledge. Basically, blameworthiness is the epistemic form of responsibility, which (as we discuss later) is tightly linked to the distribution of knowledge and communication modes available in the MAS.

While being *responsible* or *blameworthy* for a state of affairs φ merely depends on the strategic and epistemic preconditions that an agent (group) A should satisfy, being *accountable* requires (in addition) the characteristics of φ itself. In principle, accountability ascription follows and builds on a task allocation process. This is, agent A is accountable if φ remains unfulfilled even though A was able and tasked to fulfil it. In the responsibility literature, this is also known as *task responsibility* or *role responsibility* [11]. For

instance, in Figure 1(b), imagine that the $green$ vehicle has some goods on board and is tasked to meet a delivery deadline (task φ). If this task remains unfulfilled although $green$ was capable of fulfilling it, we say $green$ is to account for it.

Similar to accountability, the inherently backward-looking notion of A being sanctionable for φ depends on the nature of φ. In particular, sanctionability concerns a state of affairs that is normatively-loaded, i.e., is an undesirable situation, and for which a sanction is known (see sanctionability in normative multiagent systems [2], [14]). In other words, A is sanctionable if adhering to φ is a norm but according to the history of materialised events this norm is violated while A was able to comply with it. Note that here we are referring to a generic notion of norm as a rule that associates (un)desirability with a situation [14]. In Figure 1(b), imagine that according to the history of events vehicle $blue$ hits the pedestrian $Alice$ while there exists an established norm φ saying that hitting a pedestrian is "*to be avoided*" and violation "*can be sanctioned*" to the amount of ξ. In this case, we argue that determining sanctionability is not simply to apply the norm and see $blue$ as an ξ-sanctionable vehicle. We deem that a key point for ascribing sanctionability is to verify whether $blue$ was *"able"* to comply with φ. This is known in the responsibility literature [12] as the *avoidance potential* condition. In this regard, sanctionability is a normative form of responsibility as it concerns the avoidance potential of agents with respect to norm violations.

In the following, we present conditions, challenges, and ways forward for modelling different forms of responsibility in the context of MAS.

RESPONSIBILITY AS STRATEGIC ABILITY

As discussed, the notion of responsibility as the ability to avoid, also known as *strategic responsibility* [15], is the simplest/weakest one. It disregards the knowledge of agents about the effect of their actions (crucial for *blameworthiness*), the distribution of tasks (crucial for *accountability*), and the normative state of possible outcomes (crucial for *sanctionability*). To model and verify strategic responsibility, we require a framework that is expressive for representing the abilities of agents in a multiagent environment (i.e., what outcomes agents can ensure or avoid) regardless of actions available to other agents in the MAS. Then, in its prospective form, responsibility is about the affirmative power of agents to bring about some state of affairs φ regardless of what others can do. And in its retrospective form, i.e., when a state of affairs φ already occurred, it is to verify whether an agent (group) had the preclusive power to avoid φ.

For instance, our two-vehicle scenario in Figure 1(a) can be modelled using the transition system depicted in Figure 2. Here, different states q_0 to q_4 represent possible situations that may result from agents' actions in the multiagent environment. In this scenario, the two agents can either go *forward* or *stop* and we are interested in reasoning about responsible agents, or agent groups, for safe passing of all the vehicles as a prospective state of affairs (denoted with sp) and also for crashing (denoted with cr). To reason about forward-looking responsibility for safe-passing (sp) in q_0, we look for agents capable of ensuring sp. Neither of the two autonomous vehicles can individually ensure sp because they have no control over the other vehicle going forward at the right time. But they can ensure it collectively, hence the collective can be seen as being forward-looking responsible for sp. In the backward-looking form, e.g., when crash (cr) already occurred, we go back through the history of states (here going back from q_1 to q_0) and realise that both $blue$ and red were individually able to avoid cr in q_0 (by choosing action S). Therefore, both are backward-looking responsible for cr.

As discussed, verifying if an agent (group) is retrospectively responsible is both local and history-dependent. That is, it is a property local to the state from which we are reasoning and depends on the particular history of materialised actions. For instance, to reason about responsible vehicles for a crash that took place in q_1, it is sufficient to consider already occurred events with no attention to what potentially occurs after q_1. This temporal dependence is crucial for a realistic formulation of the more complex notions of *blameworthiness* and *sanctionability*. In principle, if a state of affairs is possible, necessarily there exists a group to make sure that it takes place in prospect and if an avoidable situation already took

This article has been accepted for publication in IEEE Internet Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MIC.2021.3107334

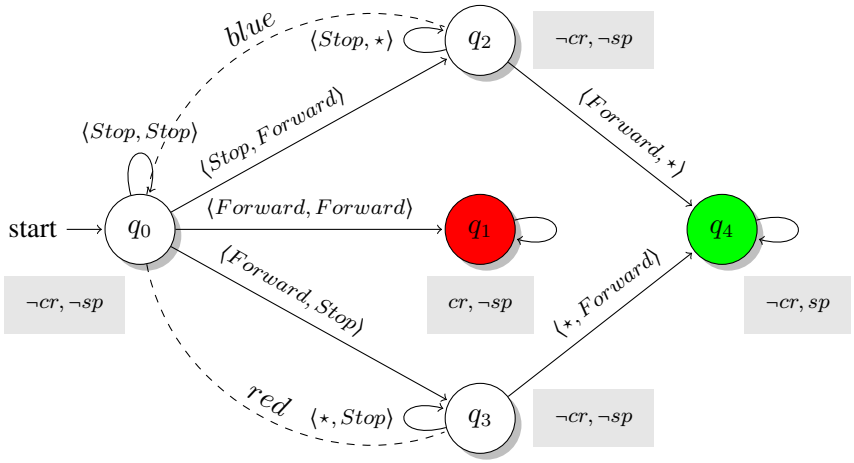


Figure 2. As presented in Figure 1(a), imagine two vehicles red and $blue$ with possible actions to go forward or stop. Labels on arrows are action profiles for $blue$ and red respectively, e.g., in q_0, if $blue$ goes forward and red stops, we go to state q_3. Moreover, $\star$ stands for any available action. We coloured the *safe-pass* state q_4 (where sp holds) in green and the *crash* state q_1 (where cr holds) in red. Epistemic limitations are represented by dashed lines labelled with the agent(s) who cannot distinguish the states a line connects.

place, there exists a group responsible for it. In upcoming sections, we show that the latter notion provides necessary but not sufficient conditions for *blameworthiness* while the former will be a base for *accountability*.

KNOWLEDGE IS KEY TO BLAMEWORTHINESS

As presented, an agent or agent group can be responsible for a state of affairs even if they lack information about the consequences of their actions. However, for modelling and reasoning about the (inherently backward-looking) notion of blameworthiness in a MAS, this epistemic aspect plays a key role. To capture the epistemic aspect of blameworthiness, we need to verify whether agents were knowledgeable about the consequences of their actions and how their actions link to the avoidance/occurrence of a state of affairs. Then an agent (group) Γ can be seen blameworthy for a materialised outcome φ, only if it had actions available to avoid φ effectively, in view of the agents' imperfect information.

The epistemic condition makes blameworthiness distinguishable from—and harder to verify in comparison to—backward-looking strategic responsibility. On the individual agents' level, the internal cognitive/epistemic states of agents are not always transparent to the reasoner, e.g., due to the encapsulation principle and the need to preserve agents' privacy. On the collective level, epistemic conditions for ascribing blameworthiness to agent groups introduces technical challenges and open research problems, mainly due to different ways that knowledge may be distributed among members of a collective. There are different ways that the knowledge about a strategy s and its effectivity for enforcing an outcome φ is distributed in a group Γ. Note that knowing a strategy should include knowing the likelihood that it will be successful—considering uncertainties in the environment and in the action deliberation phase. In principle, knowing about s and its effectivity to enforce φ may be a *common knowledge* among members of Γ that executing s enforces φ (meaning that they all know that it works and already communicate this in the group); they may be *mutually knowledgeable* about this effectivity (but not sure if others know about this); or it can be a *distributed knowledge* (meaning that if they share their knowledge, they can collectively identify that s is effective).(See [16] for other variants and dynamics of how agent groups can knowingly enforce an outcome.) As an open research problem, such a distinction between different forms of knowledge

necessitates developing *context-aware* computational notions of blameworthiness. For instance, in some domains (e.g., among vehicles owned or manufactured by the same firm) the context implies the inter-agent capacity to communicate knowledge, hence the applicability of a blame notion that relies on distributed knowledge. However, in distributed domains (e.g., in the context of the Internet of Things/Vehicles), one needs to capture potential incompatibilities among agents and even cases of miscommunication or malicious communication.

Imagine the presented two-vehicle scenario. To determine blameworthiness for an already materialised crash (cr in q_1), it is necessary to verify if a strategy to avoid cr is in possession of any vehicle or vehicular group. Here, we have that both for red and for $blue$, the individual strategy to stop in indistinguishable states (i.e., in q_0 and q_2 for $blue$ and in q_0 and q_3 for red) could avoid the crash. Hence, in this case, they are not only responsible but also blameworthy for cr.

As discussed, an agent (group) is blameworthy for an outcome φ if it happens although the group was knowingly capable of avoiding it. This is different from seeing them accountable, as they were not tasked to avoid φ, and neither are they sanctionable, as avoiding φ was not an established norm.

ACCOUNTABILITY AS A TASK-ORIENTED VARIANT

Accountability is about being responsible for a task that is not properly discharged [11]. After the allocation of a task "*bring about* φ" to the agent or agent group Γ, it is expected to see to it that φ materialises. Then, if the task is dismissed, Γ is to account for it, i.e., is accountable for not fulfilling their task. In this regard, the exact proce-dure of task allocation is an input for accountabil-ity ascription, as it determines which agents are expected to do what tasks by what deadline and how/whether the allocator is taking into account contextual requirements. Such requirements determine the way the allocator captures the agents' capacity (e.g., how many tasks one can deliver), strategic ability (e.g., whether one can deliver a task regardless of what others do), and available resources (e.g., how much time it takes to deliver a task and whether tasks are allocated efficiently, or to the most qualified agents or agent groups). This way, an agent may be accountable given one task allocation procedure but not another one.

For accountability ascription, the main concern is to verify if an agent or agent group fails to discharge an allocated task even though they were capable of delivering it. So, in addition to the already discussed conditions for modelling strategic abilities under epistemic limitations (key for reasoning about strategic responsibility and blameworthiness), accountability ascription requires taking the task allocation process into account. As discussed in [17], *justifiability* of seeing an agent (group) accountable depends on the *validity* and *suitability* of the preceding task allocation process (i.e., that all what should be done is allocated to agents capable of delivery).

In multiagent settings, the need for ascribing collective accountability motivates focusing on open problems on (1) how to capture incompatibilities among agents (as a cause of being unable to communicate and coordinate towards fulfilling an allocated task) and (2) how institutional actions like delegation [5] affect the accountability distribution. In particular, delegation does not necessarily lead to transferring full accountability to the delegatee but introduces a new accountability dimension (accountability on the breach of a duty versus accountability for unsuitable delegation of tasks).

SANCTIONABILITY IN NORMATIVE SETTINGS

In Figure 1(b), imagine that $blue$ is tasked to reach to the other side of the intersection to pick a waiting passenger and waiting for pedes-trian A results in missing the deadline. On one hand, $blue$ has the task to go forward and on the other hand, going forward may result in a crash. To resolve such a dilemma and ensure that some values (e.g., safety) should be universally preserved, we require an explicit representation of such values [9] and incentivization mechanisms to sanction those who dismiss them [14]. To that end, we require integrating a model of norms to capture the normative aspects of MAS and develop computationally verifiable frameworks to reason about *sanctionability*.

Normativity of MAS is often modelled by labelling a subset of potential situations as de-

sirable and some as undesirable (allowing space for the third category of neither desirable nor undesirable situations, also known as permitted outcomes) [2], [14]. In a norm-aware MAS—with coherent and non-conflicting norms—agents are expected to adhere to norms and behave in compliance with the spectrum of desired behaviours. In this view, an agent (group) is sanctionable if it is blameworthy for violating a norm.

For instance, in Figure 1(b) where we have autonomous vehicles in presence of pedestrians $Alice$ and Bob, one can add an explicit norm that vehicles should avoid the undesirable out-come of hitting a pedestrian, and that such a collision will be sanctioned with $100k$ dollars. Now, imagine a case in which we know that according to the history of events, $green$ hit $Alice$ and that $blue$ passed the intersection. Note that in this case, $blue$ could hit $Alice$ but did not because $green$ did it first. Thus we have that the group $\{blue, green\}$ is sanctionable as a group that could preclude the collision with $Alice$. We follow [13] and argue that capturing such counterfactual dependencies is crucial for a realistic modelling of sanctionability as a norma-tive extension of blameworthiness.

Moving from AI systems in isolation (e.g., in database systems) towards Human-Agent Collectives (HACs) [18], in which artificial agents interact with humans, necessitates focusing on sanctionability reasoning methods that capture new forms of normativity. We argue that HACs need to capture *dynamic normativity* as the collective behaviour may be interpreted as desirable in one context, culture, or society but undesirable in another one or at a different time (tempo-ral and contextual normativity). This calls for sanctionability models that are local (specific to a point in time) and context-aware (embed a dynamic set of social values/norms). Moreover, to enable applying sanctionability reasoning in real-life settings (e.g., as a base for ascribing liability in the legal domain), an open prob-lem is to integrate intentionality modelling and capture the preference of agents. For instance, in application domains such as smart mobility or disaster management, a mix of humans and artificial agents need to form coordinated teams and maintain their collective behaviour within a spectrum of acceptable behaviours. In such HACs, group decisions are not only required to be normatively acceptable, but may have legal and moral connotations as well. For instance, under time and resource limits, a human-agent rescue team may be forced to put a person at risk to rescue another patient. Developing computational method to support reasoning about sanctionability in such domains requires (1) integrating intentionality models (in particular operational tools to reason about collective intentionality [19]) and (2) eliciting the distribution of preferences/goals in human-agent teams. These elements are key to verifying if a norm is violated intentionally, and accordingly for ascribing liability for collective decisions.

CONCLUSION

As a base for engineering tools to support ethical AI decisions and responsible autonomy in MAS, we presented conceptual requirements for developing operational responsibility reasoning frameworks that are aware of the sociotechnical aspects of this notion in its various forms. Table 1 relates different forms of responsibility in MAS to sociotechnical dynamics. While responsibility is merely a strategic notion (as the strategic ability to avoid or ensure a stat of affairs φ), blameworthiness requires also capturing the epistemic dynamics of MAS and distribution of knowledge among agents. Then, accountability and sanctionability are also concerned with the nature of φ itself and whether it is an allocated task or an established norm, respectively.

Table 1. Different Forms of Responsibility in Relation to Sociotechnical Dynamics of MAS.

	Strategic	Epistemic	Task-oriented	Normative
Responsibility	✓	—	—	—
Blameworthiness	✓	✓	—	—
Accountability	✓	✓	✓	—
Sanctionability	✓	✓	—	✓

We conclude by highlighting the need for addressing responsibility voids, as a related research problem, and discussing the applicability of responsibility reasoning tools for behaviour coordination in MAS and for ensuring their trustworthiness.

Addressing Responsibility Voids in MAS. As shown in previous sections, in a MAS where multiple agents have the potential to influence

the behaviour of the system, it is mostly groups of agents that are responsible, blameworthy, accountable, or sanctionable for a state of affairs. Then the question is how to determine the extent of responsibility of each and every member of such groups. (Note that here we are using the term responsibility to refer to its various forms.) This is known as the problem of responsibility voids/gap [12] referring to situations in which a group is found to be responsible while the extent of each group member's responsibility is not clear. It is desirable to have a method that shares responsibility in a way that is fair to group members, and by fairness we mean that their individual *degree* of responsibility should reflect the contribution of each member to the responsible group. There exist recent approaches to address this problem by applying fairness-ensuring methods from microeconomics, e.g., in the logic-based framework of [15] and causal model of [20]. However, they suffer from intractability issues caused by the high computational complexity of their suggested methods. As a way forward, we envisage the applicability of rule-based methods, with lower computational complexity, for representing game forms in MAS.

Behaviour Coordination in MAS. In the literature on normative coordination, an open problem is to determine effective penalty values in a dynamic manner such that compliance to a set of norms is ensured. Basically, if the preferences of agents changes, penalty/incentive values need to be generated during run-time by taking into account the *norm set* in question, the *preferences* of agents, and the *state* of the MAS (the history of events and potential futures). To capture these features, a graded notion of sanctionability can be applied for incentive engineering in HACs and for determining the penalty/incentive value on the individual level.

Towards Trustworthy Autonomy in MAS. For an effective deployment of trustworthy autonomous systems, computational models to capture and preserve social values and ethical norms play a key role [8]. To that end, responsibility reasoning notions enable verifying if and to what extent agents are responsible for ensuring a social value, are to be blamed for knowingly violating them, or can be seen sanctionable. We ideate formalising these notions and developing responsibility reasoning frameworks using temporal logic-based semantics to capture the strategic, epistemic, and normative aspects of different forms of responsibility. Such frameworks bridge the gap in verifiable responsibility ascription tools for MAS and enable integrating responsibility reasoning into agents' decision making (e.g., to resolve potential conflicts between goals and collective responsibilities). These advancements are crucial to balance agents' individual-level preferences against societal values and, in turn, contribute to the development of responsible and trustworthy autonomy in MAS.

ACKNOWLEDGEMENT

This work was 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 a Turing AI Fellowship (EP/V022067/1).

REFERENCES

1. S. Stein, E. H. Gerding, A. Nedea, A. Rosenfeld, and N. R. Jennings, "Market interfaces for electric vehicle charging," *J. Artif. Intell. Res.*, vol. 59, pp. 175–227, 2017.
2. M. P. Singh, "Norms as a basis for governing sociotechnical systems," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 1, pp. 21:1–21:23, 2013.
3. P. K. Murukannaiah, N. Ajmeri, C. M. Jonker, and M. P. Singh, "New foundations of ethical multiagent systems," in *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '20, Auckland, New Zealand, May 9-13, 2020*, 2020, pp. 1706–1710.
4. N. R. Jennings, "On being responsible," in *Proceedings of the 3rd European Workshop on Modelling Autonomous Agents in a Multi-Agent World*, 1992, pp. 93–102.
5. T. J. Norman and C. Reed, "A logic of delegation," *Artificial Intelligence*, vol. 174, no. 1, pp. 51–71, 2010.
6. Office for Artificial Intelligence - GOV.UK, "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, 2020, accessed: 2021-04-21.

7. A. K. Chopra and M. P. Singh, "The thing itself speaks: Accountability as a foundation for requirements in sociotechnical systems," in *2014 IEEE 7th International Workshop on Requirements Engineering and Law (RELAW)*, 2014, pp. 22–22.
8. European Commission: the High-Level Expert Group on AI, "Ethics guidelines for trustworthy AI," https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai, 2019, accessed: 2021-04-21.
9. V. Dignum, *Responsible Artificial Intelligence - How to Develop and Use AI in a Responsible Way*, ser. Artificial Intelligence: Foundations, Theory, and Algorithms. Springer, 2019.
10. M. Baldoni, C. Baroglio, R. Micalizio, and S. Tedeschi, "Robustness based on accountability in multiagent organizations," in *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, 2021, pp. 142–150.
11. I. van de Poel, "The relation between forward-looking and backward-looking responsibility," in *Moral responsibility*, 2011, pp. 37–52.
12. M. Braham and M. van Hees, "Responsibility voids," *The Philosophical Quarterly*, vol. 61, no. 242, pp. 6–15, 2011.
13. H. Chockler and J. Y. Halpern, "Responsibility and blame: A structural-model approach," *J. Artif. Intell. Res.*, vol. 22, pp. 93–115, 2004.
14. M. Luck, S. Mahmoud, F. Meneguzzi, M. Kollingbaum, T. J. Norman, N. Criado, and M. S. Fagundes, "Normative agents," in *Agreement Technologies*, S. Ossowski, Ed., 2013, pp. 209–220.
15. V. Yazdanpanah, M. Dastani, W. Jamroga, N. Alechina, and B. Logan, "Strategic responsibility under imperfect information," in *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, 2019, pp. 592–600.
16. T. Ågotnes, V. Goranko, W. Jamroga, and M. Wooldridge, "Knowledge and ability," in *Handbook of Epistemic Logic*, 2015, pp. 543–589.
17. V. Yazdanpanah, M. Dastani, S. Fatima, N. R. Jennings, D. M. Yazan, and W. H. Zijm, "Task coordination in multiagent systems," in *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems*, 2020, pp. 2056–2058.
18. N. R. Jennings, L. Moreau, D. Nicholson, S. Ramchurn, S. Roberts, T. Rodden, and A. Rogers, "Human-agent collectives," *Communications of the ACM*, vol. 57, no. 12, pp. 80–88, 2014.
19. M. E. Bratman, *Shared agency: A planning theory of acting together*. Oxford University Press, 2013.
20. M. Friedenberg and J. Y. Halpern, "Blameworthiness in multi-agent settings," in *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*. AAAI, 2019, pp. 525–532.

Vahid Yazdanpanah is a postdoctoral Research Fellow in the Agents, Interaction and Complexity research group in the Department of Electronics and Computer Science at the University of Southampton. He received his PhD from the University of Twente on the application of multiagent techniques for implementing industrial collaborations. His research focus is on intelligent agent technologies and formal methods for multiagent systems.

Enrico H. Gerding is an Associate Professor in the Agents, Interaction and Complexity research group in the Department of Electronics and Computer Science at the University of Southampton. He has been an academic at Southampton since 2007. He received his PhD from the Dutch National Centre of Mathematics and Computer Science (CWI) in 2004 on the topic of automated negotiation. He has over 100 peer-reviewed publications in top conferences, journals and books in the area of artificial intelligence, specifically autonomous agents and multiagent systems.

Sebastian Stein is an Associate Professor within the Agents, Interaction and Complexity research group, which is part of Electronics and Computer Science at the University of Southampton. He completed his PhD in Multiagent Systems at Southampton in 2008 and he is currently a Turing AI Fellow. Sebastian's research is focused on techniques from mechanism design, incentive engineering, and sequential decision making, and their application for solving real-life problems in smart mobility, smart energy, crowd sourcing, and cloud computing.

Corina Cirstea is a Lecturer (Assistant Professor) in the Agents, Interaction and Complexity research group in the Department of Electronics and Computer Science at the University of Southampton. She holds a DPhil in Computation from the University of Oxford (2000). She was the holder of a Junior Research Fellowship in Computer Science at St. John's College Oxford (1999-2003) prior to joining the University of Southampton (2003). Her research interests are in logic and models of computation, more specifically in coalgebras, their close connection to modal logics, and their applications to automated verification.

This article has been accepted for publication in IEEE Internet Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MIC.2021.3107334

m.c. schraefel is a Professor of Computer Science and Human Performance, Fellow of the British Computer Society, and Research Chair for the Royal Academy of Engineering. Her research focuses on the design information systems to support the brain-body connection for quality of life, including fitness to learn, play and perform, and to understand through these paths how to enhance innovation, creativity and discovery. She also directs the WellthLab, whose vision is to help make normal better for all.

Timothy J. Norman is a Professor of Computer Science and Head of the Agents, Interaction and Complexity Group at the University of Southampton. He read Electronic and Electrical Engineering at University of Wales, Swansea, then graduated in 1997 with a Ph.D. in Computer Science from University College London in the area of AI planning and scheduling. After working as a postdoc at Queen Mary University of London, he moved to the University of Aberdeen in 1999 where he was promoted to Professor in 2009. He joined the Agents, Interaction and Complexity Group at Southampton in 2016.

Nicholas R. Jennings is the Vice-Provost for Research and Enterprise and Professor of Artificial Intelligence at Imperial College London. He is an internationally-recognised authority in the areas of AI, autonomous systems, cyber-security and agent-based computing. He is a member of the UK government's AI Council, the governing body of the Engineering and Physical Sciences Research Council, the Monaco Digital Advisory Council, and chair of the Royal Academy of Engineering's Policy Committee.