Collective Responsibility in Multiagent Settings

VAHID YAZDANPANAH, ENRICO H. GERDING, SEBASTIAN STEIN,
CORINA CIRSTEA, m.c. schraefel, and TIMOTHY J. NORMAN, University of Southampton
NICHOLAS R. JENNINGS, Imperial College London

1. INTRODUCTION

Scientific and technological advancements in autonomous agents and multiagent systems offer a promising prospect for more reliable and effective supply chain, transportation, and healthcare systems [Iqbal et al. 2016; Gerding et al. 2011; Chaib-Draa and Müller 2006]. However, at the current stage, developed autonomous systems mainly aim for technical functionality and efficiency but are incapable of reasoning about how and to what extent each agent is responsible and later on is to account for undesirable situations such as a collision among autonomous vehicles. This is mainly because they do not take into account the fact that autonomous systems and artificial intelligence technologies are embedded in a social context with other actors and stakeholders. To foster such an embedding, we follow [Jennings and Mamdani 1992] and argue that the meta-level notion of *collective responsibility* is applicable for coordinating such systems and ensuring their socio-technically desirable behaviour.

In multiagent settings, an open problem is to *determine* and *distinguish* who is responsible, blameworthy, accountable, or sanctionable in a human-agent collective [Jennings et al. 2014; Abeywickrama et al. 2019]. (In the next section, we clarify how these forms of responsibility relate.) The other challenge is on how the collective-level responsibility can be justifiably ascribed to individuals in a collective. In the literature on responsibility reasoning, this is known as the *problem of many hands* [Frankfurt 1969] where we face *responsibility voids* [Braham and van Hees 2011]. This is a class of situations where a collective is responsible for an outcome but there exist no exact method to link the outcome to individuals in the collective, hence ascribing responsibilities is problematic. This lack of collective responsibility reasoning methods, able to capture new forms of agency and autonomy, calls for developing techniques capable of capturing the strategic, temporal, epistemic, and normative aspects of collective responsibility in multiagent settings. Such tools will be a base for ensuring the responsible and trustworthy behaviour of autonomous systems and, in turn, preserving social values key to the successful deployment of human-centred artificial intelligence in society.

Against this background, this work highlights key aspects in various notions of collective responsibility and suggests an approach to employ multiagent temporal logics to formally represent and reason about different forms of responsibility in multiagent settings.

2. NOTIONS OF COLLECTIVE RESPONSIBILITY: CONCEPTUAL ANALYSIS

Responsibility, in its various forms, is a relational concept defined between at least two entities A and φ with the generic structure "*A is responsible for φ*" [van de Poel 2011]. In this structure, A is an agent (collective) 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. Then the nature of φ (e.g., whether it is a norm that agents are expected to adhere to or a task) and also the level of influence of A on φ (e.g., whether A can prevent/provide φ in prospect or could do so in retrospect) defines different forms of responsibility. To discuss different notions of responsibility, imagine a scenario in an autonomous vehicular system with two vehicles red (reaching an intersection on the North-South trajectory) and $blue$ (reaching the same intersection on the East-West trajectory), and pedestrians A and B (aiming to pass the intersection). Note that there exist no traffic light here and agents are expected to reason about their responsibilities and coordinate the situation autonomously.

Responsibility. If we only had two-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 collective (denoted by A) able to ensure that both the vehicles pass the intersection safely with no crashing (denoted by 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 a φ already took place (e.g., that a crash occurred) and one likes to realise if an agent (collective) could avoid 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 agents to possess a different form of ability with respect to φ.

Blameworthiness. 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 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 also crucial [Chockler and Halpern 2004]. Vehicle 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 collective who could avoid the crash is $\{red, blue\}$ only if they had sufficient knowledge. Basically, blameworthiness is an epistemically-bounded form of backward-looking responsibility.

Accountability. While being *responsible* or *blameworthy* for φ merely depends on the strategic and epistemic preconditions that an agent (collective) should satisfy, being *accountable* requires (in addition) the characteristics of φ itself. In principle, backward-looking accountability ascription follows and builds on a task allocation process. This is, collective A is backward-looking accountable if φ remains unfulfilled even though A was able and tasked to to fulfil it. For instance, imagine that red has some goods on board and is tasked to meet a delivery deadline (task φ). If this task remains unfulfilled although red was capable of fulfilling it, we say red is to account for it.

Sanctionability. 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 sanctioning methods in normative systems [Luck et al. 2013; Boella et al. 2006]). In other words, A is sanctionable for φ amounting to sanction value ξ if adhering to (or avoiding) 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 a sanction to violating an expected behaviour in a multiagent setting [Herzig et al. 2011; Dastani et al. 2017]. Imagine a case that according to the history of events $blue$ hits pedestrian A while there exist an established norm φ saying that (1) hitting a pedestrian is "*to be avoided*" and (2) violating φ "*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 for violating φ. 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 as the *avoidance potential* condition [Braham and van Hees 2012]. In this regard, sanctionability is a normative form of backward-looking responsibility as it concerns agents' avoidance potential with respect to the violation of a norm.

3. FORMAL RESPONSIBILITY REASONING IN TEMPORAL LOGICS

A natural approach[1] for developing a responsibility reasoning framework is to use the semantic machinery of temporal multiagent logics, in particular *Concurrent Game Structures (CGS)*[2] [Ågotnes et al. 2015; Alur et al. 2002]. In addition to being expressive for specifying temporal, strategic, normative, and epistemic aspects of various forms of collective responsibility, models that use CGS can benefit from standard model checking platforms to verify responsibilities in the modelled system (e.g., using tools in [Lomuscio et al. 2009; Jamroga and Dix 2005]). As discussed earlier, tools to reason about various forms of responsibility in multiagent settings should (not be limited to individual's responsibility but as well) allow verifying whether a collective of agents is responsible for an event/outcome. Next, we show how our CGS-based formalisation allows verifying and distinguishing various forms of collective responsibility.

In a CGS, let $\varphi \in \Pi$ be a state of affairs, $q \in Q$ an arbitrary state, and $h = q_0, \ldots, q_n$ a chain of materialised states (also known as a q-history if $q = q_n$). We say $\Gamma \subseteq \Sigma$ is forward-looking q-responsible for φ iff (1) Γ possesses an action or a sequence of actions (known as a strategy) in q to ensure φ and (2) no $\Gamma' \subset \Gamma$ can do so. Moreover, $\Gamma \subseteq \Sigma$ is backward-looking q-responsible for φ with respect to q-history h iff (1) φ holds in q but not in any other state of h, (2) Γ has a strategy in a q_i $(0 \leq i < n)$ to avoid φ, and (3) no $\Gamma' \subset \Gamma$ can do so.

To allow formal reasoning about epistemic forms, in particular for modelling blameworthiness, knowledge of agents needs to be represented explicitly. For this, the CGS concept of *"uniform strategy"* [Jamroga and van der Hoek 2004] is applicable as it captures capabilities of agent collectives under imperfect information. A group has a uniform strategy to ensure a formula only if they possess a strategy that is effective under their imperfect information. This way, we say a group is collectively blameworthy for a materialised φ only if they are backward-looking responsible while having a uniform strategy to avoid φ from happening. As discussed, accountability and sanctionability for a situation φ are both concerned with the nature of φ. Accountability is about being responsible for failing to fulfil an allocated task while sanctionability is about being blameworthy for violating a norm. In a CGS, We say $\Gamma \subseteq \Sigma$ is q-accountable for φ with respect to a q-history h iff (1) $\neg\varphi$ holds in q, (2) Γ has a uniform strategy in a q_i $(0 \leq i < n)$ to ensure φ, and (3) Γ was tasked to bring about φ in q. And for being q-sanctionable for φ, $\Gamma \subseteq \Sigma$ needs to be q-blameworthy for φ with respect to h but in addition it is necessary for φ to be normatively undesirable (e.g., determined using a norm-labelling procedure [Gasparini et al. 2016]).

Finally, to address responsibility voids and bridge the gap between collective responsibilities and individual-level responsibility [Yazdanpanah and Dastani 2015], we envisage applying game-theoretical cost-sharing techniques such as the Shapley value or Banzhaf index [Shapley 1953; Banzhaf 1964]. To handle scalability issues, caused by the computational complexity of traditional cost-sharing methods, we ideate the integration of rule-based representations such as *Marginal Contribution Networks* [Ieong and Shoham 2005].

Acknowledgements. 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).

[1] This section reports on a logic-based formalisation concerning how responsibility reasoning contributes to design and development of trustworthy autonomous systems [Yazdanpanah et al. 2021].

[2] Formally, a CGS is a tuple $\mathcal{M} = \langle \Sigma, Q, Act, \Pi, \pi, \sim_1, \ldots, \sim_n, d, o \rangle$ where: $\Sigma = \{a_1, \ldots, a_n\}$ is a set of *agents*; Q is a set of *states*; Act is a set of atomic *actions*; Π a set of atomic propositions; $\pi : \Pi \mapsto 2^Q$ is a propositional evaluation function; $\sim_a \subseteq Q \times Q$ is an *epistemic indistinguishability relation* for each agent $a \in \Sigma$ ($q \sim_a q'$ indicates that q and q' are indistinguishable to a); function $d : \Sigma \times Q \mapsto \mathcal{P}(Act)$ specifies the sets of actions available to agents at each state; and o is a transition function that assigns the outcome state $q' = o(q, \alpha_1, \ldots, \alpha_n)$ to state q and a tuple of actions $\alpha_i \in d(a_i, q)$ that can be executed by Σ in q.

REFERENCES

Dhaminda B. Abeywickrama, Corina Cîrstea, and Sarvapali D. Ramchurn. 2019. Model Checking Human-Agent Collectives for Responsible AI. In *Proceedings of the 28th IEEE International Conference on Robot and Human Interactive Communication*. 1–8.

Thomas Ågotnes, Valentin Goranko, Wojciech Jamroga, and Michael Wooldridge. 2015. Knowledge and ability. In *Handbook of Epistemic Logic*, Hans van Ditmarsch, Joseph Halpern, Wiebe van der Hoek, and Barteld Kooi (Eds.). College Publications, 543–589.

Rajeev Alur, Thomas A. Henzinger, and Orna Kupferman. 2002. Alternating-time temporal logic. *Journal of the ACM* 49, 5 (2002), 672–713.

John F. Banzhaf. 1964. Weighted voting doesn't work: A mathematical analysis. *Rutgers Law Review* 19 (1964), 317.

Guido Boella, Leendert W. N. van der Torre, and Harko Verhagen. 2006. Introduction to normative multiagent systems. *Computational and Mathematical Organization Theory* 12, 2-3 (2006), 71–79.

Matthew Braham and Martin van Hees. 2011. Responsibility voids. *The Philosophical Quarterly* 61, 242 (2011), 6–15.

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

Brahim Chaib-Draa and Jörg Müller. 2006. *Multiagent based supply chain management*. Vol. 28. Springer.

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

Mehdi Dastani, Sebastian Sardiña, and Vahid Yazdanpanah. 2017. Norm enforcement as supervisory control. In *Proceedings of the 20th International Conference on Principles and Practice of Multi-Agent Systems*. 330–348.

Harry G. Frankfurt. 1969. Alternate possibilities and moral responsibility. *The journal of philosophy* 66, 23 (1969), 829–839.

Luca Gasparini, Timothy J. Norman, Martin J. Kollingbaum, Liang Chen, and John-Jules C. Meyer. 2016. CÒIR: Verifying Normative Specifications of Complex Systems. In *Proceedings of the International Workshop on Coordination, Organizations, Institutions, and Norms in Agent Systems*. 134–153.

Enrico H. Gerding, Valentin Robu, Sebastian Stein, David C. Parkes, Alex Rogers, and Nicholas R. Jennings. 2011. Online mechanism design for electric vehicle charging. In *Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems*. 811–818.

Andreas Herzig, Emiliano Lorini, Frédéric Moisan, and Nicolas Troquard. 2011. A dynamic logic of normative systems. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*. 228–233.

Samuel Ieong and Yoav Shoham. 2005. Marginal contribution nets: A compact representation scheme for coalitional games. In *Proceedings of the 6th ACM Conference on Electronic Commerce*. 193–202.

Sajid Iqbal, Wasif Altaf, Muhammad Aslam, Waqar Mahmood, and Muhammad Usman Ghani Khan. 2016. Application of intelligent agents in health-care. *Artificial Intelligence Review* 46, 1 (2016), 83–112.

Wojciech Jamroga and Jürgen Dix. 2005. Model checking strategic abilities of agents under incomplete information. In *Proceedings of the 9th Italian Conference on Theoretical Computer Science*. 295–308.

Wojciech Jamroga and Wiebe van der Hoek. 2004. Agents that know how to play. *Fundamenta Informaticae* 63, 2-3 (2004), 185–219.

Nicholas R. Jennings and Ebrahim H. Mamdani. 1992. Using Joint Responsibility to Coordinate Collaborative Problem Solving in Dynamic Environments. In *Proceedings of the 10th National Conference on Artificial Intelligence, San Jose, CA, USA, July 12-16, 1992*. 269–275.

Nicholas R. Jennings, Luc Moreau, David Nicholson, Sarvapali Ramchurn, Stephen Roberts, Tom Rodden, and Alex Rogers. 2014. Human-agent collectives. *Communications of the ACM* 57, 12 (2014), 80–88.

Alessio Lomuscio, Hongyang Qu, and Franco Raimondi. 2009. MCMAS: A model checker for the verification of multi-agent systems. In *Computer Aided Verification*. 682–688.

Michael Luck, Samhar Mahmoud, Felipe Meneguzzi, Martin Kollingbaum, Timothy J. Norman, Natalia Criado, and Moser Silva Fagundes. 2013. *Normative Agents*. 209–220.

Lloyd S. Shapley. 1953. A value for n-person games. *Contributions to the Theory of Games* 2, 28 (1953), 307–317.

Ibo van de Poel. 2011. The relation between forward-looking and backward-looking responsibility. In *Moral responsibility*. 37–52.

Vahid Yazdanpanah and Mehdi Dastani. 2015. Quantified Degrees of Group Responsibility. In *Proceedings of the International Workshop on Coordination, Organizations, Institutions, and Norms in Agent Systems*. 418–436.

Vahid Yazdanpanah, Enrico H. Gerding, Sebastian Stein, Mehdi Dastani, Catholijn M. Jonker, and Timothy J. Norman. 2021. Responsibility Research for Trustworthy Autonomous Systems. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*. 57–62.