Tackling Challenges of Robustness Measures for Autonomous Agent Collaboration in Open Multi-Agent Systems

David Jin
Karlsruhe Institute of Technology
david.jin@kit.edu

Niclas Kannengießer
Karlsruhe Institute of Technology
niclas.kannengiesser@kit.edu

Benjamin Sturm
Karlsruhe Institute of Technology
benjamin.sturm@kit.edu

Ali Sunyaev
Karlsruhe Institute of Technology
sunyaev@kit.edu

Abstract

Open multi-agent systems (OMASs) allow autonomous agents (AAs) to collaborate in coalitions to accomplish complex tasks (e.g., swarm robots exploring new terrain). In OMASs, AAs can arbitrarily join and leave the network. Thus, AAs must often collaborate with unknown AAs that may corrupt coalitions, leading to less robust systems. However, measures to improve robustness of OMASs are subject to challenges, decreasing their effectiveness. To understand how to improve coalition robustness in OMASs and address challenges of existing robustness measures, we carried out a literature review and revealed three types of robustness measures (i.e., collaboration coordination, normative control, and reliability prediction). Moreover, we found 21 challenges for the identified robustness measures and 24 corresponding solutions. By carrying out this literature review, we forge new connections between existing measures and identify challenges and measures that apply to multiple existing measures. Hereby, our work supports more robust collaborations between AAs in open systems.

1. Introduction

Autonomous agents (AAs) can generate, retrieve, and process data to serve specific purposes, like speech recognition to control smart homes. AAs are essentially software programs that respond to states and events in their environment without immediate instructions (i.e., instructions from their creators or users for every action in response to predefined events) [1]. AAs can collaborate with other AAs in multi-agent systems (MASs) to accomplish tasks that are difficult or even impossible for single AAs [2]. For example, AAs in autonomous driving can alleviate traffic congestions and accelerate transportations when exchanging real-time traffic data with other AAs in the MAS [3]. In such settings, MASs must provide a high level of openness so that even unknown AAs can arbitrarily join and leave the MAS network. MASs with a high level of openness are referred to as open MASs (OMASs) [4].

The number of AAs, their identities, and their functionalities are usually not specified for OMASs and AAs often collaborate ad-hoc in coalitions with other AAs. In this way, OMASs enable very dynamic cooperation between AAs that can be (partially) unknown to each other. The need for AAs to collaborate with agents, whose behavior in coalitions is unknown, poses special challenges to the robustness and success of collaborations. Robustness of collaborative work refers to the extent to which the performance of a system remains stable despite the occurrences of faults or failures. For example, robust coalitions can compensate AAs that crash during collaboration or even sabotage the common effort. Insufficient robustness of coalitions can lead to abandonment of work, failure to achieve goals, and can have detrimental effects, such as traffic accidents [5].

To improve robustness of coalitions in OMASs, various measures have been presented, including reputation systems to indicate the reliability of AAs [4], rule-sets to identify and punish malicious agents [6], and fault-tolerant consensus mechanisms to achieve coalition agreements, even if not all AAs are available [7]. Nevertheless, the presented measures have shortcomings that can decrease their effectiveness. For example, reputation systems accumulate ratings on past actions of AAs to predict their future reliability in collaborative works. In this way, AAs can better decide whether other AAs are reliable enough to be part of coalitions. However, the effectiveness of reputation systems is limited, because a minimum number of ratings on an AA are required before its reliability can be predicted with sufficient accuracy (i.e., cold-start problem [8]).

To address the shortcomings of extant robustness measures, technological advances have been applied, including distributed ledger technology (DLT) [9, 10], semantic web technology [11], and machine learning [12]. For instance, DLT-based reputation systems can store ratings in a tamper-resistant way and make

URI: https://hdl.handle.net/10125/80253
978-0-9981331-5-7

ratings publicly verifiable [13, 14]. Using such technologies has revealed a multitude of potential solutions to address challenges related to robustness measures, for example, using distributed ledger technology (DLT) to store rating on AAs in a tamper-resistant way [13]. To understand how coalition robustness in OMASs can be improved through solutions based on such technologies that address challenges of robustness measures, developers require a thorough understanding of the individual measures and their shortcomings and a thorough understanding of proposed solutions.

Extant works on robustness measures for coalitions in OMASs can be found in three major research fields: social systems, psychology, and computer science. The social systems field compares coalitions of AAs with human coalitions. It introduces the concepts of organizations and institutions to model agent societies. (e.g., [6]). Social systems approaches (e.g., social circles [15]) draw on concepts from law and economics to improve social structures (e.g., [16]). The psychological field is concerned with how AAs can estimate the performance of unknown AAs (e.g., [8]). To give AAs necessary tools, approaches from the field implement human interaction concepts, such as trust and signaling. Approaches associated with the computer science field are concerned with creating infrastructures and solving computational complexity problems (e.g., [17]). Research associated with these research streams is scattered across diverse disciplines and outlets, which poses a barrier to improve robustness of coalitions in OMASs. To improve understanding of the relationships and synergies among disjointed robustness measures and to improve collaboration among AAs in OMAS, existing measures need to be consolidated. To this end, we ask the following research question: *What are the challenges and corresponding solutions of robustness measures for coalitions in OMASs?*

To answer our research question, we reviewed scientific literature on robustness challenges in OMASs. Using thematic analysis [18], we revealed three key robustness measures (i.e., collaboration coordination, normative control, and reliability prediction) to improve robustness of coalitions in OMASs. Linked to the three key measures, we identified 15 subthemes that comprise 21 challenges related to robustness measures and 24 corresponding solutions.

Our work contributes to practice by deepening the understanding of robustness of coalitions in OMASs by presenting a catalog of challenges and corresponding solutions for key robustness measures. Thereby, we support developers in choosing suitable robustness measures for coalitions under consideration of their individual benefits and drawbacks. We contribute to research by presenting consolidated knowledge about the improvement of robustness of coalitions in OMASs. Thereby, this work can serve as a starting point for the targeted improvement of the identified key robustness measures and the investigation of the individual measures' effectiveness for different use cases (e.g., collaborative machine learning).

2. Background

2.1. Autonomous agents

AAs can be characterized as *weak* or *strong*, depending on five properties: internal control, reactivity, persistence, pro-activeness, and goal orientation (see Table I) [19, 20]. Internal control describes the ability of AAs to have authority over their internal state. The internal state of an AA corresponds to the recording of its environment. Reactivity is concerned with the ability of an AA to respond to a stimulus from its environment [20]. Environments of AAs comprise signals (e.g., message receival), conditions (e.g., gravitation and physical rules), and entities (e.g., agents and objects) in their surroundings. AAs can perceive and interact with their environments, for example, by using sensors and effectors [21]. Persistence describes the ability of an AA to constantly operate and autonomously shut itself down.to store its current state at a time of choice and deactivate itself. When the agent is activated again, it resumes working from the stored state.

AAs that have internal control, reactivity, and persistence are considered *weak AAs* [22]. *Strong AAs* have the properties of weak AAs and are also pro-active and goal-oriented. A *pro-active* AA can anticipate a stimulus from the environment before receiving it and can perform actions based on predictions [23]. An AA is *goal-oriented* when it is motivated to perform actions to achieve a particular state of its environment or itself [22].

Challenges, such as data governance and hardware costs, can result in single AAs not having enough data or computing power to solve complex problems on their own. To address this challenge, multiple AAs can collaborate in coalitions. To collaborate, AAs in

Table I. Properties of principal types of AAs in multi-agent systems

Property	Weak AAs	Strong AAs
Internal control	X	X
Reactivity	X	X
Persistence	X	X
Pro-activeness		X
Goal orientation		X
Example	Motion-controlled light	Autonomous cars

coalitions need to be able to communicate with each other (e.g., to negotiate with other AAs).

2.2 Multi-agent systems

Each MAS comprises a set of agents (i.e., its population) and provides an environment to its population with which agents can interact [24]. Properties of the population and the environment mainly differentiate MASs (see Table II). *Population diversity* refers to the presence of different types of AAs. Populations diversity increases with an increasing *heterogeneity* among AAs and decreases with an increasing number of *homogeneous* AAs [25]. Populations become heterogeneous when the comprised AAs have different designs or norms applied.

In MASs, AAs can form multiple coalitions to collaborate. Coalitions can be formed by AAs ad-hoc in a dynamic fashion (e.g., between cars in VANET-based traffic applications) or pre-defined by actors controlling a set of AAs (e.g., in a robotic soccer team). *Coalition control* refers to the organization of AA collaboration in coalitions. AA collaborations can be organized in a *centralized* or *decentralized* way. A centrally organized coalition involves a central AA that determines the current state of the environment (i.e., the current perception of the environment), based on data gathered by the AAs in the coalition. The central AA makes decisions on behalf of the AAs in the coalition. In coalitions with decentralized organization, AAs coordinate themselves without central control (e.g., in peer-to-peer communication) [26].

The *goal structure* in a coalition specifies the number of goals to be achieved by AAs in the coalition. AAs in a coalition can either work together toward a *single goal* or each AA can pursue their own goals [26]. Depending on the goal structure, AAs decide on their cooperativity with other AAs.

Cooperativity describes an AA's level of collaborative engagement to achieve a shared goal. AAs in a coalition can work in a *cooperative* (i.e., AAs help other AAs to achieve goals), *competitive* (i.e., AAs work on their own goals and never collaborate at their expense), or *independent* (i.e., AAs separately work to achieve individual goals) way [24].

The *cardinality* of a population refers to the number of objects in a MAS that can either be affected or perceived by its population [25]. The cardinality of a population can be *finite*, *countable*, or *uncountable*.

Environments provided by MASs to AAs can be *static* or *dynamic*. In static environments, only AAs in the population can change the environment. In dynamic environments, changes can also be caused by objects outside of the population's control.

In MASs with a high degree of *openness*, AAs can arbitrarily join and leave the population. In conventional MASs, the population is closed, and only known AAs can join the population [4]. OMASs can provide a high level of openness to their population by allowing AAs to arbitrarily join and leave the population at any time without informing other AAs.

The environments provided by MASs can respond to actions of AAs in a deterministic or non-deterministic way. *Determinism* of environments is established when identical actions of AAs cause identical responses from the environment under identical conditions. In a *non-deterministic* environment, the action of an AA can cause uncertain responses.

Table II. Types of multi-agent systems

Property	Description	Attribute
Cardinality	The number of objects and AAs that are part of the MAS	Finite
		Countable
		Uncountable
Coalition control	The design of authority in the MAS	Centralized
		Decentralized
Cooperativity	The way how AAs work with other AAs to achieve their goal(s)	Cooperative
		Competitive
		Independent
		Prioritized
Determinism	The environment of an AA responds identically to identical AA actions in identical contexts	Deterministic
		Non-deterministic
Dynamism	Changes in the environment beyond the control of AAs	Static
		Dynamic
Goal structure	The number of goals in the MAS	Single goal
		Multiple goals
Openness	Possibilities to enter and leave the system	Open
		Closed
Population diversity	The presence of different types of AAs	Homogeneous
		Heterogeneous

To make collaborations between AAs more robust, different measures have been presented in prior works (e.g., FIRE [27], AA monitoring [28], and policies for system management [29]). However, presented robustness measures have individual challenges that must be gauged and potentially addressed by developers to achieve high robustness in AA coalitions. To support the development of more robust AA coalitions, a thorough understanding of existing robustness measures as wells as their individual challenges and corresponding solutions is required.

3. Method

To identify measures to improve robustness of coalitions in OMASs, we carried out an extensive literature review, following established guidelines [30, 31]. We first searched for literature in six scientific databases (i.e., ACM Digital Library, AIS Library, EBSCOhost, IEEEXplore, ProQuest, and

ScienceDirect). For the search, we applied the search string *"multi-agent* AND open*"* to the documents' titles, abstracts, and keywords. The search revealed a set of 1,956 documents that were potentially relevant to answer our research question.

Next, we evaluated the relevance of the retrieved documents in two rounds. First, we examined the title, keywords, and abstracts of each document and only kept journal articles and conference papers that were written in English and tackled at least one measure to improve robustness of OMASs. Second, we closer evaluated the relevancy of the remaining documents by reading their full texts, which led to the final selection of 167 relevant documents.

To identify challenges of robustness measures and corresponding solutions, we applied thematic analysis [18] that comprises six phases. In the first phase (familiarize yourself with the data), we examined each documents' meta-data and abstracts and took notes on the focus of each document to gain a general impression of the sample.

In the second phase (generate initial codes), we coded the documents to extract measures that can improve coalition robustness in OMASs, their challenges, and corresponding solutions. We harmonized the coding iteratively to avoid ambiguities and improve exclusiveness for each code. The harmonized codes form a set of 15 subthemes that include multiple challenges of robustness measures and corresponding solutions.

In the third phase (search for themes), we collated the identified subthemes into three preliminary themes (e.g., collaboration coordination). If a subtheme did not match an existing theme, we created a corresponding new theme. For example, we assigned *data scarcity* to the theme *reliability prediction*, while we assigned *coalition construction* to the theme *collaboration coordination*.

In the fourth phase (review themes), we discussed the three preliminary themes and associated subthemes with three colleagues. We refined identified inconsistencies based on the collected feedback. After coding 43 relevant documents, we could not reveal any novel challenges or solutions in the past eight documents. Given the high ratio of eight documents in a set of 43, we are confident that we have reached theoretical saturation at that point and concluded our analysis.

In the fifth phase (define and name themes), we developed concise descriptions for the three themes (see Table III) and 15 subthemes and assigned expressive names to their descriptions.

In the sixth phase (produce the report), we wrote detailed descriptions of each theme and subtheme, as presented in the following result section.

4. Challenges and Solutions to Improve Robustness Measures in Coalitions

Our literature review revealed three themes (see Table III) that represent groups of challenges of robustness measures and corresponding solutions: *collaboration coordination*, *normative control*, and *reliability prediction*. These themes comprise 15 subthemes with a total of 21 challenges of robustness measures and 24 corresponding solutions.

Collaboration coordination. Collaboration coordination refers to the methods and resources involved in the construction of AA coalitions. Inadequate coordination of AA cooperation can prevent coalitions from achieving their goals and even lead to coalition dissolution. We identified the following six subthemes associated with nine challenges and corresponding nine solutions related to collaboration coordination.

Coalition construction: Coalition construction is concerned with the formation of coalitions for successful collaboration. We identified three challenges related to coalition construction: *resource distribution*, *goal development*, and *malicious coalitions*. *Resource distribution* is concerned with the management of limited resources available to coalitions. AAs need to share resources so that their coalition can maximize its payoff. Suboptimal resource usage can reduce the coalition payoff. If the payoff is not worthwhile for AAs, they might leave the coalition, which can lead to the failure of the coalition [32]. *Goal development* refers to negotiations between AAs to agree on goals to be achieved by their coalition. These negotiations are difficult because AAs, especially strong AAs, pursue their individual goals which may oppose those of other AAs. Moreover, AAs can leave coalitions when their individual goals are at risk, thereby hindering successful collaborations. These challenges complicate the design of robust coalitions that attract and keep AAs, while the goals of the coalition and its AAs align [33]. *Malicious coalitions* (i.e., hidden coalitions) include AAs that aim to compromise AAs, coalitions, or entire OMASs. Malicious AAs that work together pose new threats to successful coalitions by having more

Table III. Identified themes of challenges for robustness of coalitions in OMASs

Theme	Description
Collaboration coordination	The methods and resources required to enable the construction of AA coalition for collaboration
Normative control	The processes and resources involved in managing coalitions or populations through rules that restrict their AAs' actions
Reliability prediction	The resources (e.g., computational costs, data, and methods) associated with the creation of predictions about the reliability of an AA

resources and the possibility to coordinate their attacks on the OMASs [34].

To improve *goal development*, social circles can be formed within coalitions [15]. Social circles are communication network structures and are composed of AAs and their mutual social dependencies [35]. Social circles can facilitate information sharing and aim to enable robust long-term collaborations between AAs in coalitions. In addition to social circles, using incentive mechanisms has been proposed to improve goal development [36]. Incentive mechanisms implement policies that specify the rewards (e.g., additional resources) that can be received by AAs, the conditions for the reward payout, and the receivers of the reward. By assigning rewards to actions that favor the accomplishment of goals, the alignment of goals of individual AAs with those of the coalition can be fostered [36]. To hinder *malicious coalitions* from compromising AAs, blocking mechanisms can be implemented [34]. Using blocking mechanisms, AAs must ask the coalition for permission to perform actions. The blocking mechanism simulates the possible results of requested actions. If actions result in an insecure state of the coalition, the mechanism does not permit the execution of the actions and, thus, prevents malicious AAs from harming the coalition.

Organization structure: Hierarchical dependencies between AAs and their engagements in coalitions constitute organizational structures of coalitions. Such organizational structures can be centralized (i.e., a rather strict hierarchy) or decentralized (i.e., a rather flat hierarchy) [37]. We identified the challenge to decide on an appropriate degree of decentralization that achieves sufficient robustness while not violating performance requirements of a coalition.

Developers must gauge the benefits and drawbacks of centralized and decentralized organizational structures. For example, centralized structures usually have less performance overhead (e.g., regarding communication and decision-making [33, 38]) compared to decentralized ones. However, decentralized structures can achieve better censorship resistance and direct communication between AAs [39].

Consensus finding: In coalitions with decentralized organizational structures, AAs need to negotiate with each other to come to agreements (e.g., on the specification of common goals). To this end, consensus mechanisms can be applied [40]. In OMASs, however, the availability of AAs in coalitions at a particular point in time is uncertain because of their high level of openness. Thus, consensus mechanisms that require responses from all AAs in a coalition may fail when indefinitely waiting for responses of currently unavailable AAs (e.g., because they left the coalition).

To find consensus in coalitions, in which AAs can be arbitrarily unreachable, crash-fault tolerant consensus mechanisms like RAFT [41] or Practical Byzantine Fault-Tolerance (PBFT) [42] can be used.

Information flow: Information flow refers to the methods and resources involved in information transfers between AAs. AAs must be able to communicate with each other to form coalitions and to collaborate. Since AAs in OMASs often communicate with unknown AAs, there is an increased risk of data being exposed or compromised by malicious AAs [17]. We identified two challenges regarding information flow: data confidentiality and data integrity. *Data confidentiality* is concerned with the prevention of unauthorized access to data exchanged between AAs [17] *Data integrity* refers to insufficient accuracy of data exchanged between AAs [43].

To improve data confidentiality, AAs can manage data locally and apply access control mechanisms [17]. However, access control mechanisms may not prevent data leaks in information flows. Data leaks can be uncovered by AAs that monitor information flows to identify information leaks [17]. To reduce transfers of confidential data, the compute-to-data paradigm can be applied (e.g., like applied in federated learning [44]). While the data-to-compute paradigm requires AAs (i.e., data donors) to make required data accessible to AAs that perform computations (i.e., data consumer), the compute-to-data paradigm requires data consumers to transfer their computation models to data donors that locally process data and only return corresponding results to the respective data consumers. To improve *data integrity*, digital signatures can be used [43]. To store data in a tamper-resistant way, distributed ledger technology (DLT) has been used in several solutions [45, 46] because DLT enables the operation of highly available, append-only databases (i.e., distributed ledger) [39].

Interaction: Interaction refers to AAs' capabilities and resources that enable interoperable communication with other entities (e.g., AAs). The increasing heterogeneity of populations can increase the likelihood for incompatibilities between AAs. Incompatibilities between AAs can mislead or prevent communication, thereby hindering collaboration coordination [47].

Interoperable knowledge representations can be developed to enable communication between heterogenous AAs [47]. For example, semantic web technologies like ontologies can be applied. In this way, AAs can communicate in different languages [48].

Resilience: Resilience represents the challenge of coalitions to appropriately operate despite the occurrences of faults or complete failures of AAs or components (e.g., the case of AA's death [28]).

AAs that are critical to the functioning of a coalition can be redundantly operated [49] so that replicates of AAs can takeover tasks of suddenly unreachable ones. In this way, coalitions can compensate unreachable AAs if at least one replication of a critical AAs is functioning.

Normative control. Normative control refers to the processes and resources involved in managing coalitions through rules that persuade included AAs to behave according to the coalition's goals. Rules specify the allowed and prohibited actions of AAs. AAs are penalized when they do not adhere to their coalition's active rules. We identified six subthemes associated with eight challenges related to normative control and eight corresponding solutions.

Specification and modification: Norm specification and modification describes the process of creating and modifying rules. Norms specify permitted, prohibited, and obligatory, actions as well as the relations between AAs [50]. We identified two challenges related to norm specification: *dynamic modification* and *reusability*. *Dynamic modification* refers to the refinement of norms and updates of the set of active norms in a coalition (e.g., by adding new norms). Dynamic modification is desirable when coalitions get into an undesired state due to actions of AAs that are not regulated by active norms. The modification of norms mostly requires switching off the coalition, publishing the refined norms, and restarting the coalition. Such restarts interrupt collaborations between AAs and may even dissolve coalitions [51]. *Reusability* refers to the flexibility of norms to be used in different situations [50].

To improve *dynamic modification*, specific action languages can be implemented to allow norms to be changed during run-time [50]. *Reusability* of norms can be improved by defining norms on a higher level of abstraction, similar to the concept of inheritance in object-oriented programming. Specifying norms on different levels of abstraction can increase their customizability and improve software maintainability [51, 52].

Adoption: After norms have been specified and published, they are available to AAs in coalitions that need to adopt these norms. However, norms can be hard to *distribute* due to their complexity and the possibility that AAs may interpret them differently [47]. After norm distribution, AAs need to identify the norms and adopt them. The *identification* of norms can be hindered by the fact that norms are not always locatable for AAs. For example, many OMASs do not have a central database for norms.

To accelerate norm *distribution*, norms should be first distributed to AAs with high network value [47]. Such AAs can be identified through network analysis. To improve norm *identification*, we found one measure that applies machine learning techniques to analyze other AAs' behaviors and infers their inherent norms [12, 52, 53].

Conflict resolution: Collaborations between AAs can be hindered by norms that contradict each other. For example, when a rule prohibits an AA to communicate with unknown AAs and another rule requires the AA to achieve a goal in collaboration with other AAs, the AA may not be able to accomplish its goal. To assure that AAs can achieve their goals, norms must not be contradictory. However, analyses of norms regarding contradictions can cause high *computational costs* that can be difficult to provide [54].

To reduce computational costs for norm analyses, a measure has been developed that creates sets of norms and analyzes the sets to identify conflicting norms. Analyzing several norms at the same time creates sets of non-conflicting norms and reduces computation costs [54].

Monitoring: To identify AAs that violate rules, the actions of AAs in coalitions can be monitored and evaluated considering the set of active norms. We identified two challenges concerning norm monitoring: *responsibility assignment* and *costs*. *Responsibility assignment* refers to the challenge of assigning responsibilities for norm monitoring to AAs in a coalition in a way that. *Costs* refers to covering the monetary efforts caused by monitoring AAs [55].

To motivate AAs to monitor each other, an incentive mechanism can be implemented (e.g., identifying a norm-violating AA will give a reward) [36]. To ensure that monitoring *costs* are covered, an access fee for AAs that enter the coalition can be introduced. The access fee is used to compensate the costs of AAs for monitoring each other [55]. However, if not enough new AAs join the coalition, monitoring costs may not be covered in the long term.

Enforcement: After identifying AAs that violated rules, norms must be enforced, which mostly pertains to punishing those AAs. By enforcing norms, AAs can be incentivized to obey the rules. In OMASs, however, AAs that violate rules are hard to punish because AAs can leave coalitions to *prevent fines* [55]. Too *severe* punishments can discourage AAs from participating in coalitions, especially when potential punishments exceed the benefits for AAs [16].

The *severity* of punishments can be set in relation to the severity of the misbehavior by using retributive justice [16].

Removal: Deprecated norms should be removed from the set of norms to avoid conflicts. To remove a norm, all AAs in a coalition must delete the norm from their local norm set. However, norms can be deeply engrained into AAs, and removing such norms

requires a collective belief-change of AAs in coalitions toward the norm to be removed [56].

To remove norms a five-step process can be applied [56]. First, AAs must start to recognize that a specific norm causes negative consequences. Second, a collective belief-change in the AAs coalition must occur. The beliefs of AAs start to change toward the norm that should be removed. Third, the AAs in a coalition collectively decide on whether to remove the norm. Fourth, if the coalition decided to remove the norm, an authority is chosen to impose sanctions on AAs that still apply the removed norm to be removed. Fifth, AAs in the coalition remove the norm from their local set of norms.

Reliability prediction. Reliability prediction is a measure to improve robustness of coalitions by enabling AAs to estimate the reliability of AAs regarding collaborations. In this work, reliability refers to the likelihood to which the actions of an AA will comply with the successful achievement of coalition goals. To increase the probability of successfully achieving goals, AAs can decline collaborations with potentially unreliable AAs. We identified three subthemes associated with four challenges and seven corresponding solutions related to reliability prediction in OMASs.

Prediction: Challenges related to prediction refer to methods and resources involved in estimating the reliability of AAs to avoid unsuccessful collaborations. *Accuracy* refers to challenges related to finding data and measures to maximize the accuracy of reliability estimations. *Identity management* refers to challenges related to the creation of identities for AAs and the management of associated information. Such Information are required to predict the AAs' reliability. Existing prediction approaches mostly use ratings associated with the identities of AAs. The identities and associated ratings are often stored in databases governed by central parties. Thereby, stored ratings are often prone to manipulations. Manipulated ratings reduce the performance of prediction approaches (e.g., in terms of accuracy), thereby decreasing their effectiveness for estimating the reliabilities of AAs.

We identified three measures that can improve *accuracy*: direct prediction approaches, indirect prediction approaches, and hybrid prediction approaches. In direct prediction approaches, AAs predict the reliability of other AAs based on their own experiences (e.g., past interactions). In *indirect prediction approaches* (e.g., reputation systems [57]), AAs acquire such experiences from other AAs. Direct and indirect approaches can also be combined in hybrid approaches (e.g., in FIRE [27]). To address challenges related to *identity management* we identified one measure. To achieve tamper-resistance of stored ratings and ensure that the ratings are associated with the correct AA identities DLT-based approaches have been found promising [58]. Because DLT-based reputation systems provide a digital platform for the reliable computation and storage of direct and indirect predictions of AAs [58, 59].

Data Scarcity: When no or only a few ratings associated with an AA are available, meaningful reliability predictions for that AA often cannot be computed [8]. The insufficient availability of data for such predictions refers to the challenge of *data scarcity*.

We identified two measures to address data scarcity: *archetyping* (e.g., [8]) and *collaterals* (e.g., [59]). In *archetyping*, AAs classify their experiences with past collaborators into archetypes of AAs. To estimate the behavior of an unknown AA, AAs can use their previously learned archetypes. Thereby, AAs need less data about other AAs to predict their behaviors. The measure *collaterals* requires AAs to deposit a specified collateral prior to joining a coalition, for example, an amount of coins of a cryptocurrency [14]. If an AA in the coalition turns out to be not reliable, its collateral can be distributed to other AAs in the coalition or destroyed [59].

Data believability: Data believability is concerned with the extent to which AAs can be sure that received data is credible. In large-scale OMASs, some AAs may never interact with each other. To however predict the reliability of AAs, AAs can share information about their previous interactions (e.g., for indirect prediction approaches). Nonetheless, information about other AAs might be wrong or biased [45], ultimately decreasing the accuracy of reliability predictions.

To improve data believability, incredible information (e.g., sabotaged witness reports or inaccurate ratings) can be detected and filtered using similarity measures between witness reports. With increasing similarity between different witness reports, the credibility of the reports rises [60].

5. Discussion and conclusion

In this work, we present three key measures that can improve robustness of AA collaboration in OMASs: *collaboration coordination*, *normative control*, and *reliability prediction*. We revealed that the three key measures are subject to 21 challenges, ranging from *data scarcity* to *norm enforcement*. To improve the effectiveness of the identified key measure by addressing the identified challenges, we presented 24 solutions.

As OMASs have a high degree of openness, OMASs can achieve a high degree of decentralization, independence from central parties that orchestrate the system, freedom in forming coalitions, and offers high

flexibility for the formation of coalitions. Despite the benefits of OMASs, our results show that reliability predictions for actions of unknown AAs are a major challenge for the robustness measures. Less dominant challenges pertain to data security (e.g., confidentiality and integrity) and data scarcity.

Due to the high frequency of published research and the number of different measures developed, we believe that the reliable identification of malicious AAs is the most pressing challenge that needs to be addressed to improve coalition robustness in OMASs. Follow-up measures, such as excluding malicious AAs from coalitions, require the previous identification of malicious AAs.

Our study indicates that DLT is promising to improve robustness of coalitions, for example, the provision of tamper-resistant storage for ratings in reputation systems and improved scalability regarding increasing or decreasing requests for ratings stored in the distributed ledger [14]. However, several DLT characteristics contradict the requirements of OMASs. For example, the public visibility of data stored on public distributed ledgers can be in conflict with data confidentiality requirements of coalitions [53]. Moreover, the fees charged by many public distributed ledgers (e.g., Bitcoin and Ethereum) for transaction processing can strongly increase costs to be paid by AAs. Especially, when AAs locally operate DLT clients, the energy consumption of these AAs will at least slightly increase [14]. Since our results show that the distribution of operational costs (e.g., for monitoring AAs) is a challenge, system designers must consider an increase in costs caused by using DLT.

Our work contributes to practice by deepening the understanding of robustness measures to improve collaboration in coalitions in OMASs. Furthermore, we support developers in choosing suitable robustness measures for coalitions by identifying potential shortcomings of existing robustness measures and providing knowledge on how to address them. We contribute to research by supporting the understanding of key robustness measures that can improve collaboration in OMASs. By providing a consolidated knowledge base for challenges and corresponding solutions, we uncover the shortcomings of existing robustness approaches. By deepening the understanding of current challenges of robustness measures and corresponding solutions, this work lays a cornerstone for the investigation of effects of identified solutions on the robustness coalitions and the corresponding OMAS.

Our work has limitations that are caused by the applied qualitative study design. We chose a broad search string to gather relevant literature to answer our research problem. However, we focused on literature on OMASs and neglected research fields that may also support us in answering our research question. We iteratively discussed and refined the coding to avoid biases. However, we cannot guarantee to have fully prevented bias in the literature analysis. Moreover, we assume theoretical saturation after having coded 43 documents and cannot guarantee completeness of our results.

The presented robustness measures and solutions to improve their drawbacks should be further investigated regarding the possibilities to quantify their effects on the robustness of coalitions. Such quantifications can improve the comparability of the identified measures and solutions to address challenges of robustness measures. Building on an empirical comparison of robustness measures and solutions, use case-specific recommendations for the use of particular measures and solutions can be made, taking into account the contexts in which they are used, thereby improving the robustness of coalitions.

Acknowledgment

This work was performed in the scope of the project SPECK and sponsored by the German Federal Office for Agriculture and Food (Project code: 281A502B19). This work was supported by KASTEL Security Research Labs.

References

[1] Bösser, T., "Autonomous Agents", in International Encyclopedia of the Social & Behavioral Sciences. 2001. Pergamon: Oxford.

[2] Wooldridge, M.J., An introduction to multiagent systems, 2nd edn., John Wiley & Sons, Chichester, UK, 2009.

[3] Lin, K., R. Zhao, Z. Xu, and J. Zhou, "Efficient Large-Scale Fleet Management via Multi-Agent Deep Reinforcement Learning", in Proceedings of the 24th ACM International Conference on Knowledge Discovery & Data Mining, London, UK. 2018.

[4] Huynh, T.D., N.R. Jennings, and N.R. Shadbolt, "An integrated trust and reputation model for open multi-agent systems", Journal of Autonomous Agents and Multiagent Systems, 13(2), 2006, pp. 119–154.

[5] Balaji, P.G., G. Sachdeva, D. Srinivasan, and C.-K. Tham, "Multi-agent System based Urban Traffic Management", in Proceedings of the 2007 IEEE Congress on Evolutionary Computation, Singapore. 2007.

[6] Criado, N., "Using norms to control open multi-agent systems", AI Communications, 26(3), 2013, pp. 317–318.

[7] Abdelrahim, M., J.M. Hendrickx, and W. Heemels, "MAX-consensus in open multi-agent systems with gossip interactions", in Proceedings of the 56th Annual

Conference on Decision and Control, Melbournce, Australia. 2017.

[8] Burnett, C., N. Timothy J., and K. Sycara, "Bootstrapping Trust Evaluations through Stereotypes", in Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems, Toronto, Canada. 2010.

[9] Tenorio-Fornés, A., S. Hassan, and J. Pavón, "Open Peer-to-Peer Systems over Blockchain and IPFS", in Proceedings of the 1st Workshop on Cryptocurrencies and Blockchains for Distributed Systems, Munich, Germany. 2018.

[10] Skowroński, R., "The open blockchain-aided multi-agent symbiotic cyber–physical systems", Future Generation Computer Systems, 94(4), 2019, pp. 430–443.

[11] Ciortea, A., S. Mayer, F. Gandon, O. Boissier, A. Ricci, and A. Zimmermann, "A Decade in Hindsight: The Missing Bridge Between Multi-Agent Systems and the World Wide Web", in Proceedings of the International Conference on Autonomous Agents and Multiagent Systems, Montreal, Canada. 2019.

[12] Serrano, E. and J. Bajo, "Discovering Hidden Mental States in Open Multi-Agent Systems by Leveraging Multi-Protocol Regularities with Machine Learning", Sensors, 20(18), 2020, 5198.

[13] Calvaresi, D., V. Mattioli, A. Dubovitskaya, A.F. Dragoni, and M. Schumacher, "Reputation Management in Multi-Agent Systems Using Permissioned Blockchain Technology", in Proceedings of the 2018 International Conference on Web Intelligence, Santiago, Chile. 2018.

[14] Lücking, M., F. Kretzer, N. Kannengießer, M. Beigl, A. Sunyaev, and W. Stork, "When Data Fly: An Open Data Trading System in Vehicular Ad Hoc Networks", Electronics, 10(6), 2021, p. 654.

[15] Golpayegani, F., I. Dusparic, and S. Clarke, "Using Social Dependence to Enable Neighbourly Behaviour in Open Multi-Agent Systems", ACM Transactions on Intelligent Systems and Technology, 10(3), 2019, pp. 1–31.

[16] Zolotas, M. and J. Pitt, "Self-Organising Error Detection and Correction in Open Multi-agent Systems", in Proceedings of the 1st International Workshops on Foundations and Applications of Self Systems, Augsburg, Germany. 2016.

[17] Bijani, S. and D. Robertson, "A review of attacks and security approaches in open multi-agent systems", Artificial Intelligence Review, 2012(42), pp. 607–636.

[18] Braun, V. and V. Clarke, "Thematic analysis", in APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological, H. Cooper, Editor. 2012. American Psychological Association: Washington.

[19] Shehory, O., Architectural Properties of Multi-Agent Systems, Pittsburgh, PA, USA, 1999.

[20] Franklin, S. and Graesser Art, "Is it an Agent, or just a Program? A Taxonomy of Autonomous Agents", in Proceedings of the Workshop on Intelligent Agents III, Agent Theories, Architectures, and Languages. 1996: Memphis, Tenessee, USA.

[21] Russell, S.J. and P. Norvig, Artificial Intelligence: A Modern Approach, 3rd edn., Prentice Hall Press, USA, 2009.

[22] Tosic, P.T. and G.A. Agha, "Towards a hierarchical taxonomy of autonomous agents", in Proceedings of the international conference on systems, man & cybernetics theme, The Hague, Netherlands. 2004.

[23] Sinapayen, L., A. Masumori, and T. Ikegami, "Reactive, Proactive, and Inductive Agents: An Evolutionary Path for Biological and Artificial Spiking Networks", Frontiers in computational neuroscience, 13, 2019, pp. 88–102.

[24] Ferber, J., Multi-agent systems: An introduction to distributed artificial intelligence, Addison-Wesley, Harlow, UK, 2005.

[25] Moya, L.J. and A. Tolk, "Towards a Taxonomy of Agents and Multi-Agent Systems", in Proceedings of the Spring Simulation Multiconference, Norfolk, Virginia, USA. 2007.

[26] Ponomarevs, S. and A.E. Voronkov, Multi-agent systems and decentralized artificial superintelligence, ArXiv, 2017 arxiv.org/ftp/arxiv/papers/1702/1702.08529.pdf.

[27] Huynh, T.D., N.R. Jennings, and N.R. Shadbolt, "FIRE: An Integrated Trust and Reputation Model for Open Multi-Agent Systems", Journal of Autonomous Agents and Multiagent Systems, 13(13), 2004, pp. 18–22.

[28] Klein, M., J.-A. Rodriguez-Aguilar, and C. Dellarocas, "Using Domain-Independent Exception Handling Services to Enable Robust Open Multi-Agent Systems: The Case of Agent Death", Journal of Autonomous Agents and Multiagent Systems, 7(1/2), 2003, pp. 179–189.

[29] Pitt, J., D. Busquets, and R. Riveret, "The pursuit of computational justice in open systems", AI & SOCIETY, 30(3), 2015, pp. 359–378.

[30] Kitchenham, B., O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering – A systematic literature review", Information and Software Technology, 51(1), 2009, pp. 7–15.

[31] Kitchenham, B. and S. Charters, Guidelines for performing Systematic Literature Reviews in Software Engineering, 2007.

[32] Shah, N., K.-M. Chao, A.N. Godwin, and A. James, "An abstract knowledge based approach to diagnosis and recovery of plan failure in multi-agent systems", Advanced Engineering Informatics, 21(2), 2007, pp. 183–190.

[33] Houhamdi, Z. and B. Athamena, "Collaborative Team Construction in Open Multi-Agents System", in Proceedings of the 2020 International Arab Conference on Information Technology.

[34] Cristani, M., E. Karafili, and L. Viganò, Blocking Underhand Attacks by Hidden Coalitions, ArXiv, 2010 https://arxiv.org/pdf/1010.4786.

[35] Tong, X., H. Huang, and W. Zhang, "Agent long-term coalition credit", Expert Systems with Applications, 36(5), 2009, pp. 9457–9465.

[36] Centeno, R., H. Billhardt, and R. Hermoso, "Persuading agents to act in the right way: An incentive-based approach", Engineering Applications of Artificial Intelligence, 26(1), 2013, pp. 198–210.

[37] Sunyaev, A., N. Kannengießer, R. Beck, H. Treiblmaier, M. Lacity, J. Kranz, G. Fridgen, U. Spankowski, and A. Luckow, "Token Economy", Business & Information Systems Engineering, 63(4), 2021, pp. 457–478.

[38] Vercouter, L., "A fault-tolerant open MAS", in Proceedings of the 1st International Joint Conference on Autonomous Agents and Multiagent Systems, Bologna, Italy. 2002.

[39] Kannengießer, N., S. Lins, T. Dehling, and A. Sunyaev, "Trade-Offs between Distributed Ledger Technology Characteristics", ACM Computing Surveys, 53(2), 2020, pp. 1–37.

[40] Franceschelli, M. and P. Frasca, "Proportional Dynamic Consensus in Open Multi-Agent Systems", in Proceedings of the 2018 Conference on Decision and Control, Miami Beach, FL, USA. 2018.

[41] Ongaro, D. and J. Ousterhout, "In Search of an Understandable Consensus Algorithm", in Proceedings of the USENIX Annual Technical Conference, Philadelphia, Pennsylvania, USA. 2014.

[42] Castro, M. and B. Liskov, "Practical Byzantine Fault Tolerance", in Proceedings of the 3rd Symposium on Operating Systems Design and Implementation, New Orleans, Louisiana, USA. 1999.

[43] Skowroński, R., "The open blockchain-aided multi-agent symbiotic cyber–physical systems", Future Generation Computer Systems, 94(4), 2019, pp. 430–443.

[44] Zhang, Z., T. Yang, and Y. Liu, "SABlockFL: a blockchain-based smart agent system architecture and its application in federated learning", International Journal of Crowd Science, 4(2), 2020, pp. 133–147.

[45] da Silva, V.T., F. Duran, J. Guedes, and C.J.P. de Lucena, "Governing multi-agent systems", Journal of the Brazilian Computer Society, 13(2), 2007, pp. 19–34.

[46] Bijani, S., D. Robertson, and D. Aspinall, "Secure information sharing in social agent interactions using information flow analysis", Engineering Applications of Artificial Intelligence, 70(1), 2018, pp. 52–66.

[47] Franks, H., N. Griffiths, and S.S. Anand, "Learning Influence in Complex Social Networks", in Proceedings of the 2013 International Conference on Autonomous Agents and Multi-Agent Systems. 2013: Richland, SC, USA.

[48] Ciortea, A., O. Boissier, A. Zimmermann, and A.M. Florea, "Give Agents Some REST: A Resource-Oriented Abstraction Layer for Internet-Scale Agent Environments", in Proceedings of the 16th Conference on Autonomous Agents and Multiagent Systems. 2017: Richland, SC, USA.

[49] Ductor, S. and Z. Guessoum, "A coordination mechanism to replicate large-scale multi-agent systems", in Proceedings of the 13th International Conference on Software Engineering for Adaptive and Self-Managing Systems, Gothenburg Sweden. 2018.

[50] Artikis, A., "Formalising dynamic protocols for open agent systems", in Proceedings of the 12th International Conference on Artificial Intelligence and Law, Barcelona, Spain. 2009.

[51] Carvalho, G., C. Lucena, R. Paes, and J.-P. Briot, "Refinement operators to facilitate the reuse of interaction laws in open multi-agent systems", in Proceedings of the 2006 International Workshop on Software Engineering for Large-scale Multiagent Systems, Shanghai, China. 2006.

[52] Chocron, P. and M. Schorlemmer, "Inferring Commitment Semantics in Multi-Agent Interactions", in Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems, Stockholm, Sweden. 2018.

[53] Savarimuthu, B.T.R., S. Cranefield, M. Purvis, and M. Purvis, "A Data Mining Approach to Identify Obligation Norms in Agent Societies", in Agents and Data Mining Interaction. 2010. Springer Berlin Heidelberg.

[54] Silvestre, E.A. and V.T. da Silva, "Conflict Detection among Multiple Norms in Multi-Agent Systems", Applied Artificial Intelligence, 32(4), 2018, pp. 388–418.

[55] Alechina, N., J.Y. Halpern, I.A. Kash, and B. Logan, "Decentralised Norm Monitoring in Open Multi-Agent Systems", in Proceedings of the 2016 International Conference on Autonomous Agents Multiagent Systems, Singapore, Singapore. 2016.

[56] Hammoud, M., A. Ahmad, A.Y. Tang, and M.S. Ahmad, "Modeling norms removal in open normative multi-agent system", in Proceedings of the 2019 International Conference on Computational Science and Technology, Kota Kinabalu, Malaysia. 2019.

[57] Aref, A. and T. Tran, "Using Fuzzy Logic and Q-Learning for Trust Modeling in Multi-agent Systems", in Proceedings of the 2019 Federated Conference on Computer Science and Information Systems, Leipzig, Germany. 2019.

[58] Bellini, E., Y. Iraqi, and E. Damiani, "Blockchain-Based Distributed Trust and Reputation Management Systems: A Survey", IEEE Access, 8, 2020, pp. 21127–21151.

[59] Harz, D. and M. Boman, "The Scalability of Trustless Trust", in Financial Cryptography and Data Security. 2019. Springer Berlin Heidelberg: Berlin Heidelberg.

[60] Xing, M., B. Li, and X. Wu, "Detecting Malicious Witness Reports in Multi-agent Systems", in Proceedings of the International Conference on Computational Intelligence and Security Workshops, Harbin, Heilongjiang, China. 2007.