Accountable Agents and Where to Find Them

Stefano Tedeschi
Università degli Studi di Torino — Dipartimento di Informatica
Torino, TO, Italy
stefano.tedeschi@unito.it

ABSTRACT

The aim of my PhD is to investigate the notion of *computational accountability* relying on approaches from the research on multi-agent systems. The main contribution will be to provide a notion of when an organization supports accountability, by exploring the process of construction of the organization itself, and to guarantee accountability as a design property.

CCS CONCEPTS

• **Computing methodologies** → **Distributed artificial intelligence**; **Multi-agent systems**; *Cooperation and coordination*;

KEYWORDS

Computational Accountability; Multi-Agent Systems; Social Commitments

ACM Reference Format:
Stefano Tedeschi. 2018. Accountable Agents and Where to Find Them. In *2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18), February 2–3, 2018, New Orleans, LA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3278721.3278783

1 CONTEXT

One might see accountability as the assumption of responsibility for decisions and actions that an individual or an organization has towards another party; it is the process by means of which principals must account for their behavior when put under examination. The concept provides a mechanism by which entities constrain one another's behavior.

In human societies, in turn, *organizations* embody a powerful way to coordinate a complex behavior of many autonomous individuals. More and more often, organizations (including companies) voluntarily adopt accountability frameworks as a way to obtain feedbacks that are useful to evaluate and possibly improve the processes they put in place, as well as their own structure [11, 12]. Modern organizations are supported in their work by software systems, that connect offices and individuals with resources and services. Such software systems together with the involved principals constitute *socio-technical systems*. In general a socio-technical system will involve autonomous and heterogeneous actors, both human and artificial ones, operating and interacting in a dynamic and distributed environment. Unfortunately, current socio-technical systems do not provide any support to the realization of accountability frameworks.

Accountability determination, indeed, is an extremely complex task, which is strongly related to the socio-cultural context in which it takes place. The examination of such a context is usually carried out by a *forum* of auditors. Moreover, in a complex system encompassing interacting parties, like those described above, the most significant cause of a given outcome may not stem from the last agent who produced a change in the result. Instead, there could be more intricate chains of actions which leaded to the final outcome.

Wishing to voice my own contribution, the aim of my PhD will be to investigate the notion of *computational accountability* [1, 3] in software systems, especially in multi-agent organizations, in order to develop a sound and complete conceptual framework and programming platform for it. By the term computational accountability I mean the realization via software of the abilities to trace, evaluate, and communicate accountability, a currently open challenge, that could be successfully faced with the support of intelligent systems. In particular, the problem will be addressed (i) by supplying a definition of computational accountability and (ii) by providing both modeling and computational tools that can be used to realize enterprise software infrastructures which support the realization of accountability frameworks.

In Artificial Intelligence (AI), *multi-agent approaches* to programming proved to be effective in the implementation and management of socio-technical systems (like those I have described), and they provide a promising basis for the development of a platform for computational accountability. Among the reasons, there is the fact that they enable to explicitly represent the interaction as well as the social relationships among the agents, which in turn allow to reason about expectations on the agents' behavior. Actually, computational accountability offers an example of how AI and ethics may interact, since it concerns the traceability, evaluation, and communication of values and good conduct, to support the interacting parties, and to help solve disputes.

Different research communities have dealt with the topic of accountability in software systems. Chopra and Singh, for instance, see accountability as an explicitly established context-specific relationship between two parties identified as account-giver and account-taker [9]. Burgermeestre and Hulstijin, in turn, focus on the entire process of accountability determination, from the establishment of relationships between different stakeholders to the investigation, discussion and evaluation of every possible relationship violation [8]. Nevertheless, a model of accountability and of how accountability relations are created and evolve is still missing.

2 ORGANIZATIONAL ACCOUNTABILITY

The approach followed in my research project consists in further developing the programming technique presented in my M. Sc.

AIES '18, February 2–3, 2018, New Orleans, LA, USA

ACM ISBN 978-1-4503-6012-8/18/02.
https://doi.org/10.1145/3278721.3278783

thesis, ADOPT. The acronym stands for Accountability-Driven Organization Programming Technique and involves the investigation of the process for the construction of an organization of agents. The first steps in this direction [4, 5, 7] concerned the development of a methodology to obtain accountability *as a design property* by relying on the same key notions that are used in defining an organization, namely the ones of *role* an agent plays, and *goals* associated with this role. The process consists of making explicit the *accountability requirements* associated with roles, which the role players should satisfy. The main intuition, here, is that, when an agent wants to play a role in an organization, it has to explicitly accept all the accountability requirements associated with the role itself, expressed as *social commitments* [10]. A social commitment $C(x, y, p, q)$ models the directed relation between two agents, a debtor x and a creditor y. The debtor commits to its creditor to bring about the consequent condition q when the antecedent p holds. In our case, debtor and creditor may amount to role players and organization, while commitment conditions will concern goals associated to roles. An agent can be considered as accountable for a goal only if it explicitly accepted it (with a commitment), possibly providing provisions, i.e. conditions under which it declares to be able to achieve the goal. After the instantiation of these commitments, the organization will be in condition to assign goals to the agents playing the various roles. If this happens, the agents become obliged to achieve the goals, provided that the related provisions hold, lest the violation of the accountability requirements.

Another point concerned the definition of an actual protocol to be followed in order to inherently design and build accountability-supporting organizations. The protocol regulates the process of enrollment of an agent inside an organization. It specifies the shape of the previously mentioned commitments and controls their creation. The gist, here, is that commitments allow to realize a relational representation of interaction, where agents directly create normative bonds with one another and use them to coordinate their activities. These bonds can be, then, inspected and used to discern who is accountable for what when an expectation is violated.

At the same time, in [6], an information model that describes which data should be available, together with their relationships, in order to identify accountabilities in a group of interacting parties has been proposed. The model refines the characterization of accountability traced with ADOPT, by identifying the main concepts that come into play in the process of accountability determination, such as *mutually held expectation* and *control*, together with the relationships among them.

3 IMPACT AND FUTURE DIRECTIONS

Agents in a social state can influence the environment and the agents around them. Furthermore, in a complex system, the most significant cause of an outcome may not stem from the last agent who affected change on the result. Social reasoning, in turn, permits the exposure of more convoluted causes of a certain outcome.

Business ethics and compliance programs are becoming more and more central, bringing consequently to the forefront the importance of accountability. Individuals have to be held accountable for their (mis)behavior and, therefore, provide feedback about the reasons of performance. An accountability platform could support this process in a transparent and automated way, with plenty of potential applications in such diverse fields as finance, (human-resource) management, corruption detection, public administration, research, and decision support. The main purpose of my work is, then, to build a system able to support and facilitate the application of a concept like accountability to modern computer systems. It is worth noting that such a system could be the foundation for the realization of other ethical principles and values, such as transparency, privacy, data protection, and so on.

Future work will mainly follow three directions. The first includes a further refinement of the accountability protocol introduced in the previous section. This is related to the formalization of a conceptual model for organizational accountability, to clarify the concepts which come into play when dealing with accountability in an organizational setting. Second, it would be interesting to integrate the protocol into JaCaMo+ [2], a commitment-based infrastructure for programming MASs, and to build a concrete accountability infrastructure and monitor for compliance. Finally, it seems extremely challenging to investigate the notion of accountability in more complex settings, such as open systems where agents can enter and exit at any time or systems with competitive agents.

REFERENCES

[1] Matteo Baldoni, Cristina Baroglio, Olivier Boissier, Katherine M. May, Roberto Micalizio, and Stefano Tedeschi. 2018. Accountability and Responsibility in Agent Organizations. In *Proceedings of PRIMA 2018 (Lecture Notes in Computer Science)*. Springer. To appear.

[2] Matteo Baldoni, Cristina Baroglio, Federico Capuzzimati, and Roberto Micalizio. 2015. Programming with commitments and goals in JaCaMo+. In *Proceedings of AAMAS 2015*. International Foundation for Autonomous Agents and Multiagent Systems, 1705–1706.

[3] Matteo Baldoni, Cristina Baroglio, Katherine M. May, Roberto Micalizio, and Stefano Tedeschi. 2016. Computational Accountability. In *Proceedings of the AI*IA Workshop URANIA 2016 (CEUR Workshop Proceedings)*, Federico Chesani, Paola Mello, and Michela Milano (Eds.). Aachen, 56–62. http://ceur-ws.org/Vol-1802/paper8.pdf

[4] Matteo Baldoni, Cristina Baroglio, Katherine M. May, Roberto Micalizio, and Stefano Tedeschi. 2017. ADOPT JaCaMo: Accountability-Driven Organization Programming Technique for JaCaMo. In *Proceedings of PRIMA 2017 (Lecture Notes in Computer Science)*, Bo An, Ana L. C. Bazzan, João Leite, Serena Villata, and Leendert W. N. van der Torre (Eds.), Vol. 10621. Springer, 295–312. https://doi.org/10.1007/978-3-319-69131-2_18

[5] Matteo Baldoni, Cristina Baroglio, Katherine M. May, Roberto Micalizio, and Stefano Tedeschi. 2017. Supporting Organizational Accountability Inside Multiagent Systems. In *Proceedings of AI*IA 2017 (Lecture Notes in Computer Science)*, Floriana Esposito, Roberto Basili, Stefano Ferilli, and Francesca A. Lisi (Eds.), Vol. 10640. Springer, 403–417. https://doi.org/10.1007/978-3-319-70169-1_30

[6] Matteo Baldoni, Cristina Baroglio, Katherine M. May, Roberto Micalizio, and Stefano Tedeschi. 2018. An Information Model for Computing Accountabilities. In *Proceedings of AI*IA 2018*, Chiara Ghedini, Bernardo Magnini, Andrea Passerini, and Paolo Traverso (Eds.). Springer, Trento, Italy. To appear.

[7] Matteo Baldoni, Cristina Baroglio, Katherine M. May, Roberto Micalizio, and Stefano Tedeschi. 2018. Computational Accountability in MAS Organizations with ADOPT. *Applied Sciences* 8, 4 (2018). https://doi.org/10.3390/app8040489

[8] Brigitte Burgemeestre and Joris Hulstijn. 2015. *Handbook of Ethics, Values, and Technological Design: Sources, theory, values and application domains*. Springer, Chapter Designing for Accountability and Transparency: A value-based argumentation approach.

[9] Amit K. Chopra and Munindar P. Singh. 2014. The thing itself speaks: Accountability as a foundation for requirements in sociotechnical systems. In *2014 IEEE 7th International Workshop on Requirements Engineering and Law*. 22–22.

[10] Munindar P. Singh. 1999. An ontology for commitments in multiagent systems:. *Artificial Intelligence and Law* 7, 1 (1999), 97–113.

[11] United Nations Children's Fund. 2009. Report on the accountability system of UNICEF. https://www.unicef.org/about/execboard/files/0915accountabilityODS-English.pdf. E/ICEF/2009/15.

[12] Mounir Zahran. 2011. Accountability Frameworks in the United Nations System. https://www.unjiu.org/en/reports-notes/JIU%20Products/JIU_REP_2011_5_English.pdf. UN Report.