Accountable Agents and Where to Find Them

Stefano Tedeschi
University of Torino
Computer Science Department
Via Pessinetto, 12, 10149, Torino, Italy
tedeschi@di.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 guarantee accountability as a design property.

Keywords: AI for Social Good, Computational Accountability, Multi-Agent Systems, Social Commitments.

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. In human societies, *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 (Zahran 2011; United Nations Children's Fund 2009). 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.

Copyright © 2018, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

The aim of my PhD will be to investigate the notion of *computational accountability* (Baldoni et al. 2016) 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, 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 (Chopra and Singh 2014). 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 (Burgemeestre and Hulstijn 2015). Nevertheless, a model of accountability and of how accountability relations are created and evolve is still missing.

Organizational accountability

The approach followed in my research project consists in further developing the programming technique presented in my M. Sc. 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 (Baldoni et al. 2017a; 2017b) 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* (Singh 1999). 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.

Impact and future directions

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, with the aim of clarifying 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+ (Baldoni et al. 2015), 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 or systems with competitive agents.

References

Baldoni, M.; Baroglio, C.; Capuzzimati, F.; and Micalizio, R. 2015. Programming with commitments and goals in jacamo+. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, 1705–1706. International Foundation for Autonomous Agents and Multiagent Systems.

Baldoni, M.; Baroglio, C.; May, K. M.; Micalizio, R.; and Tedeschi, S. 2016. Computational accountability. In Chesani, F.; Mello, P.; and Milano, M., eds., *Deep Understanding and Reasoning: A Challenge for Next-generation Intelligent Agents (URANIA)*, number 1802 in CEUR Workshop Proceedings, 56–62.

Baldoni, M.; Baroglio, C.; May, K. M.; Micalizio, R.; and Tedeschi, S. 2017a. ADOPT jacamo: Accountability-driven organization programming technique for jacamo. In An, B.; Bazzan, A. L. C.; Leite, J.; Villata, S.; and van der Torre, L. W. N., eds., *PRIMA 2017: Principles and Practice of Multi-Agent System, Nice, France, October 30 - November 3, 2017, Proceedings*, volume 10621 of *Lecture Notes in Computer Science*, 295–312. Springer.

Baldoni, M.; Baroglio, C.; May, K. M.; Micalizio, R.; and Tedeschi, S. 2017b. Supporting organizational accountability inside multiagent systems. In Esposito, F.; Basili, R.; Ferilli, S.; and Lisi, F. A., eds., *AI*IA 2017 Advances in Artificial Intelligence, Bari, Italy, November 14-17, 2017, Proceedings*, volume 10640 of *Lecture Notes in Computer Science*, 403–417. Springer.

Burgemeestre, B., and Hulstijn, J. 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.

Chopra, A. K., and Singh, M. P. 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.

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

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

Zahran, M. 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.