Value-Aligned and Explainable Agents for Collective Decision Making: Privacy Application

Doctoral Consortium

Francesca Mosca
King's College London
francesca.mosca@kcl.ac.uk

ABSTRACT

Multiuser privacy is reported to cause concern among the users of online services, such as social networks, which do not support collective privacy management. In this research, informed by previous work and empirical studies in privacy, artificial intelligence and social science, we model a new multi-agent architecture that will support users in the resolution of multiuser privacy conflicts. We design agents which are value-aligned, i.e. able to behave according to their users' moral preference, and explainable, i.e. able to justify their outputs. We will validate the efficacy of our model through user studies, oriented also to gather further insights about the usability of automated explanations.

1 INTRODUCTION AND MOTIVATION

The widespread of online services has generated an increasing concern regarding the privacy of their users. While there have been advancements and improvements in the governmental policies for individual data protection (see for instance GDPR in Europe and CCPA in California), there is still lack of support for the management of collaboratively created or owned material [6]. Collaborative editing or storing platforms (e.g., OneDrive, Dropbox, Github, etc.), smart devices (e.g., virtual assistants, self-driving cars, etc.) and Internet of Things in general, are all liable for allowing users to manage and eventually share, consciously or not, data that regards not only themselves, but also relatives, friends and colleagues. In fact, privacy is not only what we decide to share about ourselves, but also what other people can share about us [6, 16].

The most striking and studied example of collective privacy management is photo-sharing on online social networks (OSNs): nowadays and on most platforms, everyone can share a picture online without being constrained by the sharing preferences of other people who might be involved in the picture. Empirical evidence [17] shows that the majority of users have suffered multiuser privacy conflicts (MPCs) at least once, with varying severity (from mild embarrassment to the loss of job) and resolution modalities (from no solution to removal of the content).

In line with some previous literature (e.g., [7, 15, 18] and others), we want to define a multi-agent system (MAS) that is able to assist and support OSNs users while managing multiuser privacy. In such a system, that is to be eventually integrated directly in the OSNs platforms or provided as a third party service, every user would be represented by a software agent, which would interact with other agents and take decisions on behalf of the user in most situations.

We have based this line of research on the assumption of non-adversarial and collaborative behaviour among the users/agents, according to recent studies [17]: users were in general willing to modify the sharing policies assigned to certain items whenever they were made aware of the issues they generated, and they wished to have known it in advance, in order to avoid discomfort. We plan to weaken this assumption in later stages of this work, with the aim of also including in our model malicious behaviours, such as revenge-porn and cyber-bullism.

2 CONTRIBUTION

After a critical and in depth analysis of the literature on resolution of MPCs in OSNs, informed by empirical and theoretical studies in privacy, artificial intelligence and social studies, we compiled a list of requirements that models should satisfy in order to provide adequate support to OSNs users. Afterwards, we started designing MASs accordingly. A preliminary version of our model [11] represents the first attempt, to the best of our knowledge, to explicitly include a moral value component in the design of an agent for collaborative privacy management. In [10] we present an evolved model, which satisfies all the desirable requirements listed below in non-adversarial contexts.

2.1 Requirements for the Models

In recent years, scholars have suggested multiple solutions for early detection and resolution of MPCs in OSNs (see [6, 16] for more details), with many approaches consisting in agent-based models. Despite the efforts, none of the suggested solutions seems to provide an adequate support to the users. We define the adequacy of a model that aims to solve MPCs in OSNs in terms of satisfaction of a number of requirements that we identified, informed by empirical evidence and previous research [1, 8, 12, 17]:

- *explainability*: a model should be able to provide an explanation of its processes [9] to allow the users to comprehend its solutions and its effects;
- *adaptability*: a model should behave differently depending on the users' subjective preferences, because different individuals manage privacy in different ways and in different contexts [1];
- *role-agnosticism*: a model should treat all the users involved in a MPC in the same way regardless of their role, because the asymmetric access control management of uploaders and co-owners is among the main causes for MPCs [19];

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 2020, Auckland, New Zealand

https://doi.org/doi

- *utility-driven*: a model should consider solutions to MPCs according to the personal advantage or disadvantage that the involved users can face [8];
- *value-driven*: a model should support the promotion of human values, because empirical evidence suggests that users evaluate solutions and compromises while being aware that their decisions impact on the other involved users [5, 17].

We do not assume this list to be complete, but we believe these to be necessary conditions for a model to provide adequate support.

2.2 Value-aligned Agents for solving MPCs

We model the moral component of our agents according to the theory of basic values by Schwartz [14]. Values are defined by Schwartz as socially desirable concepts that allow humans to interact between themselves, representing mental goals. The human behaviour and actions are influenced by the relative order that the individual assigns to the values. Such order, representing the individual value preference, can be elicited from users through questionnaires [14], broadly validated over time and space. We prefer this value theory over others (e.g., Rokeach [13]) for a number of reasons, such as modernity, level of empirical validation, and provision of an overall value structure which directly impacts the behaviour [4].

In order to obtain agents which are value-aligned with the users they represent, we design the agents' moral component by interpreting the values in different applications. For example, considering the resolution of MPCs in [11] and [10], an agent which relatively prefers values such as *benevolence* and *universalism* (i.e. the value-direction *self-transcendence*), over *power* and *achievement* (i.e. the value-direction *self-enhancement)*, is supposed to compromise more to please its counterparts, rather than imposing its own will onto the item's co-owners.

2.3 Explainable Agents for solving MPCs

According to [9], one way to design an explainable autonomous agent consists of providing it with a *cognitive process*, i.e. the technical ability of determining the necessary information to explain the events, and a *social process*, i.e. the social ability to efficaciously convey the explanation to the user.

In [10] we introduce how the *cognitive process* can be guaranteed by applying techniques from computational argumentation. In particular, the agent can follow the *Practical Reasoning Argumentation Scheme* (PRAS) [3] while deliberating on the action to perform, by also bearing in mind its own moral values: "in the current situation, I should offer/accept the policy p in order to reach an agreement and promote my value v".

For the computational realisation of PRAS, we use an *Action-based Alternating Transition System with Values* (AATS+V) [2], which provides a useful method to model transitions between different states of the world as joint actions, when such actions are labelled with the values that they promote. A joint action is meant as a sequential or synchronous combination of single actions performed by different agents. For instance, in the MPC application we discuss in [11] and [10], a joint action is represented by the offer (i.e. a sharing policy) that the uploader presents and the response of the co-owners, which can either accept or reject such offer.

Each agent can reason with the AATS+V in order to identify the best individual action, evaluated in terms of personal utility and adherence to moral values. Also, by tracking the reasoning process, the agent gathers all the necessary knowledge to explain the events.

3 FUTURE WORK

We will work towards the extension and validation of our model for a value-aligned and explainable agent in the MPC context.

First, we will evaluate through software simulations and user studies the goodness of our solution concept compared to the MPCs solutions suggested by other researchers. Before validating our model with users, we will focus on designing the *social process* [9], by studying the best and most effective way to convey explanations, eventually counterfactual, to the interested users.

Then, we would like to weaken our assumption regarding non-adversarial and collaborative behaviour, in order to also tackle those rare but more severe conflicts where a malicious component is present, such as revenge-porn and cyber-bullism.

4 ACKNOWLEDGEMENT

The author would like to thank her PhD supervisor Jose M. Such for his guide and useful comments, and Stefan Sarkadi for his unconditional support to this research.

REFERENCES

[1] A. Acquisti, L. Brandimarte, and G. Loewenstein. 2015. Privacy and human behavior in the age of information. *Science* 347, 6221 (2015), 509–514.
[2] K. Atkinson and T. Bench-Capon. 2007. Practical reasoning as presumptive argumentation using action based alternating transition systems. *AIJ* 171, 10-15 (2007), 855–874.
[3] K. Atkinson, T. Bench-Capon, and P. McBurney. 2006. Computational representation of practical argument. *Synthese* 152, 2 (2006), 157–206.
[4] A. Bardi and S.H Schwartz. 2003. Values and behavior: Strength and structure of relations. *Personality and social psychology bulletin* 29, 10 (2003), 1207–1220.
[5] A. Besmer and H.R. Lipford. 2010. Moving beyond untagging: photo privacy in a tagged world. In *CHI*. ACM, 1563–1572.
[6] M. Humbert, B. Trubert, and K. Huguenin. 2019. A Survey on Interdependent Privacy. *Comput. Surveys* (2019), 35.
[7] D. Kekulluoglu, N. Kökciyan, and P. Yolum. 2018. Preserving privacy as social responsibility in online social networks. *ACM TOIT* 18, 4 (2018), 42.
[8] H. Krasnova, S. Spiekermann, K. Koroleva, and T. Hildebrand. 2010. Online social networks: Why we disclose. *JIT* 25, 2 (2010), 109–125.
[9] T. Miller. 2018. Explanation in artificial intelligence: Insights from the social sciences. *AIJ* (2018).
[10] F. Mosca and J. Such. 2020. Towards a Value-driven Explainable Agent for Collective Privacy. In *Extended Abstract accepted at AAMAS (2020)*.
[11] F. Mosca, J. Such, and P. McBurney. 2019. Value-driven Collaborative Privacy Decision Making. In *AAAI PAL Symposium*.
[12] F. Paci, A. Squicciarini, and N. Zannone. 2018. Survey on access control for community-centered collaborative systems. *Comput. Surveys* 51, 1 (2018).
[13] M. Rokeach. 1973. *The nature of human values*. Free press.
[14] S. H. Schwartz. 2012. An overview of the Schwartz theory of basic values. *Online readings in Psychology and Culture* 2, 1 (2012), 11.
[15] J. Such and N. Criado. 2016. Resolving Multi-Party Privacy Conflicts in Social Media. *IEEE TKDE* 28, 7 (2016), 1851–1863.
[16] J. Such and N. Criado. 2018. Multiparty Privacy in Social Media. *Commun. ACM* 61, 8 (2018), 74–81.
[17] J. Such, J. Porter, S. Preibusch, and A. Joinson. 2017. Photo privacy conflicts in social media: a large-scale empirical study. In *CHI*. ACM, 3821–3832.
[18] J. Such and M. Rovatsos. 2016. Privacy Policy Negotiation in Social Media. *ACM TAAS* 11, 1 (2016), 1–29.
[19] P. Wisniewski, H. Lipford, and D. Wilson. 2012. Fighting for my space: Coping mechanisms for SNS boundary regulation. In *CHI*. ACM, 609–618.