Situational Trust in Self-aware Collaborating Systems

Alessandro V. Papadopoulos
Mälardalen University, Sweden
alessandro.papadopoulos@mdh.se

Lukas Esterle
Aarhus University, Denmark
lukas.esterle@eng.au.dk

***Abstract*—Trust among humans affects the way we interact with each other. In autonomous systems, this trust is often predefined and hard-coded before the systems are deployed. However, when systems encounter unfolding situations, requiring them to interact with others, a notion of trust will be inevitable. In this paper, we discuss trust as a fundamental measure to enable an autonomous system to decide whether or not to interact with another system, whether biological or artificial. These decisions become increasingly important when continuously integrating with others during runtime.**

***Index Terms*—trust, integrating systems, interaction, autonomy, self-awareness**

I. Introduction

Autonomous systems operating in a common environment will inevitably affect each other either actively, by directly interaction, or indirectly, by manipulating the shared environment. SISSY systems [4] specifically integrate the actions of others constantly in order to achieve their own goals faster. This requires them to make decisions what systems to cooperate with and which ones to avoid. The decisions need to be made during runtime autonomously by the systems while operating in such a shared environment about other system for which they do not have control. In this work, we consider SISSY systems operating in the real environment and integrating other systems operating in the shared real-world environment. This means, our SISSY systems are autonomous cyber-physical systems. With the Machine-to-Machine (M2M) communication for active interaction, systems can negotiate about their actions, intentions, and goals. However, a notion of trust among systems will be required in the near future to avoid complicated negotiations and/or to assure that tasks will be accomplished. In this short position paper, we propose a notion of *knowledge-based trust*, inspired by how trust is established among humans, and how it drives the interaction among them [19]. Similar considerations were also drawn in the analysis of the interaction between businesses [2]. In both cases, the notion of trust can be enhanced (or be destroyed) over time, due to a sequence of interactions [15].

Such notion can be used in the context of Self-Adaptive Systems (SASs) to foster the collaboration among several SAS entities based on the pre-defined levels of trust towards the other systems, as well as based on the collected evidence of the outcome of the previous interactions. Figure 1 shows several SASs implementing MAPE-K loop functionalities [13]

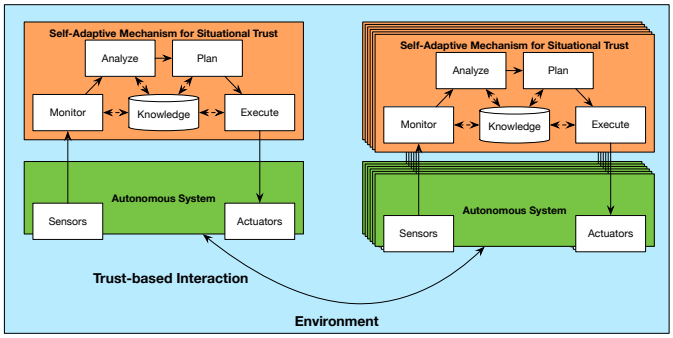

Fig. 1. Trust-based interaction of several SASs implemented as MAPE-K loops.

deployed in the same environment [11]. In order to achieve their prescribed goals the SASs can choose to cooperate, negotiate, or compete based on the environment in which they are deployed [8].

In addition to foster and utilise trust among machines to accelerate and bootstrap interactions, trust can also be used to foster the interaction between machines and humans [1], [14], [16]. However, this short position paper does not look into the trustworthy-ness of machines and how machines can establish trust towards humans. Interested readers can look towards related work in this area [17], [18], [20].

In the domain of Organic Computing (OC), the idea of trust to facilitate collaboration has been researched intensively. Kiefhaber et al. [12] use reputation to establish neighbourhoods of trust. This allows systems to get information and reputation evaluation of others through already trusted systems. Bernard et al. [5] establish trust-communities within a network of autonomous agents. The community can in turn detect free loaders and malicious agents over time and sanction them or exclude them from interactions entirely. Kantert et al. [10] proposes a reward system incorporating a value for trust in order to enable individual systems to autonomous decisions whether or not to take on tasks. To advertise tasks, systems can in turn utilise the established trust to focus their advertisement efforts. In their approach, the notion of trust is limited to

In the current design of autonomous systems, limited attention has been given to the role that trust plays in the interaction. This is quite apparent when considering the interaction between humans [21], that would take different actions based

on the level of trust that they have for each other. In this paper, we argue that such trust is also affected by additional factors, such as the environment *where* the interaction takes place, the circumstances occurring in the environment at the moment *when* the interaction takes place, and so on. For example, one may not leave personal belongings with a stranger unless it is an emergency, or it is in a context where trust can be safely assumed, e.g., a police station or with a police officers. We therefore propose to extend existing approaches on trust and reputation by a notion of where, when, who, what, why, and how actions are requested. In addition, we argue that a reflection mechanism allows an intelligent SISSY system to utilise previous experiences and established trust to further develop and improve the trust for individual other systems. This 'improvement' can mean that trust is increased or decreased.

Similar concepts apply in the context of interaction between autonomous systems, where the notion of trust is typically not part of the decision-making process, and based on prescribed behaviours developed by the system designer but not affected by the end-users, as highlighted in [3], [9]. However, even when the foundational framework for trust and how it should be generated is defined by the end-user, the autonomous system will still adjust trust within this defined framework based on ongoing interactions and experiences.

In this paper, we discuss the notion of "situational trust" as a fundamental measure to enable an autonomous system to decide whether or not to interact with another system. We analyze the main factors contributing to the assessment and establishment of trust in the interaction of multi-agent systems, independently of their intrinsic nature (biological or artificial).

II. Situational Trust

We argue that an awareness of trust in autonomous self-aware systems collaborating with each other is twofold. First, a systems aware of the trust towards other entities may lead to a more informed decision processes resulting in *better* interactions overall. Second, a system aware of the trust of others as well as of its own actions affecting this trust can choose the behaviour towards specific others accordingly.

We define *situational trust* as an estimate of how much another agent is trustworthy in a specific situation. It represents the trust between the truster, giving trust, and the trustee, receiving trust ins specific situations for specific purposes. Each situation can be characterised according to the following factors:

- *Who*: relates to the identity of the other agent.
- *Where*: relates the environment in which the interaction should take place
- *When*: is the time the interaction should take place
- *What*: represents the outcome of the interaction
- *How*: are the type of actions involved in the interaction
- *Why*: represents the underlying reasons and goals of the interaction that is also driven by the ethical values of the other system

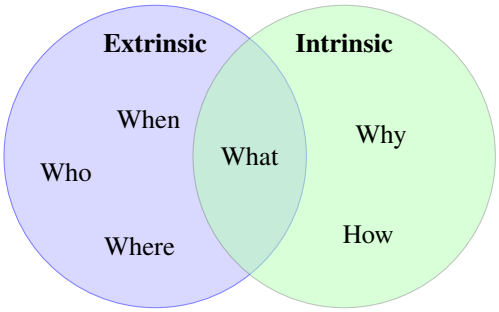

Fig. 2. *Situational trust* composed of intrinsic and extrinsic factors.

We divide these characteristics of situational trust into *Extrinsic* and *Intrinsic* parameters, assuming the trustee and truster are both able to reflect on their environment, actions, and respective impact. *Who*, *Where*, and *When* are physical aspects. Marsh et al. [14] refer to these as 'Places', the context subsuming these three aspects in which trust is established or utilised. In any case, each system involved in the trust-relationship can potentially perform an assessment of those characteristics autonomously. This can be done immediately before initiating any interactions. *What* defines the outcome expected from the interaction. Importantly, this expected outcome can be different for all systems involved. However, we cannot definitely position *What* as an intrinsic or extrinsic parameter as it depends on whether the system has to re-interpret it or not. Trust plays into this aspect as more trust potentially requires less negotiation and explanation of the expected goals and outcomes. The *How* defines what actions should be used to accomplish the interaction. While the *What* might dictate the *How*, there is often some potential in variation. In an autonomous, self-aware system these actions and their schedule of execution is defined by the enacting system itself. Trust enables the trustee to perform the actions in whatever way it deems them feasible while still achieving the goals and requests while at the same time not violating any (potentially not communicated/undefined) constraints of the truster. While the *Why* might be communicated by the truster, it relies on the re-interpretation and re-evaluation of the trustee. A trustee might not deem the request and underlying reasons and goals as important enough to perform the actions. While the extrinsic parameters are related to the sensing of the environment in the MAPE-K loop, and possibly communicating with the other agents (as shown in Figure 1), the intrinsic parameters are stored in the Knowledge component of the MAPE-K loop, and can affect both the way the agent Monitors, Analyzes, Plans, or Executes its actions.

We argue, when situational trust is required, all these aspects need to come together. These aspects form a constraint satisfaction problem with potentially multiple objectives where one or multiple characteristic might outweigh the remaining.

When an interaction shall occur between two agents, the one initiating the interaction will have a *prior situational trust* that comes from the experience of previous (direct or

indirect) interactions. Based on the current situational trust – that depends only on the specific situation in which the interaction occurs – and on the prior situational trust, the agent can create an *expected situational trust* that will drive the current interaction between the agents. Finally, based on the actual outcome of the interaction, the expected situational trust will be updated with the new piece of information, and translated into a *posterior situational trust*. The posterior will be used as a prior situational trust in the next interaction.

A. Giving Trust

Enabling systems to trust each others can lead to short-cuts in negotiations, explorations, and evaluations of each others actions. With a rising amount of trust, a system might be trusting another unconditionally. This means, the truster is prepared to take a higher risk of betrayal. These can lead to a reduction of resource utilisation in equal situations occurring. Even more so, if a system is aware of its situational trust, we can consider transfer of trust from one situation to others. This might be a transfer from simple situations with little risk involved to more complex situations, carrying more risk. A question arising is whether there is a correlation between trust and hence higher risk and the amount of required resources to achieve the same task without trust?

B. Establishing Trust

Utilising established trust allows for rapid integration without additional extensive explorations, negotiations, or evaluations. To achieve this, a system aware of others' trust towards itself and how its own actions and interactions affect this trust, can deliberately choose actions to shape the trust. This allows a self-aware system to direct its behaviour in order to increase or decrease trust in others towards itself and facilitating potential collaboration. However, the understanding between actions and resulting increase or decrease of trust also requires a complex mapping of the *situational trust* and its concomitant factors of *who*, *where*, *when, why, what*, and *how*.

III. Experience and Reflection

SISSY systems autonomously interacting and integrating the actions of others, are not limited to evaluating the current situations but can reflect on previous experiences [4]. Employing such reflective mechanisms individuals can trust or distrust selectively. For example, trusting system A in one location or time for specific reasons does not mean it also trust system B in the same situation or that it trusts system A in other circumstances (e.g., different time or location).

At the same time, establishing or diminishing trust is not a single shot action. Systems can learn, calibrate, and refine their trust for individual systems. Having established some trust with system A in one situation makes it a preferred collaborator in another situation over a system B that is unbeknownst to us.

One might argue to add some kind of decaying or forgetting factor towards trust over long periods of time. That would diminish trust over time in the absence of trust-establishing interactions. However, ongoing research works towards detection of intentionally changed behaviour through direct interaction and observation [6], [7].

Within the MAPE-K loop, the systems will perform this reflection within the 'Analyze' part of the loop. Based on historic information gathered and stored in the 'Knowledge' base, plans can be generated and interactions potentially short-cut.

IV. Conclusion

In this short paper, we examine situational trust and its characteristics. We follow the argument of Marsh et al.: "Trust matters" [14]. Even more so in autonomously integrating systems utilising the actions of others in order to improve their collective efficiency over time.We argue that self-aware systems, able to interact with others, can establish individual trust and reputation. However, this trust is dependent on the current situation and the more trust a system is having in another, the more risk it is willing to take in case the trust is misplaced. Over time, a system can build up an understanding of where, when, with whom, why and how to utilise trust to bootstrap interactions most efficiently. This historic information is constantly shaped and refined through every interaction taken with each individual system. We also argue that not all characteristics presented in this short paper are symmetric and deterministic. Systems, just as humans, are prone to misinterpretation—may it be due to incorrect sensor readings or miscommunication or different, underlying experience informing the decisions to be made.

References

[1] P. Andras, L. Esterle, M. Guckert, T. A. Han, P. R. Lewis, K. Milanovic, T. Payne, C. Perret, J. Pitt, S. T. Powers, N. Urquhart, and S. Wells. Trusting intelligent machines: Deepening trust within socio-technical systems. *IEEE Technology and Society Magazine*, 37(4):76–83, Dec 2018.

[2] A. Argandoña. Sharing out in alliances: Trust and ethics. *Journal of Business Ethics*, 21(2-3):217–228, 1999.

[3] M. Autili, D. D. Ruscio, P. Inverardi, P. Pelliccione, and M. Tivoli. A software exoskeleton to protect and support citizen's ethics and privacy in the digital world. *IEEE Access*, 7:62011–62021, 2019.

[4] K. L. Bellman, J. Botev, A. Diaconescu, L. Esterle, C. Gruhl, C. Landauer, P. R. Lewis, A. Stein, S. Tomforde, and R. P. Würtz. Self-improving system integration - status and challenges after five years of sissy. In *Proceedings of the IEEE International Conference on Self-organizing and Self-Adaptive Systems Workshop*, pages 160–167, 2018.

[5] Y. Bernard, L. Klejnowski, J. Hähner, and C. Müller-Schloer. Towards trust in desktop grid systems. In *2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing*, pages 637–642, 2010.

[6] L. Esterle and J. N. Brown. Levels of networked self-awareness. In *2018 IEEE 3rd International Workshops on Foundations and Applications of Self* Systems (FAS*W)*, pages 237–238, 2018.

[7] L. Esterle and J. N. A. Brown. I Think Therefore You Are: Models for Interaction in Collectives of Self-Aware Cyber-physical Systems. *ACM Transactions on Cyber-Physical Systems*, pages 1–24, 2020. In Press.

[8] J. Ferber and G. Weiss. *Multi-agent systems: an introduction to distributed artificial intelligence*, volume 1. Addison-Wesley Reading, 1999.

[9] P. Inverardi. Ethics and privacy in autonomous systems: A software exoskeleton to empower the user. In R. Calinescu and F. Di Giandomenico, editors, *Software Engineering for Resilient Systems*, pages 3–8, Cham, 2019. Springer International Publishing.

[10] J. Kantert, Y. Bernard, L. Klejnowski, and C. Müller-Schloer. Estimation of reward and decision making for trust-adaptive agents in normative environments. In E. Maehle, K. Römer, W. Karl, and E. Tovar, editors, *Architecture of Computing Systems – ARCS 2014*, pages 49–59, Cham, 2014. Springer International Publishing.

[11] J. O. Kephart, A. Diaconescu, H. Giese, A. Robertsson, T. Abdelzaher, P. Lewis, A. Filieri, L. Esterle, and S. Frey. *Self-adaptation in Collective Self-aware Computing Systems*, pages 401–435. Springer International Publishing, Cham, 2017.

[12] R. Kiefhaber, S. Hammer, B. Savs, J. Schmitt, M. Roth, F. Kluge, E. Andre, and T. Ungerer. The neighbor-trust metric to measure reputation in organic computing systems. In *2011 Fifth IEEE Conference on Self-Adaptive and Self-Organizing Systems Workshops*, pages 41–46, 2011.

[13] M. Maggio, T. Abdelzaher, L. Esterle, H. Giese, J. O. Kephart, O. J. Mengshoel, A. V. Papadopoulos, A. Robertsson, and K. Wolter. *Self-adaptation for Individual Self-aware Computing Systems*, pages 375–399. Springer International Publishing, Cham, 2017.

[14] S. Marsh, T. Atele-Williams, A. Basu, N. Dwyer, P. R. Lewis, H. Miller-Bakewell, and J. Pitt. Thinking about trust: People, process, and place. *Patterns*, 1(3):100039, 2020.

[15] S. Marsh and M. R. Dibben. Trust, untrust, distrust and mistrust – an exploration of the dark(er) side. In P. Herrmann, V. Issarny, and S. Shiu, editors, *Trust Management*, pages 17–33, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[16] S. P. Marsh. Formalising trust as a computational concept, 1994.

[17] J.-P. Steghöfer, R. Kiefhaber, K. Leichtenstern, Y. Bernard, L. Klejnowski, W. Reif, T. Ungerer, E. André, J. Hähner, and C. Müller-Schloer. Trustworthy organic computing systems: Challenges and perspectives. In B. Xie, J. Branke, S. M. Sadjadi, D. Zhang, and X. Zhou, editors, *Autonomic and Trusted Computing*, pages 62–76. Springer Berlin Heidelberg, 2010.

[18] R. Toegl, T. Winkler, M. Nauman, and T. Hong. Towards platform-independent trusted computing. In *Proceedings of the 2009 ACM Workshop on Scalable Trusted Computing*, page 61–66, New York, NY, USA, 2009. Association for Computing Machinery.

[19] M. Tschannen-Moran and W. K. Hoy. A multidisciplinary analysis of the nature, meaning, and measurement of trust. *Review of Educational Research*, 70(4):547–593, 2000.

[20] T. Winkler and B. Rinner. Trustcam: Security and privacy-protection for an embedded smart camera based on trusted computing. In *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 593–600, 2010.

[21] L. S. Wrightsman. Chapter 8 - interpersonal trust and attitudes toward human nature. In J. P. Robinson, P. R. Shaver, and L. S. Wrightsman, editors, *Measures of Personality and Social Psychological Attitudes*, pages 373 – 412. Academic Press, 1991.