Gamification of Trust in HRI?

Michael FUNK[a,b,1], Bernhard DIEBER[c], Horst PICHLER [c] and Mark COECKELBERGH[a]
[a] *Philosophy of Media and Technology, University of Vienna, Austria*
[b] *Cooperative Systems, Faculty of Computer Science, University of Vienna, Austria*
[c] *Institute for Robotics and Mechatronics, JOANNEUM RESEARCH, Klagenfurt, Austria*

Abstract. In this transdisciplinary paper we discuss the question whether trust in human-robot-interaction (HRI) can be gained by gamification. Therefore, the concept of credibility will be introduced. A specific focus is on the question concerning the implementation of ethical rules in robotic safety systems. With a focus on Wittgenstein as a philosopher of technology we argue that in many fields of application cultural issues play a crucial role that cannot be controlled in a top-down approach. Instead, we follow a process-oriented bottom-up understanding of trust which pays attention to different social situations of normative practices. In order to combine our transdisciplinary philosophical and engineering points of view, a model for "gamifying trust" including ethical reflection (**Figure 1.**) as well as a short sketch of a possible technical implementation are presented.

Keywords. Trust, Credibility, Robot Ethics, Transdisciplinarity, Annoying Valley

[1] Introduction

The topic of trust in technologies receives a growing interest and is intensively discussed with a specific focus on human-robot-interaction (HRI) [1] [2] [3] [4] [5] [6] [7]. Another context of the debate relates to EU frameworks of robotics implementation in societies. On a regulatory level, steps towards a standardization of normative treatment of robotics have been presented for instance in the *Ethics Guidelines for Trustworthy AI* – released in April 2019.[2] A certain interest of the European Commission is expressed with respect to the networking of certain groups of interests and stakeholders including policy makers, companies, universities, public media, the health care sector etc. Transdisciplinary communication strategies are supposed to be extended and applied in order to gain and regulate the uptake of robotics in a sustainable and responsible way. This paper is the result of the multifaceted ethical debates within one of these regulatory activities. Following the idea of transdisciplinary stakeholder connection, it combines some outputs of the EU INBOTS[3] project with research activities of the Austrian FFG project CredRoS[4]. We are following a

1 Michael Funk, Department of Philosophy, University of Vienna, Universitätsstraße 7 (NIG), A-1010 Wien, Austria. E-Mail: funkmichael@posteo.de; Web: www.funkmichael.com

2 https://ec.europa.eu/futurium/en/ai-alliance-consultation

3 http://inbots.eu/

4 https://www.joanneum.at/robotics/referenzprojekte/credros/

transdisciplinary approach since philosophy of technology (University of Vienna) is linked to robotics engineering (JOANNEUM Robotics, Klagenfurt). The approach present here remains tentative since it´s a snapshot of ongoing research, geographically focused on Austria and limited to two disciplinary perspectives. Of course, the inclusion of other academic areas like social sciences or psychology might enhance our approach in future work. This could also include a combination with other recent activities like a 2020 presented Taxonomy of *Trust-Relevant Failures and Mitigation Strategies* [7] and others.

Transdisciplinary investigations with a specific focus on HRI stand in the focus of *Robophilosophy* investigations [8] [9]. Research in the field gained a certain interest in process-oriented social ontologies and a relational understanding of diverse modes of co-working [10] [11] [12] [13]. In this context we situate our analysis of trust in technology. Therefore, the overarching question, "What is trust in robots/robotics?" receives a more concrete focus: What would a model for the description and ethically reflected processing of trust in robots look like? Following the approaches of a processual and relational ontology we argue that trust in technology is a strongly *culturally embedded* process – not a passive mental state that can only be naturalized from an external observer's perspective. It depends on social interaction and the *successful repetition of actions*. What we want to show is that trustworthy HRI cannot be planned top-down. Instead it heavily depends on bottom-up processes. However, there remains a certain influence of top-down operations in processes of trust by which also ethical assessment is supposed to be implemented. In order to analyze the dynamic interrelations between bottom-up and top-down processes, we present an epistemic model of trust and trustworthiness from a transdisciplinary point of view (**Figure 1.**):

First, we follow the common differentiation between trust and trustworthiness. Trust relates to a personal, individual moral perspective that is socially embedded (more "subjective") whereas trustworthiness is a matter of the institutional, reflexive-ethical point of view (more "objective"). An analog difference is that between acceptance and acceptability [14].

Second, we differentiate two kinds of perspectives. The *perspective of action (POA)* which is linked to individual experience of the performance – commonly known as *first-person-perspective/experience (1PP)* in the philosophy of mind. In contrast, the *perspective of description (POD)* relates to a more objective point of view from the outside – commonly known as *third-person-perspective/experience (3PP)*. Both perspectives receive a methodical significance in order to gain a more detailed understanding of processes of trust [15].

Third, related to the differentiation of trust and trustworthiness, the *POA* means in the case of trust the concrete morally significant habits of a person depending on the social embedding; in the case of trustworthiness the *POA* means the personal ethically relevant value judgment depending on normative rules. *POD*, the perspective of description, relates in the case of trust to a descriptive operation in the sense of matter of fact – in philosophical terms: descriptive ethics (*Genese*); when it comes to trustworthiness, the *POD* stands for normative ethics (*Geltung*), the matter of ought. Trust is the object of descriptive ethics because it´s about the factual values that shape human habits in real life. Trustworthiness, in contrast, is the object of normative ethics since it´s about the ought, that is, about what we *should* trust for rational reasons. With this we follow the classical philosophical difference between *Genese* and *Geltung*. For ethical assessment the very important differentiation between matter of fact (*Genese*, descriptive ethics) and matter of ought (*Geltung*, normative ethics) is implemented as

well [16] [17] [18]. Steps one, two and three belong to the genuine philosophical disciplinary contribution.[5]

Fourth, we introduce the concept of *credibility* as an engineering category for the application of trustworthiness. It relates to ethically relevant empirical criteria, but it´s not an ethical normative principle as such (!). Credibility depends on ethically reflected normative frameworks and serves as a gradual step for the implementation of trustworthiness in social processes of trust. The aim is not to technocratically control, but to set *conditions* for the pragmatically approved repetition of successful HRI. With this we follow a pragmatic truth criterion [20]. Therefore, credibility (engineering perspective) such as ethics (philosophical perspective) are processes themselves, implemented into a permanent feedback loop within the processes of trust.

Fifth, the aim of our model is to suggest a possible concrete way for the processing of trust in HRI in concrete culturally embedded situations with a focus on the interrelation between bottom-up and top-down operations. Our (both philosophical and engineering) hypothesis is that a processing of trust – and therefore the chance to gain trust in a trustworthy sense – can be realized in the form of a *gamification*. Like in computer games, the HRI might be learned by employees, professionals or layperson while playing some kind of tutorial in which trial and error bottom-up processes are initiated in a top-down safely environment. People get time to try out the chances and risks of the concrete interaction without a certain pressure, like economic success or efficiency in terms of working hours. Like children playing in the sandbox, here the way is seen as the goal.

Sixth, from a philosophical point of view, the processing of trust is linked to certain characteristics of human life. Representative for many, we shortly discuss three issues: finitude, vulnerability and Annoying Valley Effects. The last one – Annoying Valley Effects – is an homage to Mori´s Uncanny Valley [21] which we introduce as an advanced concept in order to describe the significance of trust.

Figure 1. shows a visual model of how we understand the framework as feedback loop. Processes of trust are explicated in descriptive operations – from a transdisciplinary point of view this might be a possible link to psychology or social sciences. Once the values of factual trust are explicated they turn into objects of trustworthiness – critically and rationally reflected within normative ethics. The results of these reflections lead to concrete normative requirements (like sustainability, respecting privacy etc.) that are translated into technical standards and technical (!) norms of robotic safety systems. These standards set top-down conditions for the performance (playing within the gamification of trust). The process of gamification as such is not a static list of top-down rules (not like Asimov´s robot laws), but an existential social performance traversed by contingent bottom-up activities. Ethics here is methodically understood as applied transdisciplinary pragmatism, since the contingent process of gamification requires solutions in a short period of time when unexpected practical conflicts arise. It´s an ethics for risky situations of uncertainty. Of course, processes of trust can fail and with this also certain types of HRI or social implementation of HRI. Consequently, the ethics of HRI, as conceptualized in our model, is strongly influenced by applied ethics since the 1970s and is insofar also

5 For a more detailed foundation of the philosophical perspective in robot ethics see M. Funk "What is Robot Ethics? ...and can it be standardized?" in this volume. And M. Funk *Roboter- und Drohnenethik. Eine methodische Einführung* [19].

methodically in line with the current AI HLEG *Ethics Guidelines* – for instance [22] [23].

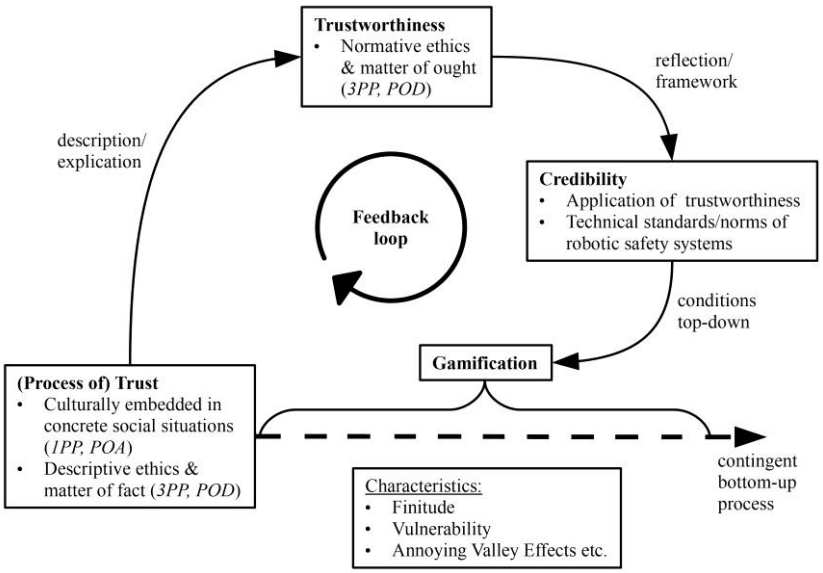

Figure 1. A conceptual framework for the gamification of trust.

In the second section we shortly summarize a philosophical foundation of the gamification of trust in HRI with respect to a Wittgensteinian approach. Finitude, vulnerability and Annoying Valley Effects are discussed as some existential characteristics of processes of trust. Section three contains the engineering perspective combining a preliminary sketch of a possible case study with a scenario analysis of how credibility and gamifying trust may be applied. We close our paper with a tentative outlook including some open questions and ideas for future research.

[2] Technology Games and Annoying Valley Effects – The Philosophical Perspective

In order to better understand the processing of trust – and therefore also the gamification of trust – we shortly summarize some basic insights of a Wittgensteinian philosophy of technology [24] [25]. Wittgenstein contributed to the philosophy of language by putting the focus on language practice and ordinary life since his 1930s works – in contrast to the previous theoretical, abstract and formal-logical turn in the *Tractatus*. Language games and forms of life are two significant concepts of this later period, where language is seen as something like playing with meaning in very specific social situations. Thereby words relate to also technologically shaped forms of life. Technique in the sense of embodied skills receives a transcendental status in Wittgenstein's approach. Going beyond a Kantian understanding the author shows for instance in his *Philosophical Investigations* how socially shared practice and the repetition of successful actions in a very pragmatic sense serves as non-propositional

truth criterion [24] [25]. Playing with language is a skillful process just like playing with technologies in order to master the world. There is a close relation between language games and technology games [26]. Wittgenstein´s language critique leads to the analysis of practice as a form of transcendental grammar – which is about philosophical sense, not about linguistic formal rules of how to create a correct sentence.

In an existential manner several characteristics like finitude and vulnerability existentially traverse transcendental grammars of human actions [27] [28]. Here we want to insist that the gamification as a processing of trust needs to pay attention to the many asymmetries in HRI [10]. For instance, a robotic body made of copper and iron is stronger and more stable than a human body of organic material. But language practice also receives elements of finitude and vulnerability. The ways we communicate are limited, forced to come to an end, and constantly endangered by misunderstandings, deceptions, illusions or manipulations. Humans exist in a fragile and logically non-perfect way. It´s very important to pay attention to these asymmetries in order ensure humanity in HRI. From a psychological point of view Mori [21] analyzed in 1970 the famous Uncanny Valley. In a process-related analogy we want to put the focus on what we call *Annoying Valley Effect*. This is an existential feature of interactive processes in which the gaining of trust can be seen as a benefit of challenging errors and disturbance of the interactions. In contrast to an accident the postulated Annoying Valley Effect relates a temporal frustration that is significant for a real processing of trust bottom-up – which is not overwhelmed by a top-down dictation. In other words: without temporary frustration (literally falling the Annoying Valley) there would be no gaining of trust.

[3] Trust and Credibility – The Engineering Perspective

The project "CredRoS – Credible and Safe Robot Systems" at the Institute for Robotics and Mechatronics of JOANNEUM Research aims at developing technological building blocks for robots that can gain the trust of their users. This includes researching fundamental techniques for extended robot safety and robot movement mechanics. It also includes work towards enabling robot transparency for existing robot applications. The following sections present some key aspects that this project currently focuses on. A project with strong technical focus but also high non-technical implications and aspects such as one in the field of trustworthy technology needs an interdisciplinary part to ensure a proper exchange of viewpoints and results with other scientific disciplines. This is why CredRoS defines a dedicated work package for exchange and interaction with other scientific disciplines such as philosophy. In course of this, we have also worked on clarifying the difference between trustworthiness and credibility.

The key requirements of trust, as defined by EU HLEG, considered in the CredRoS trustworthy robot architecture are accuracy, reliability, reproducibility, general safety, security, fallback planning, traceability, explainability, and communication. Even if an autonomous system incorporates all these key requirements, still the key question remains how to come to a sufficient level of trust between the workers and the robots around them. A worker must feel safe around the robot or rely on the robot to fulfill its task. It is well known that trust cannot be simply switched on (or commanded). Trust between humans builds over time by getting to know each other. Analogously for humans and robots trust comes with training and

practice, thus also develops over time [29]. A key question here is how to enable the build-up of trust of a worker with her robot co-worker and how technical means can support that. The terms "trustworthy" and "credible" in our daily use seem often to be interchangeable. Within this project, in some aspects we also use them this way. However, we want to make the following distinction between trust and credibility: According to the Oxford English Dictionary, "trust" means the "Firm belief in the reliability, truth, or ability of someone or something"[6] while "credible" is defined as "Able to be believed in, justifying confidence"[7]. The words have similar meanings, although we see an important distinction: Trust has a stronger subjective aspect while credibility needs some sort of justification or proof. One could paraphrase the distinction as "Credibility comes from the head while trust comes from the heart". In the technical context, we want to *characterize a credible system in the following working definition as "being constructed in accordance to established and accepted technical guidelines"* while *a trustworthy system is "a system users are willing to rely on based on their experience with this system"* (this belongs to the fourth issue mentioned in the introduction, where the philosophical conceptualization is linked to the engineering perspective).

To us, a credible system adheres to guidelines and standards from the technical perspective while at the same time being constructed with its effect on its users in mind. A trustworthy system may be unprofessionally built (but does not need to be) but earns the trust of a user by its predictability. Thus, we expect a credible system to also be able to earn the trust of a user. However, we accept that trust itself is a subjective feeling that must be earned by a robot as well as other humans (and can thus not be built-in into a robot per se). This is why we start from *established standards in robot safety* (where adhering to already provides a good part of credibility but not necessarily trustworthiness). Summarizing, we mean "credibility" to be a very technically focused term in the context of robotics but need to understand the (interdisciplinary) meaning of trust in order to go from currently established technical standards towards building flexible robot systems that can subjectively and objectively trusted by its human users in future.

Credibility: A Headstart in Trust?

Having discussed credibility as being a "certifiable" kind of trust, we would like to examine if a robot is more trustworthy to a human if it can "prove" its trustworthiness. In our everyday life, we tend to put trust into standardized processes. As a simple example, consider why we trust that food we buy at the supermarket. Besides a simple sensory examination, there is not much we can do ourselves to make sure it is edible. Still, we hardly think about this issue when we buy it. One reason is the way we rely on the standardization in our food chain[8]. In the same way, workers who interact with an industrial robot can rely on the standardized safety certification of the robot workplace and can thus start with a higher level of trust towards the system. Judging from that,

6 https://www.oed.com/view/Entry/207004

7 https://www.oed.com/view/Entry/44110

8 In [30], a similar argument is made by describing standardization to be used in absence of trust and thus replacing the need for direct trust. This supports the argument we are making here in the sense that if direct trust is not feasible (we cannot check the whole production chain of our food anymore), our trust in standardization bodies can be reflected on the process at hand.

having built a credible robot based on well-known standards like the ones for robot safety, that robot could receive an initial trust bonus in a gamification-enhanced trust-building process with a user.

Another interesting aspect here is transitivity or transferability of trust. It has been shown that users tend to transfer trust from one technical system to another under certain circumstances e.g. when the trusted system links to the new system [31] or if the new system is endorsed by a trusted entity [32]. Similar effects can be observed of course also in the context of robots – for instance in a teacher-learner constellation [33]. What we expect is also a sense of trust in a robot model. So, if a human trusts a robot of a certain model (e.g., one Pepper robot), will she also trust another robot of the same type? Technically speaking, this trust would not necessarily be justified. A robot is a flexible machine that can change its behavior based on the program that it executes. This is why a safety certified robot application in industry does not permit any changes to the program of the installation without re-certification.

Thus, the trust in a robot should primarily be focused on the specific program it is executing (just like we trust one app on a smartphone more than another) – in other words: trust depends on the process of interaction not only on the material surface structure. Transferring this to the idea of using dedicated trust-building games, this means that the gamification needs to be built into the application we want the user to trust. It´s not only about the material design of an artifact but primarily about process engineering (*Verfahrenstechnik*). We argue that having a dedicated – separate – robot program to build up trust and then switch to the productive program would not be justified. Of course, we are aware that we neither can force a user to trust a robot, nor can we force her to distrust a newly loaded program. So, in order to use gamification to support trust-building, a concept of integration into the very specific envisaged productive program is required. Therefore, we argue that credibility may help by providing a certain level of standardization or certification from a trusted authority to increase initial trust levels. What we want to investigate are the concrete objectifications of social norms that might define a robotic system as "credible" **(Figure 1.)**.

In CredRoS, we have defined a use-case that will be used throughout the project in order to study some of the possible objectifications of credibility-implementation into robots – as a form of process engineering. Within this work, we also aim to look at gamification strategies to support the trust-building processes. Therefore the use-case will reflect various types of human-robot collaboration being *co-existent* (human and robot separated by barriers working on their own tasks), *cooperative* (working on the same task but taking turns) or *fully collaborative* (working on the same task at the same time) [34] [35]. The higher the level of collaboration – that's our hypothesis –, the higher the level of trust that needs to be gained within the bottom-up processing of gamification. In order to objectify the "level", we want to experimentally apply several credibility-concepts. Methodically we are also aware that a diverse socially embedded trust building process can hardly be totally objectified or controlled in a top-down manner. We explicitly mention that point here in order to highlight that we experimentally follow a tentative approach without falling back into a logical contradiction (claiming bottom-embedding but then falling back into a top-down technocracy). In everyday language: If a user would constantly peek over her shoulder to see if a robot is approaching, this would counteract any gain in productivity and – maybe more importantly – be very uncomfortable for the users themselves. By following the transdisciplinary approach which has been shortly summarized in the first

section we try to methodically challenge the demand of setting ethically reflected conditions for the – per definition – not standardizable practice.

Scenarios of Gamifying Trust

Each of the three human-robot application categories (coexistence, cooperation, and collaboration) [34] [35] is perfectly suited to explore gamification strategies for building trust by a variety of mini games.

In the *coexistence game*, human players are mainly observers. They watch the robots in the production sequence and get explanations along the way in either virtual reality or augmented reality with real robots. Knowledge can be tested by quizzes. The only physical interaction with robots in this stage is when humans accidentally enter working areas or block movement paths, which results in emergency stops. A level of the game could be to find out how close you get to a mobile robot before it stops and re-plans its path. A player who understands that can try to force the robot to move into a certain goal area.

In the *cooperative game*, players are part of the production sequence. They have to learn their production step and receive or pass products to the robot working on adjacent steps. There are defined areas where robot and human tasks overlap, thus are hazardous, but players will soon find out that the robot stops or makes evasive movements before a collision occurs, which it will of course always do, or to guess where the robot is going.

The last game is about *real collaboration*. Robot and human work very closely together, for instance the robot fetches and presents a part and the human mounts another part on top of it. Such a close collaboration requires a tremendous amount of trust. To build this trust, players must be motivated to explore the system, to try to foresee its actions and also test its limits by acting irrational. They will learn about the robot's role in the use-case, the workspaces, where the robot is able and where it is not able to reach, the manipulation abilities of the robots, and the safety measures taken to protect humans.

Sticking to the list of key requirements of trust, they will learn that the robot acts accurately, reliably, and reproducibly, that the system is generally safe, which means that it will take all measures to prevent harm, and that fallback plans are started if necessary, for instance a mobile robot will find a way around a human blocking its path. Trust will also be increased by traceability, for instance realized by an action replay feature. Finally, communication features will offer explanations on every aspect of the robot and its behavior answering every W-question a human might want to raise: who are you, where are we, what are you doing, why are you doing this, when and what is your next step, and so on.

[4] Outlook: Gamification of Trust?

Our hypothesis is that trust in HRI is primarily a contingent bottom-up process, depending on social situations of normative practices. It can hardly be controlled top-down. From an ethical point of view it's important to highlight the fragility of the process itself, the inherent finitude and vulnerability especially of humans but also of the whole form of (human-robot-)interaction. Additionally, trust significantly depends on Annoying Valley Effects – so to say a challenging of error-moments within the trial

and error loop. However, we discuss a combination of gamification and credibility in order to (experimentally) figure out of how to set at least "objective" conditions for influencing the contingent bottom-up processing of trust. Credibility is the application of trustworthiness in technical standards and norms. Trustworthiness is the result of normative ethical reflection concerning the factual trust that has been empirically described in a previous methodical operation. We combine this interrelation between descriptive and normative ethics linked to engineering standards within a feedback-loop model in order to a) illustrate the transdisciplinary approach and b) embed gamification into a broader context of trust and trustworthiness (**Figure 1.**). Our hypothesis develops trust as intimately linked to processes of gaming. Trust is the result of gamification, so to speak. This hypothesis might stand in contrast to other currently discussed approaches, where trust is not seen as the result but as the precondition of gaming [36]. This is one of many possible open questions related to future research. Shorty summarized, we argue that:

1. Gamification of trust in HRI is the processing of trust that is intended to fulfill certain aims (means-end-relation).
2. These processes are both culturally and socially embedded and can fail for several reasons.
3. Trust is gained – from a pragmatic point of view – by the successful repetition of (human) actions. It´s a result of implicit knowledge and trial-and-error feedback loops.
4. Robots are "game changers" since they affect social actions not only by being a classical (handcraft) "tool", but interacting with the human users.
5. In order to follow the aim of gaining trust, interactive processes of human-technology-relations are initiated in a playing-like setting.
6. Gamification in this sense is not supposed to be understood as winning or losing situation. Instead it might include moments of benefit with certain levels or high scores. It´s about the process as such not about winning against a robot or losing against the boss or company.

Follow-up research might take into account:

7. This setting should be free of interests and a struggle for economic benefits – what stands in contrast to the overarching means-end-interpretation. (How is free gaming possible in means-end-oriented techno-economic circumstances?)
8. Three kinds of interaction gaming scenarios can be differentiated: coexistence games, cooperative games and collaborative games. The last one stands for a high significance of trust and most intimate HRI. It relates at the moment primarily to co-bots in industrial environments but bears a high potentiality for social applications in societal contexts.
9. From an engineering point of view, gamification is an alternative way of getting to "know" each other within cooperative processes. This goes along with the user-oriented calibration of robots and profiling of concrete users or groups of users. From an ethical point of view, then, privacy and data security become important normative requirements.
10. From a philosophical – and maybe idealistic point of view – the idea is to humanize HRI by following the *Homo ludens* concept and Friedrich Schiller's credo: "Der Mensch spielt nur, wo er in voller Bedeutung des Worts Mensch

ist, und er ist nur da ganz Mensch, wo er spielt." Playing is seen as the fundamental cultural technique of learning skills. The idea of free playing might stand in contrast to certain economic interests. (How can pure and free playing be performed in an industrial setting?)

11. A certain requirement is risk assessment in order to prevent accidents that go beyond an intended trial-and-error feedback loop. It remains an open question and motivation for further research to find out where exactly the borderline between a productive error and a dangerous accident is situated.

Acknowledgements

This work has been supported by the European Union´s Horizon 2020 research and innovation program under the CSA project "INBOTS – Inclusive Robotics for a better Society" (grant agreement No 780073), by the Austrian Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation and Technology (BMU) under the program "CredRoS – Credible and Safe Robot Systems", and by the Austrian Research Promotion Agency (FFG) under the project "dAIalog.at – Förderung von Fairness und Vertrauen in KI durch Formate der partizipativen Technikgestaltung". We also want to express our thankfulness to the three anonymous reviewers of an earlier version of this paper, who provided critical remarks and helped to improve the quality.

References

[1] V. Dignum, F. Dignum, J. Vazquez-Salceda, A. Clodic, M. Gentile, S. Mascarenhas & A. Augello, "Design for Values for Social Robot Architectures" in: M. Coeckelbergh, J. Loh, M. Funk, J. Seibt & M. Nørskov (eds.), *Envisioning Robots in Society – Power, Politics, and Public Space*, IOS Press, Amsterdam a.o., 2018, 43–52.

[2] C. Ess, "Trust, social identity, and computation", in: R. Harper (ed.), *The Complexity of Trust, Computing, and Society*, Cambridge University Press, Cambridge, 2014, 199–226.

[3] P.A. Hancock, D.R. Billings, K.E. Schaefer, JY Chen, E.J. de Visser & R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction" in: *Human Factors* **53(5)** (2011), 517–527.

[4] M. Lewis, K. Sycara & P. Walker, "The Role of Trust in Human-Robot Interaction" in: H.A. Abbass, J. Scholz & D.J. Reid (eds.), *Foundations of Trusted Autonomy*, Springer, Cham, 2018, 135–159.

[5] M. Salem, G. Lakatos, F. Amirabdollahian & K. Dautenhahn, "Towards Safe and Trustworthy Social Robots: Ethical Challenges and Practical Issues" in: *ICSR*, 2015, 584–593.

[6] K.E. Schaefer, J.Y.C. Chen, J.L. Szalmaand & P.A. Hancock, "A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems" in: *Human Factors* **58(3)** (2016), 377–400.

[7] S. Tolmeijer, A. Weiss, M. Hanheide, F. Lindner, T.M. Powers, C. Dixon & M.L. Tielman, "Taxonomy of Trust-Relevant Failures and Mitigation Strategies", *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction. 2020*; https://dl.acm.org/doi/pdf/10.1145/3319502.3374793 (accessed 2020 Apr 23)

[8] J. Seibt, "Introduction" in: J. Seibt, R. Hakli & M. Nørskov (eds.), *Sociable Robots and the Future of Social Relations*, IOS Press, Amsterdam a.o., 2014, vii–viii.

[9] M. Nørskov, "Editor´s Preface" in: M. Nørskov (ed.), *Social Robots. Boundaries, Potential, Challenges*, Ashgate, Farnham & Burlington, 2016, xv–xxii.

[10] J. Seibt, "Classifying Forms and Modes of Co-Working in the Ontology of Asymmetric Social Interactions (OASIS)" in: M. Coeckelbergh, J. Loh, M. Funk, J. Seibt & M. Nørskov (eds.), *Envisioning Robots in Society – Power, Politics, and Public Space*, IOS Press, Amsterdam a.o., 2018, 133–146.

[11] J. Seibt, G. Borggreen, K. Fischer, C. Hasse, H.-Y. Liu, & M. Nørskov, "Working with and Alongside Robots: Forms and Modes of Co-Working" in: M. Coeckelbergh, J. Loh, M. Funk, J. Seibt & M.

Nørskov (eds.), *Envisioning Robots in Society – Power, Politics, and Public Space*, IOS Press, Amsterdam a.o., 2018, 128–132.
[12] M. Coeckelbergh, *Growing Moral Relations. Critique of Moral Status Ascription*, palgrave macmillan, New York, 2012.
[13] D. Ihde, *Postphenomenology and Technoscience. The Peking University Lectures*, State University of New York Press, Albany NY.
[14] G. Ropohl, *Ethik und Technikbewertung*, Suhrkamp, Frankfurt a.M., 2016.
[15] P. Janich, *Was ist Information? Kritik einer Legende*, Suhrkamp, Frankfurt a. M., 2006.
[16] G.W. Leibniz, *Monadologie*, Reclam, Suttgart, 2012.
[17] I. Kant, *Kritik der reinen Vernunft 1. Hg. W. Weischedel*, Suhrkamp, Frankfurt a.M., 1974.
[18] O. Höffe, *Kants Kritik der praktischen Vernunft. Eine Philosophie der Freiheit*, C.H. Beck, München, 2012.
[19] M. Funk, *Roboter- und Drohnenethik. Eine methodische Einführung*, Springer, Wiesbaden, 2020.
[20] P. Janich, *Kultur und Methode, Philosophie in einer wissenschaftlich geprägten Welt*, Suhrkamp, Frankfurt a.M., 2006.
[21] M. Mori, *The Uncanny Valley: The Original Essay by Masahiro Mori. Translated by Karl F. MacDorman and Norri Kageki*, IEEE Spectrum, June 12, 2012; https://spectrum.ieee.org/automaton/robotics/humanoids/the-uncanny-valley (accessed 2020 Apr 23)
[22] T.L. Beauchamp & J.F. Childress, *Principles of Biomedical Ethics*, Oxford University Press, Oxford, 2001.
[23] A. Jonsen & S. Toulmin, *The Abuse of Casuistry*, University of Chicago Press, Berkeley, 1988.
[24] M. Coeckelbergh & M. Funk, "Wittgenstein as a Philosopher of Technology: Tool Use, Forms of Life, Technique, and a Transcendental Argument" in: *Hum Stud* **41** (2018), 165–191.
[25] M. Funk, "Repeatability and Methodical Actions in Uncertain Situations: Wittgenstein's Philosophy of Technology and Language" in: *Techné: Research in Philosophy and Technology. Special Issue* **22(3)** (2018), 351–376; (DOI) 10.5840/techne201812388 (accessed 2020 Apr 23)
[26] M. Coeckelbergh, *Using Words and Things. Language and Philosophy of Technology*, Routledge, New York & London, 2017.
[27] T. Rentsch, *Heidegger und Wittgenstein. Existential- und Sprachanalysen zu den Grundlagen philosophischer Anthropologie*, Klett-Cotta, Stuttgart, 2003.
[28] M. Coeckelbergh, *Human Being @ Risk. Enhancement, Technology, and the Evaluation of Vulnerability Transformations*, Springer, Dordrecht, 2013.
[29] R. Buchner, D. Wurhofer, A. Weiss & M. Tscheligi, "Robots in Time: How User Experience in Human-Robot Interaction Changes over Time" in: G. Herrmann, M.J. Pearson, A. Lenz, P. Bremner, A. Spiers & U. Leonards (eds.), *Social Robotics. 5th International Conference, ICSR 2013*, Springer, Cham, 138–147.
[30] G. Hardstone, L. d'Adderio & R. Williams, "Standardization, Trust and Dependability" in: K. Clarke, G. Hardstone, M. Rouncefield & I. Sommerville (eds.), *Trust in Technology: A Socio-Technical Perspective*, Springer, Dordrecht, 2006, 69–103.
[31] K. Stewart, "Trust Transfer on the World Wide Web" in: *Org. Sci.* **14(1)** (2003), 5–17.
[32] D. Belanche Gracia, L.V. Casalo, C. Flavián, & J.J.L. Schepers, "Trust Transfer in the Continued Usage of Public E-Services" in: *Information and Management* **51(6)** (2014), 627–640.
[33] S. Stadler, A. Weiss & M. Tscheligi, "I Trained this Robot: The Impact of Pre-Experience and Execution Behavior on Robot Teachers" in: *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, Edinburgh, 2014, 1030–1036.
[34] A. Kolbeinsson, E. Lagerstedt & J. Lindblom, "Foundation for a Classification of Collaboration Levels for Human-Robot Cooperation in Manufacturing" in: *Production & Manufacturing Research* **7(1)** (2019), 448–471.
[35] IFR, *Demystifying Collaborative Industrial Robots. Positioning Paper. December 2018*, International Federation of Robotics, Frankfurt a.M.; https://www.ifr.org/downloads/papers/IFR_Demystifying_Collaborative_Robots.pdf (accessed 2020 Apr 23)
[36] M. Barlow, "Trusted Autonomous Game Play" in: H.A. Abbass, J. Scholz, D.J. Reid (eds.), *Foundations of T*