Avoiding Undertrust and Overtrust

Hermann Kaindl
Institute of Computer Technology
TU Wien
Vienna, Austria
hermann.kaindl@tuwien.ac.at

Davor Svetinovic
Center on Cyber-Physical Systems
Department of Computer Science
Khalifa University of Science and Technology
Abu Dhabi, UAE
davor.svetinovic@ku.ac.ae

Abstract

An important question with regard to new systems is whether they can be trusted by human users, especially when they are safety-critical. This problem should already be addressed in the course of requirements engineering for such systems. However, *trust* is actually a complex psychological and sociological concept. Thus, it cannot simply be taken into account as a desired or needed property of a system. We propose to learn from the *Human Factors* discipline on trust in technical systems. In particular, we argue that both *undertrust* and *overtrust* need to be avoided. The challenge is to determine system properties and activities in the course of requirements engineering for achieving that. We conjecture that both actual properties like safety and subjective assessment like perceived safety will be important, and how they will have to be balanced for avoiding undertrust and overtrust.

1 The Problem

Even if a system has been shown to be *safe* (at a defined level and with defined probabilities), it is not for granted that users would trust it with regard to safety. Trust is a complex psychological and sociological concept, and trustworthiness of a trustee is a major aspect of establishing trust [HB15]. When the trustee is a (semi-)autonomous system, both *undertrust* and *overtrust* may have severe consequences. As illustrated in Figure 1, undertrust with regard to safety of a system means that the perceived safety is lower than the actual safety. Conversely, overtrust means that the perceived safety is higher than the actual safety. Ideally, the perceived safety would be as high as the actual safety.

For instance, the captain of the Costa Concordia in the famous cruise ship disaster that killed 32 passengers in 2012 is said to have undertrusted the ship's navigation system in favor of manual control according to [HB15]. Investigations discovered that the captain diverged from the ship's computer-programmed route before hitting the shallow reef that caused the sinking.

Overtrust in automation may have contributed, e.g., to the crash of a Turkish Airlines flight in 2009 according to [HB15], which killed nine people, including all three pilots. It was partially caused by the pilots' continued reliance on the plane's automatic pilot after an altitude-measuring instrument failed.

The Automated Driving Roadmap [ES17] provides an overview on the status of automated driving technologies with regard to implementation in Europe. Its overall objective is to identify challenges for implementation of

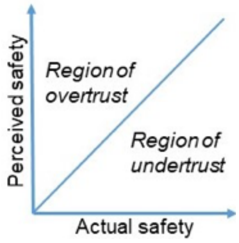

Figure 1: Undertrust vs. overtrust in case of safety; inspired by [MW07a, Figure 1] (showing the case of reliability), which was modified from [WGM00]

higher levels of automated driving functions. Regarding users' and societal acceptance, one of the mentioned challenges is trust. There are no hints, however, on how to address it.

At the recent RE'18 conference, Cysneiros et al. [CRdPL18] suggested that self-driving cars that demonstrate transparency in their operations will increase consumer trust, and they investigated transparency as a Non-Functional Requirement. However, they did not take the work on trust in automation by the Human Factors community into account, which applies psychological and physiological principles to the (engineering and) design of products, processes, and systems. The goal of Human Factors is to reduce human error, increase productivity, and enhance safety and comfort with a specific focus on the interaction between the human and the thing of interest, see `https://en.wikipedia.org/wiki/Human_factors_and_ergonomics`.

For a better understanding of the issues involved in requirements engineering regarding trust, we propose to learn from the field of Human Factors.

2 Previous Work on Human Factors

Hence, let us sketch some previous work on Human Factors that may inform future work on requirements engineering. The significance of trust is not limited to the interpersonal domain; trust can also define the way people interact with technology. Thus, the concept of trust in automation has been the focus of substantial research over the past several decades.

In their review of related work after 2002, Hoff and Bashir [HB15] surveyed many studies where trust in automation was transferred from trust relationships between humans. The truster is human, the trustee is an automated system.

Earlier work on similarities and differences between human–human and human–automation trust has been reviewed already in [MW07b]. In a nutshell, a major insight appears to be that *anthropomorphizing*, i.e., ascribing human features to an automated system, reduces the differences. This is confirmed by a more recent study with potential *(end-)users* of autonomously driving cars, which assumed that trust is directly linked to *perceived* safety [HvBPB17]. Using a car simulator without any aspect of actual safety involved, they showed that trust in a self-driving car was directly linked in their human factors study to the *perceived* safety, which was highest in the case of anthropomorphic visualization through a chauffeur avatar. It reacts to the same events as a world in miniature visualization that presents the car's perception of the surroundings, its interpretation and its actions, but is more human-like and potentially associated with more feelings.

Wang et al. [WJH09] studied the appropriateness of reliance depending on "reliability disclosure," which refers to whether participants were explicitly informed of the reliability of the feedback (as estimated by the system itself). More precisely, they provided *meta-information* on the reliability of feedback together with the feedback itself. Overall, the informed group seemed to resolve the changing reliability of the feedback better than the uninformed group. Instead of verbally informing, Neyedli et al. [NHJ11] studied displaying this meta-information in different ways, pie displays and mesh displays, respectively (in real time). They showed that this can change reliance on the automation.

3 The Challenge for Requirements Engineering

We conjecture that both actual properties like safety and subjective assessment like perceived safety will be important, and how they will have to be balanced for avoiding undertrust and overtrust. The challenge for requirements engineering is to determine system properties and activities in the course of requirements engineering for achieving a good balance between actual and perceived safety.

As illustrated in Figure 1, this means to avoid both the region of undertrust and the region of overtrust. An open question is what can be done to avoid undertrust so that, e.g., a safe system (with a very low probability of causing injuries or even fatalities) would not be used by people because they do not trust it. Conversely, another open question is what can be done to avoid overtrust so that, e.g., drivers using current autopilots of cars would not simply hand-over control but still take care of the traffic situation and stay ready for taking over.

We propose to study these questions by considering objectively determined properties like safety and to determine adequate anthropomorphism for disclosing its level. For evaluations, studies like those established in the field of Human Factors will most likely be necessary for assessing the balance.

However, based on the very recent observations in [FWCR19], it seems we cannot just rely upon the social sciences to take into account growing computer science research. It is critical for us to phrase the engineering questions in a way that they can be included in social science research, and to find a way of incorporating the results in the engineering solutions.

Besides the explicit relationship between safety and trust in safety-critical systems, the importance of trust is becoming more and more evident in *privacy* and *security-critical* systems. For example, it was observed that privacy violations are having significant impact on trust [Mar18], and the security of the online systems is a necessary prerequisite for trust in such systems. This is especially evident in a new generation of blockchain-based systems [ARS18].

In order to achieve a proper understanding of how to handle undertrust and overtrust, we will have to incorporate studies with humans into requirements engineering using a combination of standard social science research methods: case study, survey, observational, correlational, experimental, and cross-cultural methods. However, in order for these studies to be effective, it will be of critical importance that we learn how to pose the research questions in a way that the results of the social studies can be engineered into technical systems to be trusted.

Finally, let us propose a short list of research questions. We believe it will be of critical importance to handle properly both methodological and technical questions. The most pressing methodological questions that we see are:

- How do we map research questions and methods from Requirements Engineering to Human Factors, and vice versa?
- How do we make engineering questions properly included in Human Factors research, and how can the produced results be effectively engineered into the systems to be trusted?
- How do we apply computational social science methods in both Requirements Engineering and Human Factors?

The technical questions regarding trust, undertrust, and overtrust in Requirements Engineering will become even more pressing and probably best tackled in the context of emerging application areas of Artificial Intellgence, autonomous adaptive systems, and blockchain:

- How do we relate trust/undertrust/overtrust and non-functional requirements?
- How do we relate trust regarding safety/security/privacy?
- How do we develop evaluation techniques for trust regarding actual vs. perceived safety?
- How do we reduce trust manipulations in autonomous adaptive systems through techniques such as anthropomorphization?
- How do we define trust in fully open decentralized blockchain systems that store data and run code supplied by any anonymous entity?
- How do we ensure ongoing trust in evolving systems with emerging behavior and machine learning?

References

[ARS18] Israa Alqassem, Iyad Rahwan, and Davor Svetinovic. The anti-social system properties: Bitcoin network data analysis. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2018.

[CRdPL18] Luiz Marcio Cysneiros, Majid Raffi, and Julio Cesar Sampaio do Prado Leite. Software transparency as a key requirement for self-driving cars. In *2018 IEEE 26th International Requirements Engineering Conference (RE)*, pages 382–387. IEEE, 2018.

[ES17] EPoSS ERTRAC and ETIP SNET. Ertrac automated driving roadmap. *ERTRAC Working Group*, 7, 2017.

[FWCR19] Morgan R. Frank, Dashun Wang, Manuel Cebrian, and Iyad Rahwan. The evolution of citation graphs in artificial intelligence research. *Nature Machine Intelligence*, 1(2):79–85, 2019.

[HB15] Kevin Anthony Hoff and Masooda Bashir. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3):407–434, 2015.

[HvBPB17] Renate Häuslschmid, Max von Buelow, Bastian Pfleging, and Andreas Butz. Supportingtrust in autonomous driving. In *Proceedings of the 22nd international conference on intelligent user interfaces*, pages 319–329. ACM, 2017.

[Mar18] Kirsten Martin. The penalty for privacy violations: How privacy violations impact trust online. *Journal of Business Research*, 82:103 – 116, 2018.

[MW07a] P. Madhavan and D. A. Wiegmann. Similarities and differences between human–human and human–automation trust: an integrative review. *Theoretical Issues in Ergonomics Science*, 8(4):277–301, 2007.

[MW07b] Poornima Madhavan and Douglas A Wiegmann. Similarities and differences between human–human and human–automation trust: an integrative review. *Theoretical Issues in Ergonomics Science*, 8(4):277–301, 2007.

[NHJ11] Heather F Neyedli, Justin G Hollands, and Greg A Jamieson. Beyond identity: Incorporating system reliability information into an automated combat identification system. *Human factors*, 53(4):338–355, 2011.

[WGM00] Christopher D. Wickens, Keith Gempler, and M. Ephimia Morphew. Workload and reliability of predictor displays in aircraft traffic avoidance. *Transportation Human Factors*, 2(2):99–126, 2000.

[WJH09] Lu Wang, Greg A Jamieson, and Justin G Hollands. Trust and reliance on an automated combat identification system. *Human factors*, 51(3):281–291, 2009.