Universität Paderborn

Fakultät für Kulturwissenschaften

Healthy Distrust in the Context of Artificial Intelligence

Dissertation zur Erlangung des akademischen Grades Doktor der Philosophie (Dr. phil.)
im Fach Psychologie der Universität Paderborn

von

Tobias Martin Peters

Betreuerin

Prof. Dr. Ingrid Scharlau

Hiermit erkläre ich gemäß §12 der Promotionsordnung:

- dass die vorgelegte Arbeit selbstständig und ohne Benutzung anderer als der in der Arbeit angegebenen Hilfsmittel angefertigt wurde;
- dass die Arbeit bisher weder im In- noch Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt wurde;
- dass ich weder früher noch gleichzeitig ein Promotionsverfahren bei einer anderen Hochschule oder bei einer anderen Fakultät beantragt habe.

Paderborn, 19. September 2025

Unterschrift

Acknowledgments

First of all, I would like to thank Prof. Ingrid Scharlau. I would like to thank her for her advice, her support, her thoughtful input, her critical questions and, in general, for our talks and discussions. She has significantly supported and encouraged me in my thinking, writing, and academic work, and for that I am very grateful! In addition, you have broadened my perspective on a wide range of topics, not only for the 7 Deadly Sins of Psychology or the distinction between concepts and constructs, but also for mathematical proofs in cake form or at least the basics of distinguishing between yellow composites – thank you!

I would also like to thank my mother and father. Mama, Papa, I owe you both a great deal; you have had a positive influence on me and always supported me wherever you could, and I really appreciate that – thank you! Thank you also to my dear big sister and dear big brother. You have always been there for me and continue to be there – thank you!

To my friends – whether from my school days, from Hösel, Eggerscheidt, Homberg, Heiligenhaus, or elsewhere, from my bachelor's degree days in Osnabrück, from the Bumblesquad or the Rasselbundle – thank you for the countless wonderful, fun memories and ever new, great moments. Spending time with you and having your support has been and continues to be very valuable to me – thank you!

To my colleagues Chi, Kai, Philip, and Michael, thank you for all the conversations, whether content-wise or just 'Unsinn', or the occasional 'Trost' over lunch, coffee, or in between. Also, thank you to all the other ZM2 inhabitants and TRR colleagues for the time we spent together! Many thanks also to the SHKs and WHBs, namely Valeska Behr, Oliver Debernitz, Konstantin Gubenko, Fritz Lange, and Michael Neuss, who have supported me and worked with me over the last few years – thank you!

Furthermore, and a bit more formal, I gratefully acknowledge funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) for grant TRR

318/1 2021 – 438445824. I also acknowledge support for the planned publication costs of Manuscript 4 by the Open Access Publication Fund of Paderborn University.

And last but not at all least: thank you Hannah, for all your listening, your support, your advice, and for the occasional “am Kopf rütteln”. Without you, the last few years that led to this dissertation would not have been possible – thank you!

Artificial Intelligence, Healthy Distrust & Explainability

While writing Manuscript 3 of this dissertation, I was in need of a synonym for the term 'aligns well'. Following my typical approach, I used a search engine for quick inspiration. My search *aligns well synonym* gave me the result I was looking for, namely the link to some online thesaurus listing multiple alternatives. Above this link, the recently added feature 'AI-overview' appeared. This overview also provided an answer. However, instead of synonyms, I received multiple translations of the term into German in the form of a nicely worded summary, and thus, not at all what I was looking for.[1]

Why do I bother with this arguably mundane example of AI failure to begin my dissertation? I could open with lawyers who used ChatGPT without ensuring that the cited cases exist, biased crime forecasts, or fatalities due to autonomous vehicles. However, I deliberately picked this example because it touches upon a sentiment that I, among other points, want to convey in this first section that introduces my topic — healthy distrust in the context of artificial intelligence (AI). For this introduction, let us first turn to the context of AI.

Artificial intelligence can be defined as a system that is capable of solving tasks and problems, which are typically considered to require human intelligence, and that operates with some degree of autonomy and may exhibit adaptiveness (Regulation (EU) 2024/1689, 2024; Sheikh et al., 2023). Such AI systems are nowadays applied in numerous fields and have already transformed certain everyday tasks. While the progress in recent years is astounding, current and foreseeable AI technologies remain fallible. Failures remain that range from AI errors of grave consequences (the examples I did not pick) to blatantly obvious errors of little consequence, such as providing translations instead of synonyms. The former examples illustrate the importance of holding a healthy degree of distrust when interacting with AI. Otherwise, blindly trusting AI can lead to lawyers being fined, racial profiling, the endangerment of pedestrians, and many other problems. The latter example

[1] This was reproducible months later, see Figure A1 in the Appendix, 22.08.2025.

points to a benefit of healthy distrust towards AI in general, and to elaborate on this, allow me to digress from the scientific content of the following chapters to a more general discussion of AI.

In the last decade, scientific and public interest in artificial intelligence surged. Since the broader public was introduced to ChatGPT, it is even applicable to speak of a hype. The term and interest in AI are by no means novel. Already in the 1960s-70s or in the 90s, a surge of interest in and research on AI can be observed. Yet, periods of interest were also followed by a decline in interest, referred to as AI winter. For a nice history and overview of the developments, turn to Haenlein and Kaplan (2019). Interestingly, they state that “[i]n regular intervals since the 1950s, experts predicted that it will only take a few years until we reach Artificial General Intelligence—systems [...].” (p. 6) — a prediction that is also common nowadays.

Whether or not another AI winter is coming remains open. For instance, it has been called into question whether the developmental pace of the last few years will be maintained, as recently indicated by the release of OpenAI’s latest model and the accompanying disappointment (Newport, 2025). Nonetheless, the successes of the current AI phase are likely to persist rather than disappear because they have already changed and will continue to change the way we use technology. For instance, LLM-based applications have problems with being up to date and providing actual references. This has improved with technological solutions like RAG. However, while RAG can make so-called hallucinations less likely (Shuster et al., 2021), they can also lead to different forms of hallucinations or other issues (Wu et al., 2024). Although these and other problems remain, LLM-based applications have changed, for example, internet search in two ways: people using ChatGPT instead of traditional search engines, and search engines following ChatGPT’s example and introducing their own LLM-based features. Such deployment of LLM-based applications can be highly beneficial. However, they remain fallible, and looping back to my initial example, they can also be unnecessary and even less useful than previous technical solutions.

Given that AI solutions are not without problems, this brings me back to healthy distrust towards AI in general. Although not central to this dissertation, the development of AI, related technologies, and infrastructure is also a subject of criticism from different directions. This, among others, includes criticism towards the energy consumption of AI, and specifically of generative AI (Bashir et al., 2024), outsourced human labor for annotating and filtering problematic AI-generated text under questionable circumstances (Perrigo, 2023), the danger of huge investments without predictable downstream revenues (Newport, 2025), and the potential to spread dis- or misinformation (Chen & Shu, 2024).

Moreover, especially in the realm of education, outsourcing, for example, coursework to LLM-based applications runs the danger of diminishing the educational and learning effects of engaging with a topic. In a recent preprint (Kosmyna et al., 2025), university students who used LLM-based tools for an essay assignment performed worse in multiple measures than students who were permitted to use search engines or no tools at all. For all these reasons, I see benefits not only in healthy distrust during the interaction with AI but also in healthy distrust in using AI. The latter means two things: critically reflecting on the decision to use existing AI for certain tasks and the (developers') decision to apply AI-based solutions to certain tasks. While I will not continue in this direction and cannot offer any further insights for improvement, I wanted to at least highlight these issues. I regard them as pressing and therefore did not want to completely gloss over them. Furthermore, I will briefly return to them in my conclusion.

Besides these more general issues of AI, as mentioned above, individual AI systems each have their own technical shortcomings and pitfalls. For instance, so-called adversarial examples (e.g., Zhang & Li, 2019) illustrate the fallibility of AI and how different these failures can be compared to areas where humans tend to make errors. The aspect that AI remains fallible in combination with the increased application of AI to so-called high-stakes domains (e.g., medicine, law, finance) has reinforced different concerns. As discussed in

Manuscripts 1 and 2, within current AI research, recurring goals are to foster AI's fairness, accountability, interpretability, and transparency (Barredo Arrieta et al., 2020; Guidotti et al., 2019; Mohseni et al., 2021). To ensure that such goals are met, numerous guidelines and regulations have been formulated (Thiebes et al., 2021). Among those is also the EU AI Act (Regulation (EU) 2024/1689, 2024), which is currently considered the most comprehensive AI regulation and subsumes such concerns under the term of trustworthy AI. One prominent way to foster and ensure AI's trustworthiness is the usage and development of explainable AI (XAI). XAI or explainability methods can be understood as any means to allow users to better understand AI output. This typically does not only include textual or verbal explanations but also, e.g., visual highlighting via heatmaps in the case of image-based decisions or ranking of most relevant features, e.g., in the case of credit scoring. These or other forms of XAI methods are often considered to improve users' trust. This assumption was coined as the explainability-trust hypothesis (Kastner et al., 2021). Interestingly, more cautious formulations of this hypothesis are also reported. Instead of assuming an improvement of trust, they assume an improvement of appropriate trust.

Appropriate trust is related to healthy distrust, but as I detail in the following, it is not the same. So far, the explicit notion of healthy distrust does not exist in current AI literature, with a recent exception (Paaßen et al., 2025). While the research on appropriate trust shares similar motivations and is relevant to healthy distrust, it also lacks certain conceptual aspects and remains vague in crucial areas that I hope to concretize with the notion of healthy distrust.

Thus, the overarching topic of my dissertation is to approach the notion of healthy distrust. This means that my contributions take first steps towards understanding, investigating, and fostering healthy distrust. The following is structured by the chronology of the included manuscripts. Fittingly, this also means that I will start with the theoretical and conceptual basis of healthy distrust. I will then discuss recent investigations of related concepts, transition to my own empirical investigation, and finish with some concluding remarks. Throughout this, I identify and discuss shortcomings and limitations of the manuscripts and highlight connections between the individual manuscripts.

Manuscript 1 and Manuscript 2 combine multiple strands of research to approach healthy distrust from existing theoretical and empirical work. Manuscript 1 is focused on the theoretical and conceptual aspects revolving around healthy distrust. Manuscript 2 reiterates the same thoughts in an abbreviated form, combines them with further notions typically encountered in the context of (X)AI, and on the grounds of this theoretical basis, gives an overview of existing empirical research. Manuscript 3 takes on typical approaches to assess (dis)trust in human-AI interactions, namely self-reported trust and reliance, and refines them. Moreover, with the chosen experimental scenarios, I created a situation where healthy distrust is expected to occur and tested whether an instruction to distrust can improve participants' performance. Manuscript 4 extends the typical assessment approaches by introducing the perspective of visual attention as an indicator of healthy distrust. Importantly, in this paper, I also discuss the complexity of the aim of healthy distrust and related notions like appropriate trust.

Table 1
Information on the manuscripts included in the dissertation.

Manuscript 1*	**Peters, T. M.**, & Visser, R. W. (2023). The importance of distrust in AI.	published in Communications in Computer and Information Science (xAI 2023)
Manuscript 2*	Visser, R., **Peters, T. M.**, Scharlau, I., & Hammer, B. (2025) Trust, distrust, and appropriate reliance in (X)AI: A conceptual clarification of user trust and survey of its empirical evaluation.	published in Cognitive Systems Research
Manuscript 3	**Peters, T. M.**, & Scharlau, I. (2025). Interacting with fallible AI: Is distrust helpful when receiving AI misclassifications?.	published in Frontiers in Psychology
Manuscript 4	**Peters, T. M.**, Biermeier, K., & Scharlau, I. (2026) Assessing healthy distrust in human-AI interaction: Interpreting changes in visual attention.	published in Frontiers in Psychology

* The first two authors contributed equally to the manuscript and share first authorship.

- Manuscript 1
  - **TP**: Writing – review & editing, Writing – original draft, Visualization, Conceptualization. **RV**: Writing - review & editing, Conceptualization.
  - The majority of the original draft was written by TP; the contribution's content (on the underlying desideratum of appropriate trust) was extensively discussed with RV and was developed together. Both authors undertook extensive rewrites of the paper beyond the typical amount of review and editing.
- Manuscript 2
  - **RV**: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **TP**: Writing – review & editing, Writing – original draft, Visualization, Validation, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **IS**: Writing – review & editing, Validation, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization. **BH**: Writing – review & editing, Writing – original draft, Supervision, Project administration, Funding acquisition, Formal analysis, Conceptualization.
  - Both co-authors wrote parts of the original draft. The original draft of Sections 3, 4, 6.5, and 7 was written by TP; the original draft of Sections 2, 5, and

6.1-6.4 was written by RV; the original draft of Sections 1 and 8 was co-written by TP and RV. Each Section was reviewed and edited by all authors. RV conducted the literature search of the empirical overview. TP analyzed the papers included in this overview in terms of trust effects, types of trust measurement, and the conceptual consideration concerning trust and distrust. RV analyzed the papers included in this overview in terms of applied AI model, domain, XAI method, and evaluation criteria other than trust.

- Manuscript 3
  - **TP**: Conceptualization, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **IS**: Conceptualization, Funding acquisition, Resources, Supervision, Writing – review & editing.
- Manuscript 4
  - **TP**: Writing - original draft, Writing - review & editing, Methodology, Conceptualization, Investigation, Formal analysis, Software, Visualization. **KB**: Writing - review & editing, Methodology, Formal analysis, Software. **IS**: Supervision, Conceptualization, Writing – review & editing, Funding acquisition.
  - Original draft by TP, except original draft Subsection TVA-TOJ model within Section Materials and Methods by KB. Drafting the model in pymc and the implementation of the power analysis were conducted by KB. TP used and adapted this model according to the statistical comparisons he carried out.

Understanding Healthy Distrust

The first two manuscripts of this cumulative dissertation approach the notion of healthy distrust in AI on a theoretical and conceptual basis by discussing and summarizing relevant existing research. At the very beginning of research for these two manuscripts stood the idea to provide a comprehensive and interdisciplinary overview of distrust in the context of AI from the two perspectives of Machine Learning and Psychology. While my co-authors and I did approach the topic from these two perspectives, we ultimately did not focus only on distrust but also on trust. We did so because contemporary AI research was mainly investigating trust, and also, psychological research had an emphasis on studying trust. Given that this emerged as a problematic omission, this neglect of distrust became a core topic of the manuscripts.

Similar to the introductory chapter, Manuscripts 1 and 2 identify the concerns about AI that led to regulations for and interest in trustworthy AI and connect this to the approach of developing methods of XAI. Within this context, a prominent aim is the notion of appropriate trust. Appropriate trust and related concepts with differing terms, e.g., warranted trust or calibrated trust, are meant as a distinction from fostering trust because fostering trust in all cases would allow for blind trust.

This means that appropriate trust and related concepts are similar to the notion of healthy distrust. While we concluded that healthy distrust shares the same underlying aim as notions like appropriate trust, we did not equate them. Instead, we advocated for considering and fostering both appropriate trust and healthy distrust. For reaching this conclusion, let us first turn to the concepts of trust. The attentive reader may wonder why I do not turn to distrust, but this is part of the problem I am building up to.

In his influential work, sociologist Luhmann (2009) describes trust as a basic element of social life, which would not function without it. He describes trust as a necessary mechanism to reduce the complexity of interpersonal interactions, and together with other earlier work (e.g., Deutsch, 1958; Rotter, 1967, 1971), this forms the basis for subsequent, influential models of trust (and distrust). Two of these models are relevant throughout Manuscripts 1 and 2. The first is the Integrative Model of Organizational Trust by Mayer et al. (1995), and the second is the model by Lewicki et al. (1998). Both are important conceptualizations of trust (and distrust) and offer key insights.

Mayer et al. (1995) integrated the previous trust research and proposed a unified model of trust. In doing so, they clearly differentiate between trust and trustworthiness, discuss the role of uncertainty and risk, and distinguish trust from related concepts. These insights and distinctions were not new, but integrating them in this comprehensive manner was likely the reason why this model had a profound influence on subsequent research on trust. In this model, trust and distrust are understood as opposing ends of a single dimension.

In contrast to this, the main argument of Lewicki et al. (1998) was to understand distrust differently. They proposed trust and distrust as two related yet separate dimensions. Thereby, they can co-exist and are defined by different characteristics. In this conceptualization, trust and distrust are generally understood in a negatively related way (high trust often goes along with low distrust, and vice versa), but situations occur in which medium levels of the two co-exist. Importantly, this means that when measuring trust and distrust (e.g., via self-report), they should be measured as two dimensions. This also finds support in factor analyses showing superiority of the two-dimensional model (Scharowski et al., 2025; Spain et al., 2008).

At this point, it should be mentioned that there is also another, intermediate conceptualization that regards trust and distrust as one continuum but with a qualitatively meaningful middle that represents the state of suspicion (Basel & Brühl, 2024). This form of conceptualizing is neglected in Manuscripts 1 and 2. However, given that the model by Mayer et al. (1995) is the most common conceptualization in the context of automation and AI, and Basel and Brühl (2024) identify the two-dimensional concept as the most fitting, these overlooked concepts are unproblematic for the argumentation of the manuscripts. This third type of conceptualization is an interesting alternative conceptualization that should be considered in further research. Still, we again turn to the models mentioned above.

These models' insights are core contributions to the context of AI. It is not that each of these points is unheard of, but given that, as discussed in the manuscripts, research still does not consider them sufficiently, Manuscript 1 argues for considering trust and distrust in the context of AI. Given the two-dimensional concept of trust and distrust, Manuscript 1 does not call for replacing appropriate trust with healthy distrust but rather advocates appropriate trust combined with a healthy amount of distrust. Manuscript 1 identifies three main benefits of this. Firstly, explicitly naming both trust and distrust is a more accurate reflection of the two-dimensional conceptualization. Secondly, translating this concept to the assessment of users then allows for properly quantifying trust and distrust, instead of only quantifying trust. Thirdly, beyond explicitly aiming for trust and distrust, it also includes the term healthy. While we could have picked *appropriate* instead of the term *healthy*, to simply call for appropriate trust and distrust, we did not. To use this shorter description would not be wrong, but by adding healthy to distrust, a connection is made to the positive effects that distrust can have. This aims to break up the typical negative connotation of distrust, which should be overcome (Basel & Brühl, 2024; Guo et al., 2017; Lewicki et al., 1998; Mühlfried, 2018). Otherwise, a negative view of distrust contributes to a neglect of its beneficial characteristics, for which Manuscripts 1 and 2 provide multiple examples.

These observations and their foundation are discussed in Manuscript 1. As already mentioned, Manuscript 2 reiterates the same argumentation in an abbreviated form and connects it with the distinction between actual and perceived trustworthiness. Furthermore, Manuscript 2 provides an overview of the empirical investigation of trust and distrust in the (X)AI context[2]. Multiple issues contributed to the need for such an

[2] The parentheses in '(X)AI' reflect that Manuscript 2 combines insights from the contexts of AI and XAI.

overview. The main driver for the conceptual overview was the neglect of distrust and the conflation of insights from (dis)trust research and common-sense reasoning about (dis)trust, and thereby inconsistencies in the employed terminology. The primary driver for the empirical overview was the vast amount of different approaches, operationalizations, and manipulations in existing studies on trust in (X)AI, thereby necessitating an intermediate resumée to provide future research with an improved starting point.

Manuscript 2 offers a comprehensive conceptualization of trust, distrust, and appropriate reliance. It does so by clearly disentangling trust from trustworthiness and separating (dis)trust from reliance. This was needed because some work did not sufficiently distinguish between trustworthiness, the property of the system in which trust is placed, and the behavioral consequence of trust, namely, reliance. Moreover, we identified a multitude of influential factors of trust in the existing literature, which we organized and categorized. To offer an improved starting point for future research on trust and distrust in AI, Manuscript 2 lists a number of takeaways. They are related to the conceptualization and measurement of trust and distrust, to the consideration of appropriate trust, healthy distrust, and appropriate reliance, and to the effects of (X)AI systems' design.

Additionally, Manuscripts 1 and 2 make a connection to the already mentioned explainability-trust hypothesis. In XAI research, it is a common assumption that XAI methods lead to an increase in trust, or, if formulated more cautiously, to an increase in appropriate trust (Kastner et al., 2021). Manuscripts 1 and 2 highlight that explanation can identify reasons to trust and reasons to distrust. This is at least acknowledged when formulating the explainability-trust hypothesis cautiously, and therefore, the cautious formulations should be preferred. If explanations would simply foster trust, they would convince users to trust instead of improving their decision on when to trust and when to distrust.

Arguably, when researchers state that explanations should foster trust, they most likely do not mean that any explanation should lead users to always trust. But this should be more explicit, e.g., by aiming for appropriate trust and healthy distrust. One advantage could be that such an aim makes it more apparent that the success of XAI methods may be worth evaluating not only by trust but also by distrust, or that a decrease in trust can also be a positive result. If it is only evaluated by trust, only the function to identify reasons to trust is directly assessed, and it becomes complicated if an explanation provides reasons for both trust and distrust.

With regard to at least formulating the explainability-trust hypothesis more cautiously, our overview in Manuscript 2 indicates improvement. In the period of 2024, all papers except one aimed for appropriate trust and not trust, while in the years before, only half of the papers did. However, we find that such research still has a problem: The complexity of the aim of appropriate trust.

In our overview of existing research in Manuscript 2, we argued that the concepts appropriate, warranted, or calibrated trust or similar phrasing all share the aim of appropriate reliance. Typically, appropriate reliance is defined as the alignment between perceived and actual trustworthiness (Mehrotra et al., 2024). This entails that appropriate reliance is ensured if a hypothetical person would only rely on correct AI but not on incorrect AI. This also entails that someone appropriately relies if they only rely on an AI system suitable for a given task in a suitable manner and do not rely on AI systems that are unsuitable for the task or rely on a system in an unsuitable manner.

Manuscript 2 oversimplifies appropriate reliance as ensuring that a potential user relies on correct and does not rely on incorrect AI. This is oversimplified because such users would need a perfect understanding of the task for which they receive AI support. This perfect understanding is necessary because it would allow them to rely only on correct AI, but not on incorrect AI. Arguably, such users would then not need AI support, because with their perfect knowledge, they could also decide without the AI support or, at most, only use the AI support as a helpful tool, like a mathematician would use a calculator. Importantly, this then would not necessitate any trust or distrust because no uncertainty is present. Thus, the requirements for trust and distrust to be conceptually relevant are not fulfilled (Mühlfried, 2018).

This is an issue that we did not touch upon in Manuscript 2, but one which is discussed and approached in Manuscript 4 by explicating this complex aim of appropriate trust and healthy distrust, and thereby of appropriate reliance. Manuscript 4 takes a first step in establishing assessments that fully encompass the complexity of this aim. Therefore, we will return to this issue in a moment.

An aspect that also has to be critically acknowledged and discussed is that we and the existing research transfer concepts of trust and of distrust from an interpersonal context into the context of human-machine interaction. In Manuscript 1 and briefly in Manuscript 2, we already discussed this step and referenced evidence (Hoff & Bashir, 2015) that this is a valid application. Nonetheless, one needs to be cognizant of this step. While, for example, the Computers as Social Actors paradigm (Nass et al., 1994) supports the validity of using interpersonal concepts also in this domain (Stanton & Jensen, 2021), it can not be taken for granted that everything from the interpersonal applies to this context.

In the context of trust in automation, trust has a long history of being relevant and defining for the way people interact with technology (Hoff & Bashir, 2015; Lee & See, 2004). However, Glikson and Woolley, 2020 note that there is a key difference between prior generations of technology and current AI. While the capabilities of prior generations of technology were limited by the programmer's knowledge, AI can learn to make better or different decisions than a human. Thus, with different generations of technology, the way that insights on interpersonal trust translate to human-machine interaction may change because, e.g., the degree of anthropomorphism changes.

The relationship between anthropomorphism and the degree to which insights from interpersonal trust are valid for the AI context should not be regarded in a simple linear manner. Prominently, the uncanny valley effect (Mori et al., 2012) highlights that there is a qualitative difference when increasing anthropomorphism. Even though the uncanny valley is not directly linked to trust, it still points to the benefit of being conscious about the fact that one draws from an interpersonal concept, and some aspects can and should not easily be mapped to the context of automation and AI.

With acknowledgment of this fact, existing research has identified different commonalities between interpersonal trust and trust in automation or AI. The first key commonality is the function of trust: In both interpersonal and technological contexts, trust is assumed to reduce the complexity of an interaction. The second and third commonalities are risk and uncertainty. Both are considered as prerequisites for trust to be of matter. They may be slightly differently phrased, e.g., cost or vulnerability for risk. Therefore, on top of being aware of the transition from interpersonal to human-machine trust, it is advisable to ensure that there is complexity to be reduced and that risk and uncertainty are given when investigating trust and distrust. If they are not present, one does not investigate (dis)trust. This is, of course, difficult to fully simulate in an experimental setting. However, it is possible to respect these aspects in the design of experimental scenarios, to which I will return when discussing Manuscript 3, whose methodological design was strongly influenced by the insights gained in Manuscripts 1 and 2.

To sum up, approaching the notion of healthy distrust theoretically and conceptually in Manuscripts 1 and 2 formed a solid foundation for the other two manuscripts. Part of this foundation is the solid insight that there is no easy and prominent conceptual basis for healthy distrust. Instead, multiple aspects support our notion of healthy distrust. Across the current research on human-AI interaction, the need for a better understanding of healthy distrust can be observed. This is rarely explicit, but the way that XAI methods are envisioned to empower their users, it is well described as trust with a sufficient amount of distrust. The existing research on appropriate trust, together with the more fitting conceptualization of trust and distrust as two related yet separate dimensions, led us to advocate appropriate trust together with healthy distrust.

The two manuscripts discussed in the following chapter take steps to substantiate this proposition by empirical investigation of scenarios in which healthy distrust should occur. In these manuscripts, we improve and extend common assessments in human-AI interaction and test a way to foster healthy distrust.

Investigating Healthy Distrust

Manuscript 3

Manuscript 3 applies the theoretical and methodological insights obtained in Manuscripts 1 and 2 to an empirical investigation of two experimental scenarios. These scenarios are situated in the context of image classification, and I designed them in a way that healthy distrust is expected to occur. The two scenarios mainly differ in the material that is used. The first, the Real or Fake material (RoF), is about distinguishing real photographs from AI-generated images. The second, the Forms material, is about categorizing abstract forms into one of two categories. For the categorization of the forms, familiarization with the material and the categorization rules is necessary to decide to which category (Type A vs. Type B) each form belongs. [3] The RoF material does not require such a familiarization because it consists of images that people regularly encounter.

The stimuli were designed and tested for sufficient difficulty, and participants' performance in the experimental task was incentivized with a monetary bonus. These two steps respect two insights from Manuscripts 1 and 2: The fact that (1) the stimuli are sufficiently difficult to categorize makes the participants' decision uncertain, and (2) that the participants could receive a bonus introduces a form of risk. Thereby, the two prerequisites of trust — risk and uncertainty — are ensured to be present during the experiments. Monetary incentives are one of the measures that Miller (2022) suggests as possible ways to have risk present in one's experiments.

Notably, the possibility that participants could win or lose the bonus by performing well or badly does neither equate nor resemble the risk of decisions in those high-stakes scenarios that typically motivate research on trustworthy AI. Instead of simulating a condition similar to, say, an AI-advised medical diagnosis, this rather ensures that the task is not completely inconsequential to the participants. Therefore, monetary incentivization

[3] For more details and examples of the material, turn to Manuscript 3 or to https://github.com/PsyLab-UPB/Img-Clsfct-Dstrst, where all stimuli can be found.

does mitigate low-effort responses, but it must not be regarded as a way to create a true high-stakes scenario. I would argue it is a balanced tradeoff between the objective of making it conceptually valid to consider trust and distrust, and having a feasible experimental setup. If necessary, the felt risk by the participants could, for example, be increased by improving task engagement among participants through gamification of the experiments. Improving the introduction and manipulation of risk in experiments on human-AI interaction could be an important part of future research because of the role that risk may have on the desired balance between appropriate trust and healthy distrust. I will elaborate on that in the conclusion.

Returning to the experiments of Manuscript 3, for both scenarios, the experiment was split into two sessions. To increase the generalizability of the results, both scenarios were conducted once in a laboratory and once in an online setting. In Session 1, participants had to categorize the images on their own — they made an unadvised decision. In Session 2, they had to categorize the same images once again, but this time they received AI advice. The participants were informed about the AI advice and received, depending on the condition they were in, different information about the AI advice. What the participants did not know was that the advice was not AI-generated but was either correct or wrong based on a probability value that differed depending on the phase of Session 2: the pre-error, error, and post-error phase. In the pre- and post-error phase, 90% of the advice was correct. In the error phase, this probability was gradually decreased to only 45% correctness. With this error-prone advice in combination with the AI cover story, I wanted to create a scenario in which healthy distrust should occur.

Furthermore, I instructed half of the participants to be skeptical of the AI-advice and check for themselves if it should be used for their decision. This instruction was meant to increase the participants' healthy distrust. The other half of the participants received the same information but were not instructed to be skeptical and were told that they should use the advice. Comparing these two conditions (information-only vs. distrust), I tested whether such an instruction can foster healthy distrust. For this test, the performance of the participants in the two conditions was compared. Against my expectation and as discussed in the following, also against common expectations, the performance between the conditions did not differ. Thus, the instruction to distrust the error-prone AI advice was not helpful in comparison to the information-only instruction.

For investigating healthy distrust, this means that a simple instruction prior to the participants' interaction with the mock-up AI advice does not foster their healthy distrust. I picked this manipulation of the instruction as a first and very feasible means to potentially foster healthy distrust. Instead of indicating that such instructions are ineffective, other alternative explanations can, of course, be possible. The first alternative explanation for the ineffectiveness that comes to mind is that the participants ignored the instructions. Expecting this potential issue, the laboratory participants received the instructions twice, once verbally and a second time in writing, and the online participants had to summarize the instructions they received in their own words. Although I cannot rule out that the participants still attributed little relevance to the instruction, this ensured that they at least apprehended the general tone of the instruction.

Another alternative explanation is that the participants' interaction trajectory with the AI advice was a stronger influence on their reliance on the advice than the instruction. The probability of receiving incorrect advice and its changes were the same across the two conditions. Experiencing this may have altered the participants' reliance irrespective of the initial instruction, leading to similar performance across conditions. Manuscript 3's approach did not allow for a more fine-grained analysis beyond the data split by the experiments' phases. Therefore, the second alternative explanation remains speculative. Repeating the instruction and manipulating whether such repetitions occur randomly or before errors would be adaptations that I would recommend for further investigation.

In a similar manner, this could be applied to investigating the use of disclaimers in interactions with LLM-based applications, such as ChatGPT. First works in that direction already exist, as Manuscript 4 also points out. Given the frequent use of such disclaimers in LLM-based chat applications, it would be expected that they not only shift responsibility to the applications' users but also foster a more critical usage of such applications' output. However, as Manuscript 3 mentions, the existing research remains inconclusive. Despite a different context, Manuscript 3's results add to that. Given the widespread adoption of LLMs, it is important to investigate whether my results directly translate to the interaction with LLM-based applications. Again, comparing disclaimers to repeated warnings (random or specific) would be an interesting direction for future research. Furthermore, I would expect that the context of the interaction, for example, its topic, and thereby the involved risk and uncertainty, to be an interesting moderator.

Beyond the investigation of fostering healthy distrust via instructions, Manuscript 3 contributes to refining the assessment of appropriate reliance. To improve appropriate reliance, the two problems, under- and overreliance, need to be mitigated. For that, I developed a quantification of appropriate reliance by utilizing parameters of the Signal Detection Theory (SDT; e.g., Hautus et al., 2021) obtained from AI-advised and unadvised decisions. For details, please turn to Manuscript 3. The advantages of this analysis are that it provides a performance estimate independent of potential response biases (e.g., participants tend to categorize almost all images as real even though some of them are fake), that the analysis respects individual performance differences, and allows for inspecting the mitigation of under- and overreliance.

The first advantage is a core merit of using SDT and is already very well explained, for example, by Hautus et al. (2021). The second advantage is the merit of my two-session approach. For analyzing appropriate reliance, I used the d' difference between Session 2 and Session 1. Thereby, I analyzed only the performance difference from having AI advice in Session 2. For the last advantage, in Manuscript 3's experiments, a ground truth was needed. I calculated the d' difference once for the trials where the advice was correct and once for the trials where the advice was incorrect. Based on the direction and magnitude of these differences, conclusions can be drawn about whether under- or overreliance, or both, are mitigated. As mentioned, employing the analysis in the described way requires a ground truth to distinguish between correct and incorrect advice. Especially, more realistic applications often do not allow for that. Therefore, it is important to note that it is also possible to follow a similar approach in the absence of ground truth. Instead of splitting trials by advice correctness, it would also be possible to compare a well-performing system with a poorly performing one, or to compare one XAI method to another.

Furthermore, throughout Session 2, the participants were repeatedly asked to report their trust and distrust. Informed by Manuscripts 1 and 2, I assessed both trust and distrust to reflect the two-dimensional concept, and we did so repeatedly, given Manuscript 2 insights on (dis)trust's dynamicity. The self-report shows that the participants' trust decreased and their distrust increased after the error phase. In addition, given the multiple time points, it was also possible to analyze when this decrease and increase took place. Interestingly, after the first third of the error phase, where the advice correctness was at 75%, I did not find a meaningful difference compared to the self-report in the pre-error phase. The differences were observed when comparing the self-report obtained after the entire error phase to the previous time points. Moreover, neither trust nor distrust returned to the values reported at the beginning of the experiments, even though the AI advice in the post-error phase was as often correct as in the pre-error phase.

As discussed in Manuscript 3, this aligns well with the so-called asymmetry principle (Poortinga & Pidgeon, 2004; Slovic, 1993) in existing trust research, which describes trust as easy to lose but hard to gain, and distrust research, where the reverse is assumed (Guo et al., 2017; Vaske, 2016). Unsatisfyingly, these results and the way I measured them do not provide evidence for or against the two-dimensional concept of trust and distrust. Given the single-item approach that I chose, a correlation between the two items is valid in the one-dimensional concept, but also in the two-dimensional concept. To this end, more extensive and validated questionnaires are required. I chose the single-item approach because of the manuscript's focus on repeated interaction with AI and the need for multiple assessments. In these experiments, we wanted to avoid presenting the same extensive questionnaire five times throughout the experiments by picking the less obtrusive single-item judgments.

Manuscript 4

Manuscript 4 approaches the investigation of healthy distrust from a different angle. Instead of refining typical measurements in the context of human-AI interactions, it introduces a new potential indicator of healthy distrust. In doing so, Manuscript 4 makes the complex aim of appropriate trust and healthy distrust apparent. We already approached this in the previous section via the closely related concept of appropriate reliance. Similarly, Manuscript 4 discusses that appropriate trust is described in a way that it is given when humans trust an AI system, while they also notice errors and have an intuition when to expect errors. This points to the fact that desired parts of appropriate trust are conscious perceptions in combination with a mere, potentially unconscious, intuition. As problematized in Manuscript 4, the latter part may be difficult to operationalize via self-report, even if trust and distrust are assessed. Therefore, my co-authors and I argue that self-report and reliance do not suffice to fully encompass such an aim. We suggest attention as an additional indicator.

Given our scenario and methods, and the connection between vigilance and distrust that is discussed in Manuscript 4, we focused on visual attention. The experimental paradigm is similar to the ones from Manuscript 3. The important difference is that Manuscript 4's setup includes 2 images that are classified by mock-up AI instead of 1 image. Two images are needed to assess visual attention via the TVA-TOJ paradigm. For details of this paradigm, turn to Manuscript 4's methods section or to the more extensive details that we also reference in said manuscript (Bundesen, 1990; Bundesen et al., 2015; Tünnermann et al., 2017).

Experiment 1 of Manuscript 4 investigated the influence of the stimuli's categorization difficulty on visual attention, and Experiment 2 investigated the influence of the AI classification's correctness on visual attention. To briefly summarize the results of Manuscript 4: The categorization difficulty influences the attentional weight w but not the attentional capacity C. For the classification's correctness, we observed the reverse, that is it influenced the attentional capacity but not the weighting.

So what does that mean for investigating healthy distrust? — Unfortunately, in terms of a clear indication of healthy distrust, rather little. If the classification's correctness had increased w_{probe}, this would have been easy to interpret, but it did not. The observed reduction in attentional capacity due to misclassification would have been easier to interpret if we had observed a correlation between this reduction and the participants' performance on judging the classifications in our post-hoc analysis; yet again, it did not.

However, it does not mean that we cannot take away anything from the results for investigating healthy distrust. Firstly, the results indicate that our manipulation worked because it affected TVA's parameters; otherwise, we would not have observed the C difference or the w effect of the categorization difficulty. As discussed, in Manuscript 4, if the decrease in attentional capacity due to misclassification were to be replicated, it would be highly valuable to explore the potential benefits of it. Moreover, regarding the results on w, the effect of categorization difficulty informs a direction for future research, namely, to improve the manipulation of the image classifications.

This is also mentioned in Manuscript 4's discussion, where we suggested introducing systematic errors for a certain subtype of the stimuli. This could be tested to affect w. Furthermore, to connect back to Manuscript 3, it would also be interesting to have such a systematic error in place and combine it with the temporal increase of errors of the experiments in Manuscript 3. Additionally, further investigation on disclaimers or on warnings throughout the interaction could be tested for their potential to foster healthy distrust, e.g., by manipulating whether or not they help to identify the pattern behind the systematic error.

With a more critical stance, at least two things should be considered about the TVA-TOJ usage of Manuscript 4: Firstly, despite the efforts to closely combine the TOJ and the classification judgment task, which went into Manuscript 4's setup, the possibility remains that the participants separated the two tasks from one another. This would be problematic because it can diminish the effect that manipulations of the classification have on the attention measurement. Changing the order of classification judgment and TOJ, or additionally tracking the participants' gaze, could corroborate the discussed results.

Secondly, it has to be considered that the TVA-TOJ approach is established for investigating changes in visual attention due to low-level manipulations (e.g., saliency manipulations) that directly manipulate the stimuli's sensory information (though application beyond such typical manipulations exists (see, e.g., Banh et al., 2024; Biermeier et al., 2024)). Given that our manipulation changes the correctness of a label regarding the combination of multiple features of the stimuli, it might not work as well. While I think the connection between healthy distrust and attention is a plausible and interesting connection to investigate, it may be difficult to operationalize via a paradigm that typically investigates more direct manipulations of visual attention. Therefore, it may not be possible to obtain more convincing results for using visual attention as an indicator of healthy distrust with this approach.

An additional part of Manuscript 4's investigation was the research question of whether participants' performance improves when they are given the option to deliberate on their judgment. To this end, participants had, in a quarter of their trials, the option to delay their response. They could choose to be presented with the stimulus once again, and only then decide whether the classification was correct. Problematically, only a few participants used this response option, and if they did, they did so only a few times. Therefore, we did not obtain sufficient data to carry out the planned analysis for that research question. Potentially, it was too ambitious to add this research question to the already existing experimental design. For a sufficient trial count for the other analyses, the experiment was already long and rather complicated. Adding this additional option may have been overwhelming for the participants, which resulted in them ignoring the additional response option. We tried to minimize the additional demand by varying the response options only block-wise, not trial-wise, but it appears that it would have been better to split this into two separate experiments.

In summary, and despite the discussed limitations, Manuscripts 3 and 4 make multiple contributions to investigating healthy distrust. They apply insights from the first two manuscripts, refine the investigation of appropriate reliance, extend typical assessments of human-AI interaction by a formal assessment of visual attention, and identify against my and common expectations that instructions do not foster healthy distrust.

Future directions

All four manuscripts approach the topic of healthy distrust via (appropriate) trust. I did this to align the contributions with the existing approaches, notions, and research in the AI context. While doing that, we always advocated for not only aiming for appropriate trust but also for healthy distrust. In a closely related preprint (Paaßen et al., 2025), the benefit of this explication is discussed. In parallel, it is not my intention to replace the aim of appropriate trust with aiming for healthy distrust. I intend to add the notion of healthy distrust to the manifold research on appropriate trust.

As discussed in Manuscripts 1 and 2, existing research describes appropriate trust in a way that not only trust but also distrust is sometimes warranted. But often this research's operationalization of trust and distrust remains underdeveloped. Distrust is neglected because either only trust is assessed or scales about trust and distrust are interpreted one-dimensionally, even though they are better interpreted two-dimensionally. It is not wrong to assess trust, but if only trust is assessed, the full dynamics of trust and distrust are not measured. Of course, the fact that trust and distrust are often negatively correlated means that trust can be a decent proxy for distrust. However, for appropriate trust and healthy distrust, the cases where trust and distrust are not proxies of one another appear especially relevant. Therefore, this argumentation leading to the suggestion of considering trust and distrust is an overarching argument across the four papers and this dissertation.

Beyond conceptual improvements and more accurate measurements, aiming for appropriate trust and healthy distrust would also broaden the development of relevant research questions and the interpretation of results. While the more cautious formulations of the explainability-trust hypothesis implicitly share this aim, the implicitness and the focus on evaluating trust trigger research questions such as 'Does my method lead to the right amount of trust?' However, the more accurate question would be, 'Does it lead to the right amount of trust, and does it also lead to the right amount of distrust?' Aiming for appropriate trust and healthy distrust can shift the interpretation of results. For example, observing lower trust due to an XAI method may in fact be a positive result, or observing no changes in trust may in fact be that the XAI method sometimes fosters trust but also sometimes fosters distrust.

The question then, of course, is what is the right amount of (dis)trust, and this brings me back to a point of the previous section that I promised to elaborate on. I regard the amount of uncertainty and risk within the interaction with AI as playing a crucial role in calibrating the right amount of trust and distrust for the given context. Beyond ensuring that risk and uncertainty are present in the investigation of human-AI interaction to fulfill the requirements that (dis)trust is conceptually relevant, I advocate for further investigation of potential moderating effects by manipulating the context of the interaction to test for the influence of differing risk and uncertainty. There are scenarios where practically no distrust can be healthy (e.g., using generative AI for a creative task) and scenarios where having a lot of distrust can be healthy (e.g., medical advice). The opposite also holds for trust. Therefore, the influences of the context's risk and uncertainty are relevant for future research on trust and distrust in human-AI interaction.

So far, we have not been able to sufficiently back up the separation of trust and distrust. Our assessment of trust and distrust in Manuscript 3 is too shallow to provide convincing evidence for the two-dimensional concept. As already mentioned, more extensive and validated questionnaires are required. Given the single-item approach we chose, a correlation between the two items is valid in the one-dimensional concept but also in the two-dimensional concept. In that direction, analyses of trust questionnaires provide evidence for a better fit of two dimensions instead of one (Scharowski et al., 2025; Spain et al., 2008), and other studies also report qualitative benefits of considering trust and distrust when interacting with LLM-based applications (Colville & Ostern, 2024). Despite this, and seeing that multiple overviews come to the conclusion that this is not fully resolved, the conceptualization of trust and distrust is still debated. However, in these overviews, the two-dimensional concept is consistently described as the most fitting given the existing evidence, and that therefore, research should continue in this direction (Basel & Brühl, 2024; Vaske, 2016). For instance, in their chapter on distrust, Basel and Brühl (2024) conclude that the view on trust and distrust as closely related but distinct should be preferred. As they stress, this entails that trust and distrust can have positive and negative consequences, and that this, most importantly, conceptually allows for the investigation of positive effects of distrust.

Moreover, Basel and Brühl (2024) nicely describe how distrust plays a crucial role in politics, science, and economics. They consider a general distrust problematic. However, they regard elements that institutionalize specific distrust as beneficial and as contributors to ultimately trusting the relevant systems. For example, financial audits, double-blind peer review, and checks and balances are examples of distrust embodied into systems that function based on the trust that is put into them. The knowledge that these safeguards based on distrust are in place ultimately fosters trust.

Similarly, Poortinga and Pidgeon (2003) highlight that already in 1983, Barber argued that distrust can be an essential element of political accountability in participatory democracy, which they describe as an effective distrust. Another related argumentation can be found in Sperber et al.'s (2010) work on epistemic vigilance. They describe that the function of trust is buttressed by epistemic vigilance to distinguish it from blind trust. All of this is closely connected to what I also propose across the manuscripts, namely that a healthy amount of distrust may be the key ingredient for what is typically described as appropriately trusting. With a healthy distrust, one may have given enough room to their skepticism and doubt to inform and secure their trust.

To put my argumentation on a more abstract level, what I summarized and discussed until now indicates that trust needs a control mechanism. By knowing about this control, by experiencing it, and by seeing it in effect, trust can be further corroborated. Especially in the case of power imbalance, this is a crucial mechanism to be established.

Conceptually, distrust is related to this. While such mechanisms are not straightaway defined as distrust, distrust appears to me as the most fitting overarching term for the differing descriptions. Moreover, by picking distrust, one stays in line with the already employed terminology of trust in the AI context. Even though the idea is shared across contexts, no unified common ground exists. I do not think that I can provide such a common ground, but I hope to have underlined the commonalities that exist and backed up my notion of healthy distrust. For establishing and improving said common ground, I recommend, like Basel and Brühl (2024), the accumulation of knowledge about trust and distrust as two related dimensions and the positive effects of distrust.

Conclusion

With my dissertation, I provide an in-depth analysis of the notion of healthy distrust within the currently highly relevant context of AI. I identify the theoretical and conceptual foundations of healthy distrust, as well as the gaps within them. Translating these insights into my empirical investigations led to useful refinements and an extension of the measurement of healthy distrust and related notions. While I regard healthy distrust as a virtue for many contexts, I think that, especially, the research about human-AI interaction in general and about XAI specifically can benefit from it. These research areas already share motivations and aims close to the notion of healthy distrust, and only lack more clarity on it.

Given the current interest in and relevance of human-AI interaction, I regard this context as well-suited to progress the understanding of healthy distrust. By considering both trust and distrust, by focusing also on the beneficial consequence that distrust has, and accurately assessing it, I hope that a healthy distrust during human-AI interactions can be fostered. Moreover, returning to the introduction of this text, I hope that these insights also translate to and stimulate a healthy distrust in using AI. With this, I envision a sensible usage of AI applications, and humans who are empowered to decide when to trust and when to distrust an AI system.

References

Banh, N. C., Tünnermann, J., Rohlfing, K. J., & Scharlau, I. (2024). Benefiting from binary negations? Verbal negations decrease visual attention and balance its distribution. *Frontiers in Psychology*, *15*, 1451309. https://doi.org/10.3389/fpsyg.2024.1451309

Barber, B. (1983). The logic and limits of trust. *Rutgers Univ. Pr.*

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Basel, J., & Brühl, R. (2024). Misstrauen. Eine interdisziplinäre Bestandsaufnahme. In *Psychologie von Risiko und Vertrauen: Wahrnehmung, Verhalten und Kommunikation* (pp. 271–301). Springer. https://doi.org/10.1007/978-3-662-65575-7_11

Bashir, N., Donti, P., Cuff, J., Sroka, S., Ilic, M., Sze, V., Delimitrou, C., & Olivetti, E. (2024). The climate and sustainability implications of generative AI. *An MIT Exploration of Generative AI*. https://doi.org/10.21428/e4baedd9.9070dfe7

Biermeier, K., Scharlau, I., & Yigitbas, E. (2024). Measuring visual attention capacity across xReality. *Proceedings of the 17th International Conference on PErvasive Technologies Related to Assistive Environments*, 98–106. https://doi.org/10.1145/3652037.3652050

Bundesen, C. (1990). A theory of visual attention. *Psychological Review*, *97*(4), 523–547. https://doi.org/10.1037/0033-295x.97.4.523

Bundesen, C., Vangkilde, S., & Petersen, A. (2015). Recent developments in a computational theory of visual attention (TVA). *Vision Research*, *116*, 210–218. https://doi.org/10.1016/j.visres.2014.11.005

Chen, C., & Shu, K. (2024). Combating misinformation in the age of LLMs: Opportunities and challenges. *AI Magazine*, *45*(3), 354–368. https://doi.org/10.1002/aaai.12188

Colville, S., & Ostern, N. K. (2024). Trust and distrust in GAI applications: The role of AI literacy and metaknowledge. *ICIS 2024 Proceedings*. https://eprints.qut.edu.au/254224/

Deutsch, M. (1958). Trust and suspicion. *Journal of Conflict Resolution*, *2*(4), 265–279. https://doi.org/10.1177/002200275800200401

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5), 1–42. https://doi.org/10.1145/3236009

Guo, S.-L., Lumineau, F., & Lewicki, R. J. (2017). Revisiting the foundations of organizational distrust. *Foundations and Trends in Management*, *1*(1), 1–88. https://doi.org/10.1561/3400000001

Haenlein, M., & Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California management review*, *61*(4), 5–14. https://doi.org/10.1177/0008125619864925

Hautus, M. J., Macmillan, N. A., & Creelman, C. D. (2021). *Detection theory: A user's guide*. Routledge.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Kastner, L., Langer, M., Lazar, V., Schomacker, A., Speith, T., & Sterz, S. (2021). On the relation of trust and explainability: Why to engineer for trustworthiness. *Proceedings, 29th IEEE International Requirements Engineering Conference*

Workshops : REW 2021 : September 20-24 2021, online event, 169–175. https://doi.org/10.1109/REW53955.2021.00031

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025). Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task. *arXiv preprint arXiv:2506.08872*. https://doi.org/10.48550/arXiv.2506.08872

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, *46*(1), 50–80. https://journals.sagepub.com/doi/abs/10.1518/hfes.46.1.50_30392

Lewicki, R. J., McAllister, D. J., & Bies, R. J. (1998). Trust and distrust: New relationships and realities. *Academy of Management Review*, *23*(3), 438–458. https://doi.org/10.5465/amr.1998.926620

Luhmann, N. (2009). *Vertrauen : Ein Mechanismus der Reduktion sozialer Komplexität* (4th edition). Stuttgart : Lucius & Lucius.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*(3), 709–734. https://doi.org/10.5465/amr.1995.9508080335

Mehrotra, S., Degachi, C., Vereschak, O., Jonker, C. M., & Tielman, M. L. (2024). A systematic review on fostering appropriate trust in human-AI interaction: Trends, opportunities and challenges. *ACM Journal on Responsible Computing*, *1*(4), 1–45. https://doi.org/10.1145/3696449

Miller, T. (2022). Are we measuring trust correctly in explainability, interpretability, and transparency research? *ArXiv*, *abs/2209.00651*. https://doi.org/10.48550/arXiv.2209.00651

Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on*

Interactive Intelligent Systems (TiiS), *11*(3-4), 1–45. https://doi.org/10.1145/3387166

Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine*, *19*(2), 98–100. https://doi.org/10.1109/MRA.2012.2192811

Mühlfried, F. (2018). *Mistrust–ethnographic approximations*. transcript Verlag. https://doi.org/10.14361/9783839439234

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. *Proceedings of the SIGCHI conference on Human factors in computing systems*, 72–78. https://doi.org/10.1145/259963.260288

Newport, C. (2025, August). *What if A.I. doesn't get much better than this?* The New Yorker. Retrieved September 9, 2025, from https://www.newyorker.com/culture/open-questions/what-if-ai-doesnt-get-much-better-than-this

Paaßen, B., Alpsancar, S., Matzner, T., & Scharlau, I. (2025). Healthy distrust in AI systems. *arXiv preprint arXiv:2505.09747*. https://doi.org/10.48550/arXiv.2505.09747

Perrigo, B. (2023, January). *OpenAI used Kenyan workers on less than $2 per hour to make ChatGPT less toxic*. Time Magazine. Retrieved September 9, 2025, from https://time.com/6247678/openai-chatgpt-kenya-workers

Peters, T. M., Biermeier, K., & Scharlau, I. (2026). Assessing healthy distrust in human-AI interaction: Interpreting changes in visual attention. *Frontiers in Psychology, 16, 1694367*. https://doi.org/10.3389/fpsyg.2025.1694367

Peters, T. M., & Visser, R. W. (2023). The importance of distrust in AI. In L. Longo (Ed.), *Explainable Artificial Intelligence. xAI 2023. Communications in Computer and Information Science, vol. 1903* (pp. 301–317). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-44070-0_15

Peters, T. M., & Scharlau, I. (2025). Interacting with fallible AI: Is distrust helpful when receiving AI misclassifications? *Frontiers in Psychology*, *16*, 1574809. https://doi.org/10.3389/fpsyg.2025.1574809

Poortinga, W., & Pidgeon, N. F. (2003). Exploring the dimensionality of trust in risk regulation. *Risk Analysis: An International Journal*, *23*(5), 961–972. https://doi.org/10.1111/1539-6924.00373

Poortinga, W., & Pidgeon, N. F. (2004). Trust, the asymmetry principle, and the role of prior beliefs. *Risk analysis: an international journal*, *24*(6), 1475–1486. https://doi.org/10.1111/j.0272-4332.2004.00543.x

Regulation (EU) 2024/1689. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). Retrieved October 24, 2024, from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689

Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust. *Journal of Personality*. https://doi.org/10.1111/j.1467-6494.1967.tb01454.x

Rotter, J. B. (1971). Generalized expectancies for interpersonal trust. *American Psychologist*, *26*(5), 443. https://doi.org/10.1037/h0031464

Scharowski, N., Perrig, S. A. C., von Felten, N., Aeschbach, L. F., Opwis, K., Wintersberger, P., & Brühlmann, F. (2025). To trust or distrust AI: A questionnaire validation study. *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 361–374. https://doi.org/10.1145/3715275.3732025

Sheikh, H., Prins, C., & Schrijvers, E. (2023). Artificial intelligence: Definition and background. In *Mission AI: The new system technology* (pp. 15–41). Springer. https://doi.org/10.1007/978-3-031-21448-6_2

Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. (2021). Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*. https://doi.org/10.48550/arXiv.2104.07567

Slovic, P. (1993). Perceived risk, trust, and democracy. *Risk analysis*, *13*(6), 675–682. https://doi.org/10.1111/j.1539-6924.1993.tb01329.x

Spain, R. D., Bustamante, E. A., & Bliss, J. P. (2008). Towards an empirically developed scale for system trust: Take two. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *52*(19), 1335–1339. https://doi.org/10.1177/154193120805201907

Sperber, D. A., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, *25*(4), 359–393. https://doi.org/10.1111/j.1468-0017.2010.01394.x

Stanton, B., & Jensen, T. (2021). Trust and artificial intelligence. https://doi.org/10.6028/nist.ir.8332-draft

Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy artificial intelligence. *Electronic Markets*, *31*(2), 447–464. https://doi.org/10.1007/s12525-020-00441-4

Tünnermann, J., Krüger, A., & Scharlau, I. (2017). Measuring attention and visual processing speed by model-based analysis of temporal-order judgments. *JoVE*, (119), e54856. https://doi.org/doi:10.3791/54856

Vaske, C. (2016). *Misstrauen und Vertrauen*. Universität Vechta. https://doi.org/10.23660/voado-31

Visser, R., Peters, T. M., Scharlau, I., & Hammer, B. (2025). Trust, distrust, and appropriate reliance in (X)AI: A conceptual clarification of user trust and survey of its empirical evaluation. *Cognitive Systems Research*, *91*, 101357. https://doi.org/10.1016/j.cogsys.2025.101357

Wu, K., Wu, E., & Zou, J. Y. (2024). Clasheval: Quantifying the tug-of-war between an LLM's internal prior and external evidence. *Advances in Neural Information Processing Systems*, *37*, 33402–33422. https://proceedings.neurips.cc/paper_files/paper/2024/file/3aa291abc426d7a29fb08418c1244177-Paper-Datasets_and_Benchmarks_Track.pdf

Zhang, J., & Li, C. (2019). Adversarial examples: Opportunities and challenges. *IEEE transactions on neural networks and learning systems*, *31*(7), 2578–2593. https://doi.org/https://doi.org/10.1109/TNNLS.2019.2933524

In addition to the synopsis, this dissertation includes four published manuscripts, reproduced on the following pages. The digitally published version of the synopsis ends here.

Manuscript 1

Peters, T. M., & Visser, R. W. (2023). The importance of distrust in AI. In L. Longo (Ed.), *Explainable Artificial Intelligence. xAI 2023. Communications in Computer and Information Science, vol. 1903* (pp. 301–317). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-44070-0_15

Manuscript 2

Visser, R., Peters, T. M., Scharlau, I., & Hammer, B. (2025). Trust, distrust, and appropriate reliance in (X)AI: A conceptual clarification of user trust and survey of its empirical evaluation. *Cognitive Systems Research*, *91*, 101357. https://doi.org/10.1016/j.cogsys.2025.101357

Manuscript 3

Peters, T. M., & Scharlau, I. (2025). Interacting with fallible AI: Is distrust helpful when receiving AI misclassifications? *Frontiers in Psychology*, *16*, 1574809. https://doi.org/10.3389/fpsyg.2025.1574809

Manuscript 4

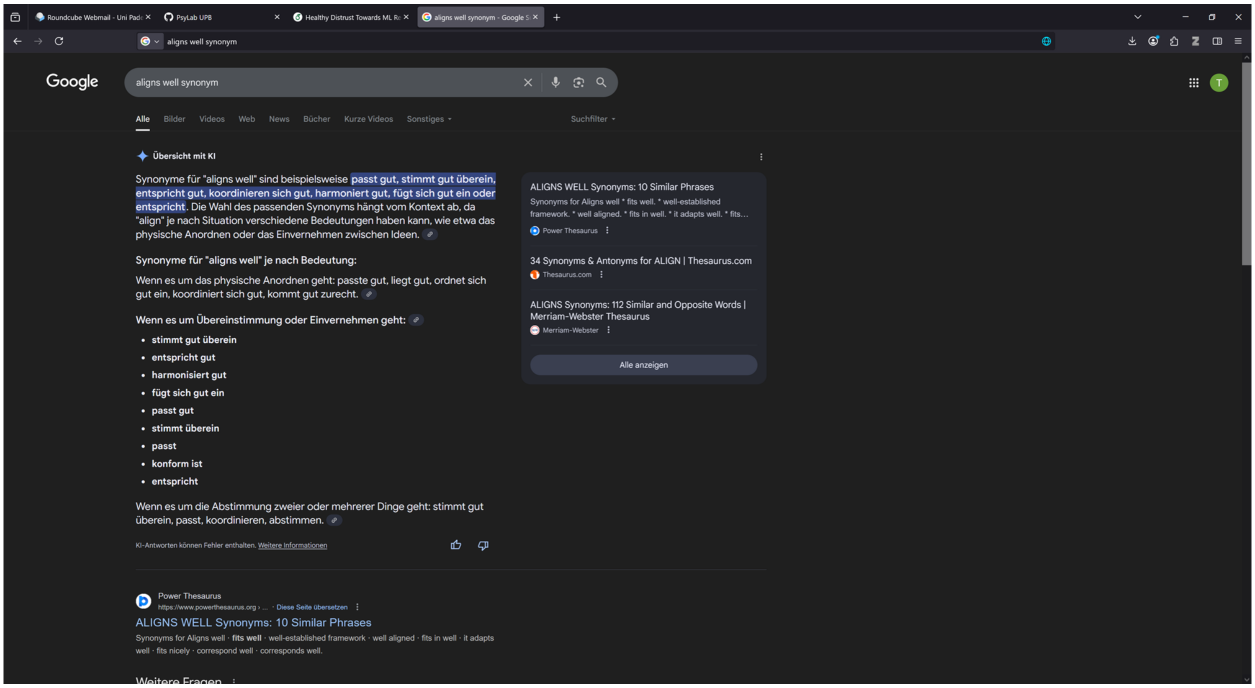
Peters, T. M., Biermeier, K., & Scharlau, I. (2026). Assessing healthy distrust in human-AI interaction: Interpreting changes in visual attention. *Frontiers in Psychology, 16, 1694367.* https://doi.org/10.3389/fpsyg.2025.1694367

Appendix

Screenshot — internet search for synonym

Figure A1

Screenshot of the internet search for synonyms for 'alings well' that is discussed in the introduction.



Übersicht mit KI

Synonyme für "aligns well" sind beispielsweise passt gut, stimmt gut überein, entspricht gut, koordinieren sich gut, harmoniert gut, fügt sich gut ein oder entspricht. Die Wahl des passenden Synonyms hängt vom Kontext ab, da "align" je nach Situation verschiedene Bedeutungen haben kann, wie etwa das physische Anordnen oder das Einvernehmen zwischen Ideen.

Synonyme für "aligns well" je nach Bedeutung:

Wenn es um das physische Anordnen geht: passte gut, liegt gut, ordnet sich gut ein, koordiniert sich gut, kommt gut zurecht.

Wenn es um Übereinstimmung oder Einvernehmen geht:

- **stimmt gut überein**
- **entspricht gut**
- **harmonisiert gut**
- **fügt sich gut ein**
- **passt gut**
- **stimmt überein**
- **passt**
- **konform ist**
- **entspricht**

Wenn es um die Abstimmung zweier oder mehrerer Dinge geht: stimmt gut überein, passt, koordinieren, abstimmen.

KI-Antworten können Fehler enthalten. Weitere Informationen

ALIGNS WELL Synonyms: 10 Similar Phrases
Synonyms for Aligns well * fits well. * well-established framework. * well aligned. * fits in well. * it adapts well. * fits…
Power Thesaurus

34 Synonyms & Antonyms for ALIGN | Thesaurus.com
Thesaurus.com

ALIGNS Synonyms: 112 Similar and Opposite Words | Merriam-Webster Thesaurus
Merriam-Webster

Alle anzeigen

Power Thesaurus
https://www.powerthesaurus.org › ... · Diese Seite übersetzen

ALIGNS WELL Synonyms: 10 Similar Phrases

Synonyms for Aligns well · **fits well** · well-established framework · well aligned · fits in well · it adapts well · fits nicely · correspond well · corresponds well.

Weitere Fragen