The Importance of an Ethical Framework for Trust Calibration in AI

Amelie Schmid and Manuel Wiesche, *TU Dortmund University, 44221, Dortmund, Germany*

The transformative power of AI raises serious concerns about ethical issues within organizations and implicates the need for trust. To cope with that, numerous ethical frameworks are generally published, but only on a theoretical level. Furthermore, proper trust calibration in AI is of high relevance for the workers. Up to now, only limited studies have been carried out to investigate how an ethical framework can foster proper trust calibration of workers in practice. To close this gap, an ethical framework is investigated that ensures trust calibration by targeting AI reliability and AI safety. Finally, the effectiveness of the applied framework is evaluated based on 17 interviews within an international automotive supplier. As a result, this ethical framework led to a major increase in trust. This is a groundbreaking outcome since workers are willing to accept a lower level of AI safety and AI reliability at the same time.

AI is one of the main drivers in the field of technology and has a major impact on organizations and society. Its transformative power to improve life and business comes along with serious concerns about ethical aspects. AI is often described as a "black box," which is basically caused by the fact that it is difficult for workers to understand the algorithm, its underlying rules, and the logic behind them.[1] This leads workers feeling that they are faced with nontransparent decision-making processes of AI. Subsequently, challenges related to the explainability and robustness of the AI model have to be tackled.[2] This can be overcome by the introduction of ethical frameworks, which are of increasing importance. Up to now, numerous ethical frameworks containing fundamental ethical principles have been developed by researchers, organizations, and governments.[3]

However, these frameworks typically maintain a high level of abstraction, lacking specific guidance on their practical implementation and operationalization. These findings highlight a compelling need to explore the implementation of an ethical framework within an organizational context.

Furthermore, to unleash the full potential of AI within organizations, a high degree of trust in the emerging technology is required as a prerequisite. However, the lack of trust represents one of the main factors preventing workers from taking full advantage of AI. To compensate for the lack of trust and to ensure business success, special attention needs to be paid to the process of building trust, also called "trust calibration." Trust calibration is targeting the avoidance of overtrusting or undertrusting the AI. The former might lead to serious safety issues and the latter to a prevention of AI usage.[4]

Both aforementioned impact factors are governed by numerous influencing factors that specifically depend on the research stream and its perspective. User-related studies in information systems exhibit that the *reliability* of the AI system serves as an important baseline for proper trust calibration. The reliability refers to the confidence of the worker that the technology will operate in a suitable and consistent manner.[5,6] In contrast, in computer science, it is evident that AI *safety* represents an important pillar within ethical frameworks. It is straightforward that AI safety is closely related to reliability since AI safety requires, as a prerequisite, that the AI system be reliable.[3,7] Thus, both terms can even be considered as synonyms.

Since AI safety and AI reliability can be used as synonyms, it is obvious that both influencing factors have, in contrast, a direct impact on both AI ethics and AI trust. Following this logic in a literature review, it turned out that only limited studies have been carried out to investigate how an ethical framework can foster a proper trust calibration of workers in practice. Thus,

1541-1672 © 2023 IEEE
Digital Object Identifier 10.1109/MIS.2023.3320443
Date of publication 29 September 2023; date of current version 13 November 2023.

the following research question will be answered: *How do organizations use ethical frameworks to calibrate trust in AI-based systems?*

The research question is approached by a two-part exploratory analysis, using 17 semistructured interviews. First, the applied ethical framework within one of the biggest automotive suppliers is illustrated. Second, the effectiveness of this implementation is analyzed. The process of trust calibration in AI as a function of time is evaluated in combination with the underlying ethical framework and further measures.

BACKGROUND

AI Trust

It is well known that trust is a crucial success factor for the acceptance and the sustainable adoption of AI.[8,9] To reach that goal within organizations, workers have to calibrate their level of trust according to the reliability of the AI system.[6] During the process of trust calibration, overtrust (trust that surpasses AI's reliability) or undertrust (trust that falls behind AI's reliability) should be prevented to avoid a negative impact on successful human–AI cooperation.[4,9] It must be noted that, in contrast to former technologies, like traditional automation, the behavior of AI-based technologies is rather unknown.[10] These so-called "black-box" models result in a limited understanding of the generated decisions and lack explainability.[2] Moreover, workers might not be capable of quantifying uncertainties of AI models by themselves in an adequate way. This creates tension regarding the actual robustness of the AI model, which slows down or even stops the process of trust calibration.[9]

Designers can facilitate a proper trust calibration by providing instant feedback concerning the reliability of the AI system toward the worker. This feedback of the designers includes, e.g., the explanation of factors that affect AI reliability or further information regarding the task performance.[11] In addition, an iterative and human-centered AI design process can further enhance explainability and robustness, which strengthens the implementation of trustworthy AI.[2] However, best practices as well as deeper insights into the dynamics behind the trust calibration in AI are still missing. This raises the demand for a profound understanding with respect to the dynamic process of trust calibration in AI within organizations as well as the roles of reliability, explainability, and robustness.

AI Ethics and the AI Lifecycle

For many years, AI ethics has been a controversial topic, which can be explained by the rapid technological development and high degree of complexity as, e.g., in software for automated and autonomous driving. The purpose of AI ethics is to serve as a fundament for ethical decision making by machines and computers. Such ethical frameworks give guidance for software developers as well as workers and ensure the ethical behavior of AI.[12] So far, various ethical frameworks, including their underlying ethical principles, have been proposed by researchers, organizations, and governments. All of these frameworks are limited to a theoretical level, and all of them make use of generic principles for AI ethics, such as transparency, justice, responsibility, privacy, or accountability.[3,7,13]

In practice, the application of AI systems follows different stages of the AI lifecycle: 1) requirements engineering, 2) design, 3) implementation, 4) verification, and 5) operation.[14] In contrast, ethical frameworks are predominately focusing on AI in general and are, apart from that, neglecting the existence of different phases. This is explained by the high degree of complexity of AI, which, obviously, makes it hard to implement suitable ethical frameworks in practice. It results in one of the most fundamental challenges within organizations, namely, to deal with remaining potential ethical risks.

As a countermeasure, we propose, for the first time, a new, practical approach by integrating the concept of trust calibration into the context of AI ethics. This connection can be established because the ethical principle of "safety" is closely linked to the important dimension "reliability" for trust calibration. Moreover, both characteristics, AI safety and AI reliability, are of striking importance for verification and validation as part of the AI lifecycle. In this fourth phase, it is checked whether the AI system fulfills the requirements and meets its purpose in the intended way.[14] The demand for a tailored ethical framework that specifically addresses AI reliability and safety becomes evident, emphasizing the importance of proper trust calibration in practice.

Research Setting

The entire analysis is based on a comprehensive study within an international automotive supplier. In this context, an AI-based system that serves as a quality gate in manufacturing is considered. The selected organization is developing and manufacturing semiconductors and sensors as well as electronic control units at several locations worldwide. To further improve the high-quality standards, AI-based visual inspection was implemented into the running operations at more than 100 manufacturing lines in seven international plants. Ensuring a high-quality manufacturing process is crucial, as the goods are used for safety-critical applications within vehicles whose failures or malfunctions might cause serious injury to people, severe damage to equipment and property, or environmental harm. As a matter of fact, ethical aspects are of high relevance for the internal quality gates.

APPLIED FRAMEWORK

Within the organization, an ethical framework is applied with the purpose of considering the safety criticality of the different failure modes (FMs) in manufacturing (Figure 1). This approach leads, in the end, towards a quantitative assessment of the overall ethical impact (EI) with a special focus on AI safety and AI reliability. The common ground of this framework is the "FM and effects analysis" (FMEA), which is an automotive standard according to the International Automotive Task Force 16949. It is a commonly used methodology within manufacturing to identify potential FMs and to evaluate their impact on the system. In the software context, the FMEA methodology gains increasing attention in front of the verification and validation of AI systems and contributes to ensuring AI safety and to understanding possible ethical failures and risks.[14,15]

In the present practical context, the FMEA methodology serves as a risk comparison tool and provides a systematic analysis of the potential EI regarding the new AI-based system.

The applied ethical framework consists of four elements: 1) FMs, 2) occurrence (O), 3) detectability (D), and 4) severity (S) of the respective FM. In the framework, the EI of the different FMs is represented by the product of the elements O, D, and S. To quantify the EI, the following procedure is used. First, an adequate timeframe is selected for the evaluation. During that timeframe, the O of the different FMs is documented. Second, the respective values for D and S are determined, for which merits are set between one and 10 according to the FMEA systematic. In this context, a small number represents a high probability of being detected by the AI system and a low impact for customers in the case of nondetection. A high number represents a low probability of being detected and a high impact on the customer in the case of nondetection. Finally, the individual scores of the different FMs are summed up over all defect types, resulting in an overall risk number for the AI-based system. This represents the overall EI, which, in short, is calculated by

$$\mathrm{EI} = \sum_{i=1}^{n} \mathrm{EI}_i(\mathrm{FM}_i) = \sum_{i=1}^{n} (\mathrm{O}_i * \mathrm{D}_i * \mathrm{S}_i). \quad (1)$$

In this way, the result of the new AI-based system can be compared to the previous method without the usage of AI.

EMPIRICAL EVALUATION

An empirical study was conducted to investigate both the effectiveness of the applied ethical framework and the respective trust calibration during the entire AI lifecycle within the global production network. In this context, the roles of AI reliability and AI safety were the focal point. Due to the sensitive nature of trust calibration in AI, it was essential to select an appropriate methodology for the evaluation. To consider that, an exploratory research strategy was chosen by using grounded theory methodology.[16]

Data Collection

Primarily, data were collected using semistructured interviews. To investigate the ethical framework as well as the trust calibration from different perspectives, 17 workers of different organizational roles and from different hierarchical levels were selected. Among others, the project leader of the AI project, representatives of quality management and engineering, and the

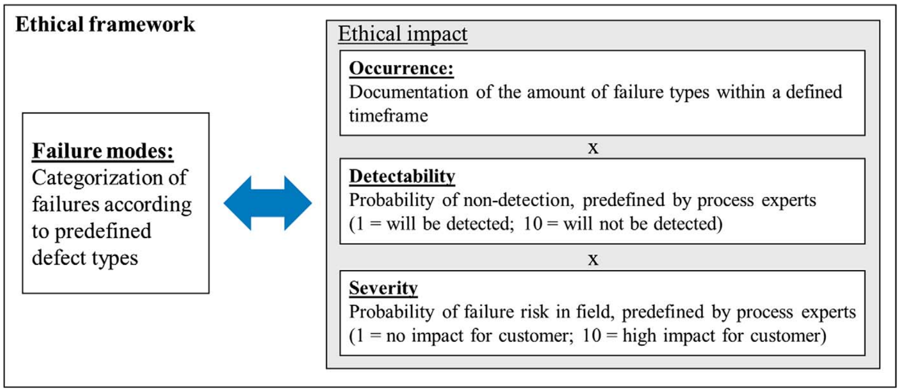

Ethical framework
Failure modes:
Categorization of failures according to predefined defect types
Ethical impact
Occurrence:
Documentation of the amount of failure types within a defined timeframe
x
Detectability
Probability of non-detection, predefined by process experts
(1 = will be detected; 10 = will not be detected)
x
Severity
Probability of failure risk in field, predefined by process experts
(1 = no impact for customer; 10 = high impact for customer)

FIGURE 1. An illustration of the ethical framework based on failure mode and effects analysis.

corresponding leaders who released the new AI-based process were considered. Moreover, we selected employees from different international plants with differing progress along the AI lifecycle. By means of that, we could make comparisons along the different phases. It is well noted that all data were collected with the informed consent of all participants. During the entire data collection and analysis process, special attention was paid to respecting participants' confidentiality and anonymity. As a consequence, identifiable details were obscured, and only selected sanitized parts of the data are shared.[17]

We collected further qualitative data to deepen the fundamental understanding and to capture the phenomenon with more clarity. Thus, numerous data, such as presentations for management reviews, documentation of the new process release, several checklists and best practices provided as informational material for the implementing plants, AI performance dashboards, the results of team internal workshops, and the AI model training process, were investigated.

Data Analysis

To obtain a profound overview of the key messages during the data collection process, a short summary with the most relevant statements was written after each interview.[16] These memos helped to recall all relevant data during the data collection process.[17] Consecutively, we were able to get back to the raw data during the data analysis process.

Based on grounded theory, we used open, axial, and selective coding for analyzing the interview transcripts. In the first step, open coding was applied to assign appropriate labels to the interview passages, which summarized the key message in a short phrase. The emerging main categories were still closely linked to the raw data. However, the generic expectations of the workers regarding AI reliability and AI safety along the five predefined phases of the AI lifecycle could be deduced. In the second step, axial coding was applied to interlink the categories and identify subcategories in a systematic way. In addition, the underlying mechanisms of trust calibration were elaborated. During the data analysis process, we realized varied mechanisms at the different phases of the AI lifecycle. That is the reason why we used selective coding to cluster the mechanisms accordingly and to synthesize the data regarding the trust calibration. Finally, we applied constant comparison to further develop and combine the dimensions, resulting in the final consolidation.[16,17]

RESULTS

The main target of the present work is to investigate the implementation of an ethical framework in practice and to analyze its effectiveness towards trust calibration. Trust calibration and the evolvement of AI reliability and AI safety are studied along the previously mentioned five phases of the AI lifecycle[14]: 1) requirements engineering, 2) design, 3) implementation, 4) verification, and 5) operation. As a result, an iterative, human-centered process model is developed, oriented along the AI lifecycle within an organization. The major contribution of the model is the interplay between trust calibration, the ethical framework, and AI reliability and AI safety (Figure 2). Three major findings can be deduced by the process model:

- The trust calibration of the workers along the five phases of the AI lifecycle.
- The major impact of the ethical framework during AI validation and verification.
- The development of the accepted level of AI reliability and AI safety.

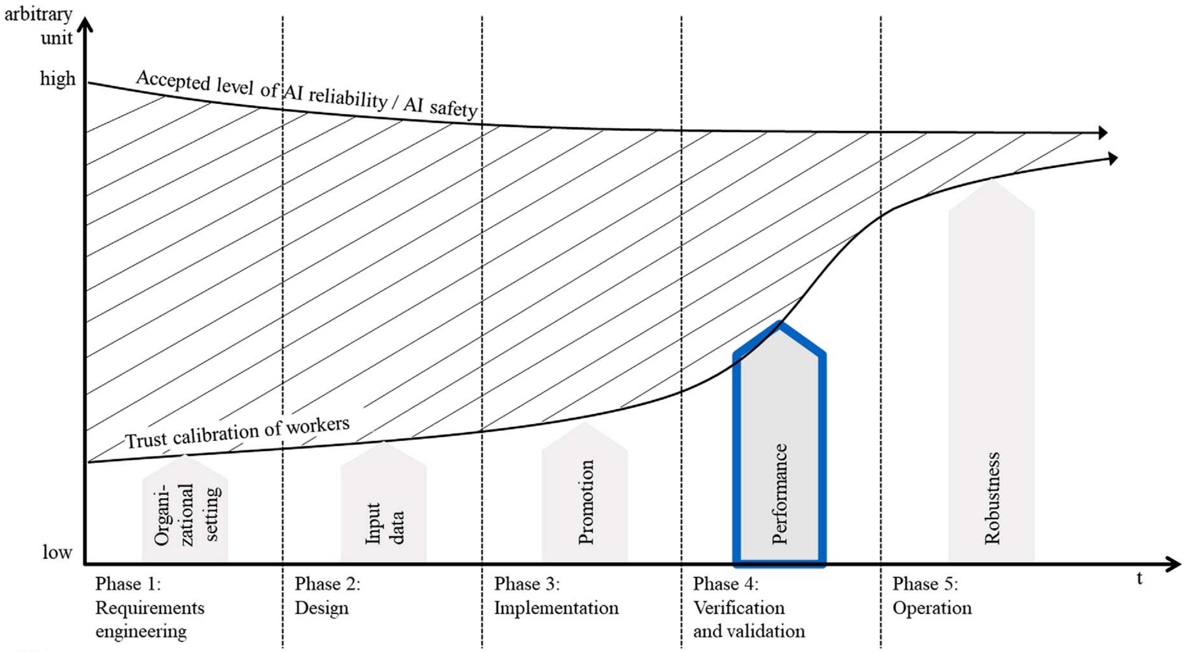
As a constraint, it must be noted that the illustration does not show any defined values but, rather, the general progression and tendency of the different aspects.

The Ethical Framework for Effective Trust Calibration

To support trust calibration and to overcome ethical concerns related to the AI-based system, an ethical framework is used. In particular, the framework puts special focus on the concerns regarding AI reliability and AI safety since these are of great importance for trust calibration. Based on that, the ethical framework applies predominantly during the fourth phase, AI verification and validation.

The accepted level of AI reliability and safety and the level of trust change dynamically as a function of time during the AI lifecycle. At the beginning of the implementation, the workers seek a very high degree of reliability and safety, while—at the same point in time—there exists a low level of trust. This is caused by the fact that the workers are not willing to accept any deviation of the AI model with respect to reality, which is also clearly communicated as a requirement to the AI development team.

> "In the beginning, they [the AI developers] always came to us and presented us some kind of key figures. Then they told us, 'Now we have [. . .] a slip-through of 1, 2 or 3%.' Then I said, 'You can forget that. We demand 0% slip-through from our machines, and now we are suddenly supposed to allow 1%.' And then I said, 'That's not going to happen, it won't work.'" (#02)


arbitrary unit
high
low
Accepted level of AI reliability / AI safety
Trust calibration of workers
Organi-zational setting
Input data
Promotion
Performance
Robustness
t
Phase 1: Requirements engineering
Phase 2: Design
Phase 3: Implementation
Phase 4: Verification and validation
Phase 5: Operation

Note:
The blue frame shows the immense influence of the ethical framework during the AI verification and validation. The five grey arrows show the supportive measures towards the level of trust. The phases of the AI lifecycle are separated by a dashed line.

FIGURE 2. Trust calibration along the AI lifecycle and the impact of the ethical framework during the fourth phase.

However, the results show that, during the AI lifecycle, the workers are willing to deviate from that initial requirement. This means that they are willing to accept a lower degree of reliability and safety as compared to their original statement. This finds entry in the project plan, e.g., by the adaption of the reliability requirement from 100% toward a lower level.

Strikingly, this optimization of the reliability targets has no negative effect on the level of trust, which is, nonetheless, still increasing. At this point in the AI lifecycle, there is a considerable impact of the ethical framework, as it helps the workers to enhance their understanding regarding the AI-based system. This results in the largest leap regarding trust calibration upon this fourth phase.

> "The AI classifies in the background, and you can check: What does the AI say? What does the operator say? Is there a mismatch, and what is the impact?" (#03)

The ethical framework provides a simple comparison pattern to give evidence of AI reliability and AI safety as compared to the previous process without using AI. Based on that, a risk evaluation is carried out, which specifies the possible EI of the different FMs.

> "I would say the most important tool that we provide for the plants to monitor the functioning of the AI on a day-to-day basis is a comprehensive dashboard." (#08)

It can be concluded that the ethical framework during AI validation and verification is crucial for developing a deeper understanding regarding the AI-based system and its related pitfalls. After the successful completion of the fourth phase, the workers are confident enough to scale the technology, even with the constraint of a slightly lower level of reliability and safety. The ethical framework concentrates on AI reliability and safety, which insured proper trust calibration. Thus, the trust in AI increased to a such high extent that the workers engage with AI in a nearly natural manner. This can be stated even though the level of reliability and safety had to be adjusted and lowered.

> "We could insist and say, 'It has to be perfect.' But if reality itself is not perfect, then we have to make a compromise." (#12)

During the evaluation of the ethical framework, further underlying measures for trust calibration were detected, which are discussed in the following section.

Supportive Measures Along the AI Lifecycle

The trust-building process is dynamic; as a consequence, the same holds for the underlying measures and mechanisms. Besides the impact of the ethical framework, five additional focus areas for underlying mechanisms are identified. These target different objectives along the AI lifecycle and are illustrated by the thick gray arrows in Figure 2: 1) organizational setting, 2) input data, 3) promotion, 4) performance and 5) robustness.

The identified mechanisms during the first phase are mainly related to the overall organizational setting. What can contribute to the trust calibration in this phase is 1) the high internal reputation of the AI development team or 2) the overall strategic focus of the organization on AI topics as well as 3) the comprehensive announcement of corresponding management decisions.

> "For our organization, AI has been of great strategic importance for many years. It got a great push through the foundation of the corporate internal AI campus." (#05)

Considerable responsibility is assigned to the workers in the second phase with respect to the input data. As an example, they are requested to define an individual domain expert, who has to define a test dataset, which is used as a baseline for the AI training. By this involvement, the responsibility of the workers regarding the input data becomes clear. It thereby helps to enhance the workers' understanding of the technology and the relevance of high-quality input data. As one outcome, e.g., the workers invented the label "unknown" for input data that they could not classify by themselves. Thus, the workers had to admit that they cannot provide the previously claimed high-reliability standard themselves.

> "The information, what data they want to use, from what time periods, on what lines, what products—that's up to the plant [the worker] to decide." (#01)

In this phase, it is also important to emphasize what kind of input data and, thus, which parts are not covered by the AI model. In the provided documents, it is evident that a precise documentation was made even for "out-of-scope packages," including the underlying reasons for that, e.g., due to the bad quality of input data. Thus, the workers are starting to get to know the boundaries of the AI-based system.

The relevant aspect within the third phase is promotion. It is essential to share the knowledge from the project teams and pilot workers with the extended user group, e.g., further plants. During the exchange and training meetings, the workers are educated in understanding the different errors. In this way, the workers are enabled to deal with them and find ways to fix inaccuracies. Consequently, the workers themselves are able to explain the new AI technology to outside parties.

> "I am very busy making sure that we bring this topic out of the project phase and into the normal process. This means that knowledge must be built up at the plant." (#01)

During the fourth phase, it has already been stated that the ethical framework makes a major contribution to the fundamental trust calibration. Further supportive measures regarding AI performance can be identified, like 1) the creation of a glossary, which includes the performance-related figures and their interpretation, or 2) the implementation of an appropriate risk management. Nevertheless, some ethical risks of potential issues remain. Different strategies for risk management need to be provided, like a detailed reaction plan including containment and corrective actions. It includes a profound documentation of different levels of countermeasures, e.g., on how to switch off the AI or how to exclude certain parts from AI.

> "So, the most important thing is that we have immediate measures to prevent the error from persisting, and then we have to somehow eliminate it in the long term." (#08)

The AI implementation develops toward AI operation within phase five, supported by mechanisms regarding the robustness. This includes, e.g., the constant adaptation of the model to new products. The workers must understand that the AI model needs constant improvement and training to keep it running. Moreover, the model needs to remain robust in terms of security and safety aspects.

> "There are changes regarding the products, new components, etc., so this results in subsequent effort. This is not just a one-time process." (#03)

DISCUSSION

Our research builds on studies focusing on ethical frameworks for AI[3,7,14] in combination with studies specifically focusing on workers' trust calibration.[4,5] For the first time, the common ground of AI reliability and AI safety is justified, and both perspectives are brought together. As an outcome, we investigate an ethical framework that is tailored to AI validation and verification in an organizational context for proper trust calibration. It turned out that workers are stepwise aware of the lower reliability and safety. Nonetheless, they do not refuse but, rather, engage and even promote the technology in the organization. Thus, the ethical framework leads to a major increase in trust, which comes along with a willingness of the workers to accept a lower level of AI reliability and safety. This major finding extends the current body of knowledge and is in clear contrast to a previous work, in which it was stated that workers might refuse the application of the technology if they cannot trust in its reliability and safety to the full extent.[18]

The current body of knowledge is further extended by the findings regarding the iterative interplay between ethical frameworks, trust calibration, and AI reliability and safety. Up to now, it has been reported in the literature that a continuous trust calibration during the AI lifecycle, by following an iterative and human-centered design, is mandatory.[2,14,19] In this study, a practical example that tackles such a continuous trust calibration is provided for the first time. We can state that the adoption of an ethical framework enhances both the explainability and robustness of AI models. This is highly relevant, as both factors, explainability and robustness, contribute to promote reliability and, thus, trust in the AI-based systems.[2] Reasons lie in the fact that the provision of analogous patterns on AI reliability and safety supports this trust calibration. For workers, the cause-and-effect relationship of the results becomes clear as the explainability is enhanced.[1] Moreover, in practice, the ethical framework serves as systematic risk management that tackles ethical concerns in a structured manner. [9,20]

Finally, the illustrated human-centered AI design process serves as a first practical example for the demanded hybrid approach of AI development. It combines the benefits of data-driven machine learning with human domain knowledge.[1] In this way, it is possible to ensure that humans retain control. As a consequence, the full advantage of AI is realized, and AI further contributes to discovering previously hidden connections.

LIMITATIONS AND FUTURE RESEARCH

The present study is subject to some limitations. The qualitative approach and the selective sample size result in restrictions on the overall generalizability of the results, as the data collection is limited to the domain-specific context with a focus on the automotive industry. To level that out, 17 interview participants were selected, considering various positions, working experience, and hierarchies. Basically, this can be extended by a future quantitative approach with the target of verifying the generalizability of the presented results.

Future research could expand our process model by integrating further factors for trust calibration. It might be reasonable to put the focus on other phases of the AI lifecycle and investigate suitable ethical frameworks for later phases, like operation.

CONCLUSION

It is well known that there is an increasing importance of AI within our life and within organizations. In the present study, for the first time, the implementation of an ethical framework for the dynamic process of trust calibration in AI was investigated, and a practical human-centered design process was developed. To combine the two perspectives of AI ethics and AI trust, we used AI reliability and AI safety as a common ground. Based on 17 interviews within an automotive supplier, the effectiveness of the ethical framework was evaluated.

The present article provides two main outcomes with respect to trust calibration by means of an ethical framework. First, it led to a major increase in the trust of the workers. This is due to the fact that it was tailored to the critical phase of AI verification and validation. Thus, the presented measures enable workers to explain and understand AI-generated decisions.[2] Second, it turned out that workers are willing to tolerate a lower level of AI reliability and safety as a function of their increasing experience with the AI-based system.

ACKNOWLEDGMENTS

This study was conducted in cooperation with Robert Bosch GmbH. The authors would like to thank all interview partners for their valuable input throughout the research process. Their insights and expertise were instrumental in shaping the direction of this project.
The work involved qualitative interviews with participants selected on a voluntary basis. During the entire data collection and analysis process, special attention was paid to respect the participants' confidentiality and anonymity. As a consequence, identifiable details were obscured and only selected sanitized parts of the data are shared.

REFERENCES

1. A. Holzinger and H. Muller, "Toward human–AI interfaces to support explainability and causability in medical AI," *Computer*, vol. 54, no. 10, pp. 78–86, Oct. 2021, doi: 10.1109/MC.2021.3092610.

2. A. Holzinger et al., "The next frontier: AI we can really trust," in *Communications in Computer and Information Science, Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, M. Kamp, Ed., Cham, Switzerland: Springer International Publishing, 2021, pp. 427–440.
3. J. Fjeld, N. Achten, H. Hilligoss, and A. Nagy, "Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI," Harvard Library, Cambridge, MA, USA, 2020. [Online]. Available: http://nrs.harvard.edu/urn-3:HUL.InstRepos:42160420
4. K. Okamura and S. Yamada, "Empirical evaluations of framework for adaptive trust calibration in human-AI cooperation," *IEEE Access*, vol. 8, pp. 220,335–220,351, 2020, doi: 10.1109/ACCESS.2020.3042556.
5. D. H. McKnight, M. Carter, J. B. Thatcher, and P. F. Clay, "Trust in a specific technology: An investigation of its components and measures," *ACM Trans. Manage. Inf. Syst.*, vol. 2, no. 2, pp. 1–25, Jul. 2011, doi: 10.1145/1985347.1985353.
6. X. Li, P. Ye, J. Li, Z. Liu, L. Cao, and F.-Y. Wang, "From features engineering to scenarios engineering for trustworthy AI: I&I, C&C, and V&V," *IEEE Intell. Syst.*, vol. 37, no. 4, pp. 18–26, Jul./Aug. 2022, doi: 10.1109/MIS.2022.3197950.
7. A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nature Mach. Intell.*, vol. 1, no. 9, pp. 389–399, Sep. 2019, doi: 10.1038/s42256-019-0088-2.
8. A. F. Salam, S. Pervez, and S. Nahar, "Trust in AI and intelligent systems: Central core of the design of intelligent systems," in *Proc. AMCIS*, 2021, p. 22.
9. S. Thiebes, S. Lins, and A. Sunyaev, "Trustworthy artificial intelligence," *Electron. Markets*, vol. 31, no. 2, pp. 447–464, Jun. 2021, doi: 10.1007/s12525-020-00441-4.
10. E. Glikson and A. W. Woolley, "Human trust in artificial intelligence: Review of empirical research," *Acad. Manage. Ann.*, vol. 14, no. 2, pp. 627–660, Jul. 2020, doi: 10.5465/annals.2018.0057.
11. K. A. Hoff and M. Bashir, "Trust in automation: Integrating empirical evidence on factors that influence trust," *Human Factors*, vol. 57, no. 3, pp. 407–434, 2015, doi: 10.1177/0018720814547570.
12. C. Allen, W. Wallach, and I. Smit, "Why machine ethics?" *IEEE Intell. Syst.*, vol. 21, no. 4, pp. 12–17, Aug. 2006, doi: 10.1109/MIS.2006.83.
13. J. Zhou, F. Chen, A. Berry, M. Reed, S. Zhang, and S. Savage, "A survey on ethical principles of AI and implementations," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Canberra, ACT, Australia, 2020, pp. 3010–3017, doi: 10.1109/SSCI47803.2020.9308437.
14. Q. Lu, L. Zhu, X. Xu, J. Whittle, and Z. Xing, "Towards a roadmap on software engineering for responsible AI," in *Proc. 1st Int. Conf. AI Eng., Softw. Eng. AI*, Pittsburgh, PA, USA, 2022, pp. 101–112, doi: 10.1145/3522664.3528607.
15. C. Ebert and M. Weyrich, "Validation of autonomous systems," *IEEE Softw.*, vol. 36, no. 5, pp. 15–23, Sep./Oct. 2019, doi: 10.1109/MS.2019.2921037.
16. B. G. Glaser and A. L. Strauss, *The Discovery of Grounded Theory: Strategies for Qualitative Research*. New York, NY, USA: Aldine, 1967.
17. R. Hoda, "Socio-technical grounded theory for software engineering," *IEEE Trans. Softw. Eng.*, vol. 48, no. 10, pp. 3808–3832, Oct. 2022, doi: 10.1109/TSE.2021.3106280.
18. N. Lankton, D. H. McKnight, and J. Tripp, "Technology, humanness, and trust: Rethinking trust in technology," *J. Assoc. Inf. Syst.*, vol. 16, no. 10, pp. 880–918, Oct. 2015, doi: 10.17705/1jais.00411.
19. M. Hengstler, E. Enkel, and S. Duelli, "Applied artificial intelligence and trust—The case of autonomous vehicles and medical assistance devices," *Technol. Forecasting Social Change*, vol. 105, pp. 105–120, Apr. 2016, doi: 10.1016/j.techfore.2015.12.014.
20. L. Y.-H. Chen, "A conceptual framework for AI system development and sustainable social equality," in *Proc. IEEE/ITU Int. Conf. Artif. Intell. Good (AI4G)*, Geneva, Switzerland, 2020, pp. 101–106, doi: 10.1109/AI4G50087.2020.9310984.

AMELIE SCHMID is a Ph.D. student at the Chair of Digital Transformation, TU Dortmund University, 44221, Dortmund, Germany, in collaboration with Robert Bosch GmbH, Reutlingen, Germany. Her research interests include the AI-based transformation of work, implications for AI ethics and trust, and the implementation of AI in practice. Schmid received her master's degree in organizational design from Nuertingen-Geislingen University, Germany. Contact her at amelie.schmid@tu-dortmund.de.

MANUEL WIESCHE is a full professor and chair of Digital Transformation at TU Dortmund University, 44221, Dortmund, Germany. His research interests include IT workforce, IT project management, digital platform ecosystems, and IT service innovation. Wiesche received his doctoral degree and habilitation degree from TUM School of Management, Technical University of Munich, Germany. Contact him at manuel.wiesche@tu-dortmund.de.