

## Designing for Appropriate Trust in Human-AI Interaction

Mehrotra, S.

**DOI**

[10.4233/uuid:5a0c475b-5494-4f7a-a91c-796975233d95](https://doi.org/10.4233/uuid:5a0c475b-5494-4f7a-a91c-796975233d95)

**Publication date**

2024

**Document Version**

Final published version

**Citation (APA)**

Mehrotra, S. (2024). *Designing for Appropriate Trust in Human-AI Interaction*. [Dissertation (TU Delft), Delft University of Technology]. <https://doi.org/10.4233/uuid:5a0c475b-5494-4f7a-a91c-796975233d95>

**Important note**

To cite this publication, please use the final published version (if applicable).  
Please check the document version above.

**Copyright**

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

**Takedown policy**

Please contact us and provide details if you believe this document breaches copyrights.  
We will remove access to the work immediately and investigate your claim.

# **DESIGNING FOR APPROPRIATE TRUST IN HUMAN-AI INTERACTION**



# DESIGNING FOR APPROPRIATE TRUST IN HUMAN-AI INTERACTION

## Dissertation

for the purpose of attaining the degree of doctor  
at Delft University of Technology,  
by the authority of Rector Magnificus prof. dr. ir. T.H.J.J. van der Hagen,  
chair of the Board for Doctorates,  
to be defended publicly on  
6 September 2024, 10.00 hours

by

**Siddharth MEHROTRA**

Master of Science in Media Informatics,  
RWTH Aachen University, Germany,  
born in Shahjahanpur, India

This dissertation has been approved by the promotor.

Composition of the doctoral committee:

|                       |   |
|-----------------------|---|
| Rector Magnificus     | chairperson                                       |
| Prof. dr. C.M. Jonker | Delft University of Technology, <i>promotor</i>   |
| Dr. M.L. Tielman      | Delft University of Technology, <i>copromotor</i> |

Independent members:

|                             |   |
|-----------------------------|---|
| Prof. dr. S. Roeser         | Delft University of Technology                        |
| Prof. dr. G.C.H.E. de Croon | Delft University of Technology                        |
| Prof. dr. P. Yolum Birbil   | Utrecht University                                    |
| Prof. dr. M. Winikoff       | Victoria University of Wellington, New Zealand.       |
| Prof. dr. M.A. Neerincx     | Delft University of Technology, <i>reserve member</i> |

SIKS Dissertation Series No. 2024-33. The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.



**Keywords:** Trust, Appropriate Trust, Values, Ethics, and Explainable AI

**Printed by:** Gildeprint

**Cover design by:** Dr. Agathe Balyan and Dr. David Maxwell  
*Krishna (the modern times AI systems) guiding Arjuna (the user) how to reach the kingdom by constructing the bridge. Arjuna carefully following Krishna's advice to trust his decision shows appropriate trust in Krishna.*  
Image of Krishna and Arjuna used under the terms of the Wikimedia Commons License, from thedawnwithin.com. Typeset in Laila by the Indian Type Foundry, under the terms of the Open Font License.

**Style:** TU Delft House Style, with modifications by Moritz Beller

An electronic version of this dissertation is available at <http://repository.tudelft.nl/>.

# CONTENTS

|  |             |
|--|-------------|
| <b>Summary</b>   | <b>ix</b>   |
| <b>Samenvatting</b>  | <b>xi</b>   |
| <b>Summary in Hindi</b>  | <b>xiii</b> |
| <b>1 Introduction</b>  | <b>1</b>    |
| 1.1 Motivation . . . . .   | 2           |
| 1.2 Main Research Question . . . . .   | 3           |
| 1.2.1 Approach and Sub-questions . . . . .   | 4           |
| 1.3 Three Lenses: Formal, Social, and Application . . . . .                          | 5           |
| 1.4 Scoping this thesis . . . . .  | 10          |
| 1.5 Contributions . . . . .  | 10          |
| 1.6 Thesis Organization . . . . .  | 12          |
| <b>I Formal Lens</b>   | <b>15</b>   |
| <b>2 Building Appropriate Trust in AI systems – A Formal Perspective</b>             | <b>17</b>   |
| 2.1 Introduction . . . . .   | 18          |
| 2.2 Trust as a belief of Trustworthiness . . . . .                                   | 19          |
| 2.3 Appropriate Trust . . . . .  | 21          |
| 2.4 How are trust beliefs established? . . . . .                                     | 23          |
| 2.4.1 Ability . . . . .  | 23          |
| 2.4.2 Integrity . . . . .  | 23          |
| 2.4.3 Benevolence . . . . .  | 24          |
| 2.5 Summary of Contributions . . . . .   | 25          |
| 2.6 Limitations . . . . .  | 25          |
| 2.7 Follow-up Discussions and Thesis Positioning . . . . .                           | 26          |
| <b>3 A Systematic Review on Appropriate Trust in Human-AI Interaction</b>            | <b>29</b>   |
| 3.1 Introduction . . . . .   | 30          |
| 3.2 Background and History of Appropriate Trust . . . . .                            | 32          |
| 3.2.1 1980-1990s: Over- and under-trust in automation . . . . .                      | 33          |
| 3.2.2 1990s: Introduction of HCI as a field and focus on appropriate trust . . . . . | 34          |
| 3.2.3 2000s: Emergence of appropriate trust as a key topic of research . . . . .     | 34          |
| 3.2.4 Parallel Developments: Influential Domains . . . . .                           | 35          |
| 3.3 Systematic Review Methodology . . . . .  | 36          |
| 3.3.1 Search String . . . . .  | 37          |
| 3.3.2 Selection Criteria . . . . .   | 38          |
| 3.3.3 Corpus Overview and Analysis . . . . .   | 39          |

|                       |   |           |
|-----------------------|---|-----------|
| 3.4                   | Definitions and Related Concepts . . . . .                                | 40        |
| 3.4.1                 | A Belief-Intentions-Actions (BIA) Mapping. . . . .                        | 40        |
| 3.5                   | Results of the Systematic Review. . . . .                                 | 44        |
| 3.5.1                 | Measures (How to measure Appropriate Trust?) . . . . .                    | 45        |
| 3.5.2                 | Tasks. . . . .  | 48        |
| 3.5.3                 | Methods for building appropriate trust (How to achieve it?) . . . . .     | 50        |
| 3.5.4                 | Results of calibration interventions. . . . .                             | 55        |
| 3.6                   | Discussion . . . . .  | 58        |
| 3.6.1                 | Key Challenges. . . . .   | 58        |
| 3.6.2                 | Open Questions . . . . .  | 61        |
| 3.6.3                 | Novel perspectives . . . . .  | 62        |
| 3.6.4                 | Limitations. . . . .  | 62        |
| 3.6.5                 | Summary. . . . .  | 63        |
| 3.7                   | Conclusion. . . . .   | 66        |
| <b>II Social Lens</b> |   | <b>69</b> |
| <b>4</b>              | <b>Effect of Value Similarity on Trust in Human-Agent Interaction</b>     | <b>71</b> |
| 4.1                   | Introduction . . . . .  | 72        |
| 4.2                   | Related Work. . . . .   | 73        |
| 4.3                   | Method. . . . .   | 74        |
| 4.3.1                 | Creation of value profiles. . . . .                                       | 74        |
| 4.3.2                 | Agents and the scenario . . . . .   | 75        |
| 4.3.3                 | Participants . . . . .  | 76        |
| 4.3.4                 | User study test bed. . . . .  | 77        |
| 4.3.5                 | Procedure . . . . .   | 77        |
| 4.4                   | Results . . . . .   | 77        |
| 4.4.1                 | Manipulation check . . . . .  | 77        |
| 4.4.2                 | Correlation between Value Similarity and overall Trust . . . . .          | 78        |
| 4.4.3                 | Benevolence, and Willingness as attributes of overall trust . . . . .     | 79        |
| 4.4.4                 | Qualitative data analysis . . . . .                                       | 79        |
| 4.5                   | Discussion . . . . .  | 80        |
| 4.5.1                 | Why were our manipulations unsuccessful? . . . . .                        | 81        |
| 4.5.2                 | Trust in AI systems. . . . .  | 82        |
| 4.5.3                 | Limitations & Future Work. . . . .  | 83        |
| 4.6                   | Conclusion. . . . .   | 83        |
| <b>5</b>              | <b>Integrity Based Explanations for Fostering Appropriate Trust in AI</b> | <b>85</b> |
| 5.1                   | Introduction . . . . .  | 86        |
| 5.2                   | Appropriate Trust . . . . .   | 89        |
| 5.2.1                 | Prior work on Appropriate Trust. . . . .                                  | 89        |
| 5.2.2                 | Our Approach on Appropriate Trust - a formal perspective . . . . .        | 90        |
| 5.3                   | Integrity . . . . .   | 92        |
| 5.3.1                 | Prior work on Integrity. . . . .  | 92        |
| 5.3.2                 | Our Approach on Integrity: Integrity-laden Explanations . . . . .         | 93        |

- 5.4 Method. . . . . 98
  - 5.4.1 Participants . . . . . 98
  - 5.4.2 Task Design . . . . . 98
  - 5.4.3 Measures. . . . . 100
  - 5.4.4 Experimental Setup . . . . . 102
  - 5.4.5 Procedure . . . . . 103
- 5.5 Results . . . . . 103
  - 5.5.1 Effect of different principles of integrity on Appropriate Trust . . . 103
  - 5.5.2 Effect of different principles of integrity on Subjective Trust . . . 106
  - 5.5.3 Effect of different principles of integrity on usefulness . . . . . 108
- 5.6 Discussion . . . . . 111
  - 5.6.1 Expressions of Integrity and Appropriate Trust. . . . . 111
  - 5.6.2 Subjective trust, helpfulness and comfort. . . . . 112
  - 5.6.3 Understanding human psychology advice utilization. . . . . 113
  - 5.6.4 Reflections on design considerations for building appropriate trust. 114
  - 5.6.5 Limitations and Future Work. . . . . 115
- 5.7 Conclusion. . . . . 116

**III Application Lens 117**

**6 Fostering Appropriate Trust in AI-based Predictive Policing Systems 119**

- 6.1 Introduction . . . . . 120
- 6.2 Background and Related Work . . . . . 122
  - 6.2.1 Appropriate Trust . . . . . 122
  - 6.2.2 AI-based Predictive Policing & Trust . . . . . 123
- 6.3 Study Design. . . . . 124
  - 6.3.1 Designing AI system’s Explanations . . . . . 124
  - 6.3.2 Traditional Investigation Notes. . . . . 126
- 6.4 First User Study - "Expert Users" . . . . . 126
  - 6.4.1 Participants . . . . . 126
  - 6.4.2 Methodology. . . . . 126
  - 6.4.3 Results . . . . . 130
- 6.5 Second User Study - "Lay Users" . . . . . 131
  - 6.5.1 Methodology. . . . . 131
  - 6.5.2 Results . . . . . 131
- 6.6 Discussion . . . . . 133
  - 6.6.1 Effect of explanations on appropriate and subjective trust . . . . . 133
  - 6.6.2 Usefulness of explanations, user expertise & exploratory measures. 134
  - 6.6.3 Policy Implications - Challenges and Recommendations . . . . . 135
  - 6.6.4 Limitations and Future Work. . . . . 136
- 6.7 Conclusion. . . . . 136
- 6.8 Impact Statements . . . . . 137
  - 6.8.1 Ethical Considerations Statement . . . . . 137
  - 6.8.2 Researcher Positionality Statement. . . . . 137
  - 6.8.3 Adverse Impact Statement . . . . . 138

|           |   |            |
|-----------|---|------------|
| <b>IV</b> | <b>Conclusions</b>  | <b>139</b> |
| 7         | <b>Conclusions</b>  | <b>141</b> |
| 7.1       | Conclusions . . . . .   | 142        |
| 7.2       | Limitations. . . . .  | 144        |
| 7.2.1     | Complexity of Synthesis . . . . .   | 144        |
| 7.2.2     | Reductionist Perspective . . . . .  | 145        |
| 7.3       | Contributions . . . . .   | 146        |
| 7.3.1     | Scientific. . . . .   | 146        |
| 7.3.2     | Societal. . . . .   | 147        |
| 7.4       | Future Work . . . . .   | 149        |
| 7.5       | Take-home message . . . . .   | 150        |
| <b>V</b>  | <b>Appendices</b>   | <b>151</b> |
| <b>A</b>  | <b>Effect of Value Similarity on Trust in Human-Agent Interaction</b>       | <b>153</b> |
| A.1       | Algorithm 1 . . . . .   | 154        |
| A.2       | VSQ and HCTS. . . . .   | 154        |
| <b>B</b>  | <b>Integrity Based Explanations for Fostering Appropriate Trust in AI</b>   | <b>155</b> |
| B.1       | Appendix 1. . . . .   | 155        |
| B.2       | Appendix 2. . . . .   | 157        |
| B.3       | Appendix 3. . . . .   | 157        |
| <b>C</b>  | <b>Fostering Appropriate Trust in AI-based Predictive Policing Systems:</b> | <b>161</b> |
| C.1       | Preliminary Study . . . . .   | 161        |
| C.1.1     | Method of the Preliminary Study. . . . .                                    | 161        |
| C.1.2     | Insights from the Preliminary Study . . . . .                               | 161        |
| C.2       | Pilot Usability Study . . . . .   | 162        |
| C.3       | Additional Figures . . . . .  | 162        |
|           | <b>Bibliography</b>   | <b>167</b> |
|           | <b>SIKS Dissertations</b>   | <b>203</b> |
|           | <b>Acknowledgments</b>  | <b>215</b> |
|           | <b>Curriculum Vitæ</b>  | <b>219</b> |
|           | <b>List of Publications</b>   | <b>221</b> |

## SUMMARY

Trust is essential to any interaction, especially when interacting with technology that does not (*metaphorically*) think like we do. Nowadays, many AI systems are being developed that have the potential to make a difference in people's lives, from health apps to robot companions. However, to reach their potential, people need to have appropriate levels of trust in these AI systems, *i.e.*, people should not over- or under-trust AI as it can lead to misuse and disuse. Therefore, AI systems need to understand how humans trust them and what to do to promote appropriate trust.

In this research, as a first step towards eliciting appropriate trust, we must understand what factors influence trust in AI agents. Despite the growing attention in research on trust in AI agents, a lot is still unknown about people's perceptions of trust in AI agents. Therefore, this research studied what makes people trust or distrust AI.

Additionally, as mentioned above, human's trust in the AI must be appropriate. The challenge is to ensure that humans tune their trust in the AI agent since we do not have control over humans. Therefore, in this research, we leverage the idea that if AI agents can reflect on their own trustworthiness through explanations, we may be able to influence humans to fine-tune their trust in them appropriately. With the information regarding the AI agent's trustworthiness, a human can adapt to the qualities and limitations of the AI agent and, consequently, adjust the utilization of the agent accordingly.

The topic of this thesis relates to hybrid intelligence, meaning mutual trust is crucial for effective human-AI interaction. To do this, in this thesis, we developed artificial agents that can reason about and promote appropriate mutual trust.

To explore our research questions, this thesis makes use of three lenses namely: a formal, a social and an application lens. This methodological approach ensured a holistic exploration of appropriate human trust, drawing on formal theories, social considerations, and practical insights.

The formal lens delves into the technical intricacies by understanding the logic of building appropriate trust in AI systems. The social lens shifts focus to the human element, encompassing the ways in which humans interact with and rely on AI systems in various aspects of their lives. The application lens situates us in real-world contexts, enabling the development of tailored solutions and guidelines that cater to the unique appropriate trust requirements of various domains, from healthcare to law and beyond.

Overall, by combining these the lenses this thesis presents a holistic overview of designing for appropriate trust in human – AI interactions.



---

## SAMENVATTING

Vertrouwen is essentieel voor iedere interactie, vooral als het om interactie gaat met technologie die niet (*metaphorically*) denkt zoals wij. Tegenwoordig worden er veel AI-systemen ontwikkeld die de potentie hebben om een verschil te maken in het leven van mensen, van gezondheidsapps tot gezelschapsrobots. Om hun potentieel te verwezenlijken moeten mensen echter een passend niveau van vertrouwen hebben in deze AI-systemen, dat wil zeggen dat mensen AI niet te veel of te weinig moeten vertrouwen, aangezien dit tot misbruik en onbruik kan leiden. AI-systemen moeten dus begrijpen hoe mensen hen vertrouwen en wat ze moeten doen om passend vertrouwen te bevorderen.

In dit onderzoek moeten we, als eerste stap op weg naar het wekken van passend vertrouwen, begrijpen welke factoren het vertrouwen in AI-agenten beïnvloeden. Ondanks de groeiende aandacht in onderzoek naar vertrouwen in AI-agenten, is er nog steeds veel onbekend over de perceptie van mensen over vertrouwen in AI-agenten. Dit onderzoek onderzocht daarom wat ervoor zorgt dat mensen AI vertrouwen of wantrouwen.

Bovendien moet, zoals hierboven vermeld, het vertrouwen van de mens in de AI passend zijn. De uitdaging is om ervoor te zorgen dat mensen hun vertrouwen in de AI-agent afstemmen, aangezien we geen controle over mensen hebben. In dit onderzoek maken we daarom gebruik van het idee dat wanneer AI-agenten kunnen nadenken over hun eigen betrouwbaarheid, we mensen mogelijk kunnen beïnvloeden om hun vertrouwen in hen op de juiste manier te verfijnen. Met de informatie over de betrouwbaarheid van de AI-agent kan een mens zich aanpassen aan de kwaliteiten en beperkingen van de AI-agent en daarmee de inzet van de agent overeenkomstig aanpassen.

Het onderwerp van dit proefschrift heeft betrekking op hybride intelligentie, wat betekent dat wederzijds vertrouwen cruciaal is voor effectieve mens-AI-interactie. Om dit te doen hebben we in dit proefschrift kunstmatige agenten ontwikkeld die kunnen redeneren over passend wederzijds vertrouwen en dit kunnen bevorderen.

Om onze onderzoeksvragen te onderzoeken, heeft dit proefschrift gebruik gemaakt van drie lenzen, namelijk: een formele, een sociale en een toepassingslens. Deze methodologische aanpak zorgde voor een holistische verkenning van passend menselijk vertrouwen, waarbij gebruik werd gemaakt van formele theorieën, sociale overwegingen en praktische inzichten.

De formele lens duikt in de technische complexiteit door de logica te begrijpen van het opbouwen van passend vertrouwen in AI-systemen. De sociale lens verschuift de focus naar het menselijke element en omvat de manieren waarop mensen in verschillende aspecten van hun leven omgaan met en vertrouwen op AI-systemen. De toepassingslens plaatst ons in de context van de echte wereld, waardoor de ontwikkeling van op maat gemaakte oplossingen en richtlijnen mogelijk wordt gemaakt die tegemoetkomen aan de unieke passende vertrouwensvereisten van verschillende domeinen, van gezondheidszorg tot recht en daarbuiten.

Door deze lenzen te combineren presenteert dit proefschrift een holistisch overzicht van ontwerpen voor passend vertrouwen in mens-AI-interacties.

## SUMMARY IN HINDI

विश्वास किसी भी बातचीत के लिए आवश्यक है, विशेष रूप से जब हम ऐसी तकनीक के साथ बातचीत कर रहे हों जो हमारी तरह नहीं सोचती। आजकल, कई कृत्रिम बुद्धिमत्ता (एआई) प्रणालियाँ विकसित की जा रही हैं जो लोगों के जीवन में बदलाव ला सकती हैं, जैसे स्वास्थ्य ऐप्स और रोबोट साथी। हालाँकि, अपनी पूरी क्षमता तक पहुँचने के लिए, लोगों को इन एआई प्रणालियों पर उचित स्तर का विश्वास होना चाहिए। न तो बहुत अधिक और न ही बहुत कम विश्वास होना चाहिए, क्योंकि यह दुरुपयोग या अनुपयोग का कारण बन सकता है।

इस शोध में, हम यह समझने की कोशिश करते हैं कि लोग एआई एजेंट्स पर कैसे और क्यों विश्वास करते हैं। हालाँकि एआई एजेंट्स पर विश्वास के बारे में शोध बढ़ रहा है, लेकिन अभी भी लोगों की धारणाओं के बारे में बहुत कुछ अज्ञात है। इसलिए, यह शोध यह अध्ययन करता है कि लोग एआई पर क्यों विश्वास करते हैं या क्यों नहीं करते हैं।

इसके अतिरिक्त, जैसा कि ऊपर उल्लेख किया गया है, मनुष्य का एआई पर विश्वास उचित होना चाहिए। चुनौती यह सुनिश्चित करने की है कि मनुष्य एआई एजेंट पर अपना विश्वास समायोजित करें, क्योंकि हमारा मनुष्यों पर नियंत्रण नहीं है। इसलिए, इस शोध में, हम इस विचार का लाभ उठाते हैं कि यदि एआई एजेंट स्पष्टीकरण के माध्यम से अपनी विश्वसनीयता पर विचार कर सकते हैं, तो हम मनुष्यों को उन पर अपना विश्वास उचित रूप से समायोजित करने के लिए प्रभावित कर सकते हैं। एआई एजेंट की विश्वसनीयता के बारे में जानकारी के साथ, एक मनुष्य एआई एजेंट के गुणों और सीमाओं के अनुरूप अपने को ढाल सकता है और परिणामस्वरूप, एजेंट के उपयोग को उसी के अनुसार समायोजित कर सकता है।

इस शोध का विषय हाइब्रिड इंटेलिजेंस से संबंधित है, जिसका अर्थ है कि प्रभावी मानव-एआई संवाद के लिए आपसी विश्वास महत्वपूर्ण है। इसे करने के लिए, इस शोध में, हमने ऐसे कृत्रिम एजेंट विकसित किए हैं जो उचित आपसी विश्वास के बारे में तर्क कर सकते हैं और उसे बढ़ावा दे सकते हैं।

औपचारिक दृष्टिकोण एआई प्रणालियों में उचित विश्वास निर्माण के तर्क को समझकर तकनीकी जटिलताओं में गहराई से जाता है। सामाजिक दृष्टिकोण मानवीय तत्व पर ध्यान केंद्रित करता है, जिसमें मनुष्य अपने जीवन के विभिन्न पहलुओं में एआई प्रणालियों के साथ कैसे बातचीत करते हैं और उन पर निर्भर करते हैं, यह शामिल है। अनुप्रयोग दृष्टिकोण हमें वास्तविक दुनिया के संदर्भों में स्थित करता है, जो स्वास्थ्य सेवा से लेकर कानून और उससे आगे तक विभिन्न क्षेत्रों की अनूठी उचित विश्वास आवश्यकताओं के अनुरूप समाधान और दिशानिर्देश विकसित करने में सक्षम बनाता है।

कुल मिलाकर, इन दृष्टिकोणों को मिलाकर यह शोध मानव-एआई संवाद में उचित विश्वास के डिजाइन का एक समग्र अवलोकन प्रस्तुत करता है।



# 1

## INTRODUCTION

सर्व परवशं दुःखं सर्वमात्मवशं सुखम् ।

आ नो भद्राः क्रतवो यन्तु विश्वतः ।

एतद् विद्यात् समासेन लक्षणं सुखदुःखयोः ॥

*“Everything that is in another’s control is hard to trust. All that is in self-control is happiness. Let noble thoughts to understand the meaning of trusting others come to me from all directions.” - Bhagavad Gita [305]*

The above verses from the Bhagavad Gita (a Hindu scripture) emphasise self-control and inner strength for finding happiness and trust. They suggest that relying on external factors or depending too much on others for trust can be challenging, as these factors are not within our control. Instead, the path to happiness and trust lies in nurturing noble thoughts and understanding the significance of trusting others through one’s own inner qualities and self-control.

In a world where many factors lie beyond our control, trust becomes an anchor amidst the turbulent sea of uncertainties. Trusting wisely, especially in developing interpersonal relationships, is like exercising self-control. By discerning and embracing the risk associated with trusting others, we can navigate the complexities of trust in a way that leads to happiness and meaningful connections [345]. Trust in building social relationships involves assessing the elements of risk and vulnerability. However, this concept transcends the boundaries of interpersonal relationships [160].

In today’s society, where Artificial Intelligence (AI) has gained a vital role, it is of paramount importance to grasp and employ the concept of trust judiciously. Our trust in AI systems has profound implications for our daily interactions [236]. In this thesis, we are motivated by the rise of many successful applications of AI systems that guide our interactions with the ever-evolving world of technology, with our overarching objective of ensuring human’s appropriate trust in AI systems.

Although there are many ways to define appropriate trust [255], in this thesis we take this to mean that the trust a human has in a system is aligned with the actual trustworthiness of the system [114]. We argue that human and AI system together should strive for appropriate trust of the human in the AI system because, with appropriate trust in AI, people may be simultaneously aware of the potential and limitations of AI. Thus, with appropriate trust, they will use AI wisely and appropriately.

## 1.1 MOTIVATION

Since its birth in the 1950’s, AI has proliferated as a research field. Using different approaches, such as logic-based expert systems since the 80’s and 90’s [158, 296, 411] and machine learning since its early days [181, 268], the field has attempted to “not just understand but also to build intelligent entities” [317]. In our current age of the 2020s, AI systems

can automatically learn relevant patterns from data relieving humans of the burden of manually expressing these patterns in a formal language [168].

Many successful applications of AI algorithms have influenced people in their everyday lives. Some software examples are health care diagnostics [178] and drug discovery [235], automated financial investments [16], and recommender systems that learn your preference to recommend products or services [300]. Following the popularity of smartphones, many people carry AI applications with them daily in their pockets, such as voice assistants [234] and intelligent map planners [285]. Hardware applications have also increased, such as autonomous factory robots working assembly lines [322], consumer robots that vacuum clean your house [130], and cars that drive themselves [342].

Intelligent AI applications provide services and products to the general public. However, there have also been unintended negative side-effects of using AI in practice. For example, a Boeing Max737 aircraft crashed in Indonesia where the pilot over-trusted the auto-pilot mode [303] and defence personnel under-trusted the warning of an autonomous anti-ballistic missile which landed on a busy neighbourhood, causing the loss of hundreds of lives [110]. Both over-trust and under-trust of AI-embedded systems by humans have led to severe issues [155]. For example, Amazon's AI recruiting tool being biased against women's recruitment for leadership roles [13], a railroad accident in which the crew neglected speed constraints [346], and the use of facial recognition technology in law enforcement to target Black and Latino communities [271]. One of the major reasons of disuse and misuse of AI is people's over- or under-trust in it, or in other words, lack of appropriate trust in AI [256].

Obviously, designers of computer systems, want their users to trust their systems, as they do their best to create trustworthy systems. Hence the objective of designing computer systems or algorithms for increasing human trust [51], which spans at least the last five decades [233]. Following the existing literature on building trust in computer systems, various researchers have adopted the learning to design for increasing human trust in AI systems [368]. However, with a mere focus on increasing human trust in AI systems, we ignore the fact that these systems can fail or behave unpredictably, introducing the risk of misuse (over-trust [294]) and disuse (neglect or under-trust in AI [293]). Therefore, there is a gap in the current literature, which focuses on how we can design for appropriate trust in AI.

## 1.2 MAIN RESEARCH QUESTION

Appropriate trust is a complex topic as it requires consideration of context's influence, the AI system's goal-related characteristics, and the cognitive processes that govern the development and erosion of trust [61]. The current landscape of AI research predominantly emphasises building trust without delving into the subtleties of appropriateness. This research gap represents a critical juncture where we aspire to bridge the divide and offer insights into the nuanced world of designing for appropriate trust, ensuring that AI systems align harmoniously with human needs and aspirations. Therefore, in this thesis we explore the following main research question:

## How can we design for appropriate human trust in human-AI interaction?

### 1.2.1 APPROACH AND SUB-QUESTIONS

To explore our main research question, this thesis will make use of three lenses namely: a formal, a social and an application lens. This methodological approach ensures a holistic exploration of appropriate human trust, drawing on formal theories, social considerations, and practical insights. This comprehensive approach enables a nuanced understanding of the multifaceted nature of trust in human-AI interactions, facilitating the development of effective design recommendations and strategies that align with the complexities of real-world contexts.

*The formal lens* delves into the technical intricacies by understanding the logic of building appropriate trust in AI systems. A formal perspective is crucial for comprehending AI trust, with models like Belief, Desire, and Intentions (BDI) [306] and Ability, Benevolence, and Integrity (ABI) [242] serving as essential tools for understanding the intricacies of trust within AI systems.

*The social lens* shifts focus to the human element, encompassing the ways in which humans interact with and rely on AI systems in various aspects of their lives. As human trust is inherently rooted in the social sphere, it necessitates an exploration of the social aspects of trust.

*The application lens* situates us in real-world contexts, enabling the development of tailored solutions and guidelines that cater to the unique appropriate trust requirements of various domains, from healthcare to law and beyond. This lens helps us bridge the insights gained from both the formal and social lenses to specific domain-specific tasks involving Human-AI interaction.

Our aim is to approach the main RQ by formalizing appropriate trust for AI systems to understand it, taking into account the social factors influencing it, and how such formalisms can be used when humans interact with AI agents. This multi-faceted approach ensures a comprehensive understanding of the topic. The formal lens allows us to analyze the intricate dynamics of appropriate trust within AI systems with formal methods, providing a structured way to define and measure it. The social lens uncovers the intricate web of human perceptions, subjective biases, and actions that shape our trust in technology, ensuring our solutions resonate with the human experience.

Lastly, the application lens brings a real-world context, identifying specific areas where fostering appropriate trust in AI is paramount, guiding us to develop practical solutions tailored to the diverse needs. Taken together, these three lenses can help in exploring our main research question (refer Figure 1) and provide a holistic framework for understanding appropriate trust in the context of AI, ranging from its formal foundations to its social and practical implications.

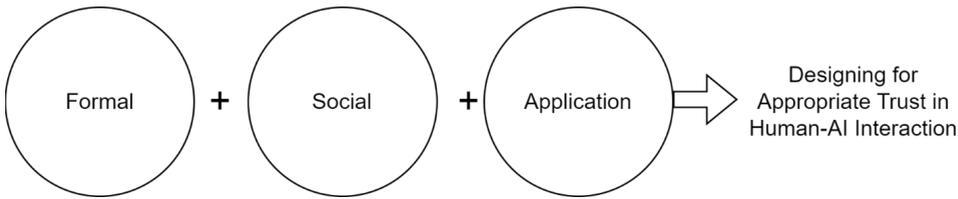


Figure 1.1: The three lenses: Formal, Social & Application used in this study.

By applying these lenses, our main research question is divided into different sub-questions. The following describes these sub-questions and how they are supported and motivated by findings and theories reported in the literature.

### 1.3 THREE LENSES: FORMAL, SOCIAL, AND APPLICATION

#### FORMAL LENS: FORMAL UNDERSTANDING AND BACKGROUND

##### *Formal Understanding*

We posit that designing for appropriate trust is ineffective without understanding first what we mean by it. The formal lens we use provides a structured and systematic approach to defining and conceptualizing trust in human-AI interaction, which is foundational for developing a clear and precise understanding of appropriate trust. Firstly, we undertake a rigorous analysis and definition of trust concepts. By employing formal methodologies, we can establish clear and unambiguous definitions, fostering a shared understanding among researchers. Secondly, the formal lens enables the development of models and frameworks that can be systematically tested and validated. This enhances the reliability and reproducibility of research findings. Furthermore, a formal approach facilitates the identification of variables and parameters that influence trust dynamics facilitating their operationalization in a structured manner. Therefore, we posit our first research question as:

**RQ1: How can we formally define appropriate human trust in human-AI interactions?**

As mentioned before, human trust in an AI agent should be as appropriate as possible. The challenge is how to encourage that humans tune their trust towards an AI agent. Metaphorically speaking, by embracing the challenge of designing AI agents to introspect and communicate on their own trustworthiness and adjusting their behavior accordingly, we can exert an influence on humans to fine-tune their trust appropriately. With the information regarding the agent's trustworthiness, a human can adapt to the qualities and limitations of the agent and, consequently, adjust the utilization of the AI agent accordingly. Without this knowledge, it would be difficult to coordinate within a task environment which requires a high level of mutual understanding and adaptability [320].

The initial phase of our exploration involved studying trust as a belief of trustworthiness. In this phase we first endeavoured to understand trustworthiness, and secondly how beliefs of trustworthiness are formed. By synthesizing theoretical constructs, we adopted the ABI

model [242] to define the trust of an agent  $x$  in agent  $y$  as the belief  $B$  of agent  $x$  regarding the trustworthiness of  $y$  with respect to  $x$ .

Based on the theoretical constructs and our definition of trust, we established that in human-AI interaction it is important that 1) the agent appropriately trusts the human and 2) the human appropriately trusts the agent, but it is also important that 3) the agent has a belief about whether the human appropriately trusts the agent, and that 4) the human believes the agent appropriately trusts the human, and why. We schematized these notions through formalism of nested beliefs and proposed to view them in a human-AI agent dyadic relationship in Chapter 2.

Once we formally defined appropriate trust, we studied mutual trust between an AI agent and a human as a dyadic relationship, including belief formation in shaping human trust. Subsequently, our focus shifted to understanding agent beliefs, wherein we sought to define an AI agent's integrity and benevolence. By establishing a theoretical foundation concerning the ethical principles and their congruence with human values, this exploration laid the essential groundwork for comprehending one of the components of appropriate trust in AI systems. Finally, we synthesize these distinct yet interconnected strands of inquiry, offering a cohesive narrative that explicates the integral components and relationships in formally defining appropriate trust.

Our exploration serves as a crucial bridge, connecting the theoretical foundations of defining appropriate trust, such as the need for humans to be able to form beliefs, to the practical aspects of how individuals perceive and trust AI agents in real-world interactions. The insights derived from this examination not only contribute directly to the broader objective of formalizing trust in human-AI interactions but also emphasize the importance of understanding the cognitive processes involved, including belief formation, in shaping dyadic trust. Furthermore, we consider all the aspects that shape dyadic trust: the roles (trustee/trustor) of humans and AI agents and the different perceptions (beliefs) about each team member's trustworthiness.

### ***A Systematic Review***

To utilize the formal lens of understanding human-AI interactions, we must situate our inquiry within the broader context of existing research. The formulation of research questions around the formal definition of appropriate trust, the integrity and benevolence of AI systems in building appropriate trust have provided a structured approach to understanding the complexities inherent in human-AI interactions. Therefore, it becomes imperative to acknowledge the multifaceted landscape of trust in human-AI interaction within the realm of formal analysis. Thus, a comprehensive background analysis is warranted to grasp the current state of the formal aspects of the field and the diverse methodologies employed to foster appropriate trust.

To achieve an appropriate level of trust in human-AI interaction, different approaches have been taken in the literature with mixed results. Through empirical studies, researchers from multiple scientific fields have implemented various strategies for trust calibration [402]. The level of trust placed in a system can be aligned to some extent with the perceived trustworthiness of that system [56, 247, 406]. Despite these endeavours, the field lacks a comprehensive understanding, and consensus regarding the definition of appropriate trust. This lack of clarity is compounded by different perspectives, diverse definitions, and varied concepts related to appropriate trust, trustworthiness, and appropriate reliance,

as highlighted by Vereschak et al. [379]. A structured literature review can provide an overview of the state of the art of the seemingly scattered nature of the field and different terms associated with appropriate trust. Research gaps can be identified on the basis of such a structured literature review. This lead us to the second research question:

**RQ2: What's state-of-the-art in fostering appropriate trust in AI systems?**

Studying the state-of-the-art of a field requires examining its evolution, definitions, related concepts, measures, and methods. Therefore, we adopted the formal lens to conduct a systematic review to answer our research question. We formulated an extensive literature review of former and current trust models in building appropriate trust in AI systems to scope the field. This analysis not only illuminates the existing landscape but also provides a structured foundation for a better conceptual understanding of appropriate trust, enabling the identification of crucial elements and gaps that contribute to a more nuanced and refined formalization. Furthermore, this review lead to a better conceptual understanding of appropriate trust, and a formalization of such. Based on the findings of the literature review, we analysed how we could employ these models to promote appropriate trust in the agents, and what is still missing.

## **SOCIAL LENS: VALUES & INTEGRITY OF AI SYSTEMS**

### ***Human & AI Agent Values***

The social lens is crucial for revealing the need for AI systems to have a relationship with people [289]. This lens focuses our attention on the profoundly human aspects of this interaction, encompassing not only the ways in which individuals interact with and rely on AI systems but also on beliefs, values, and societal factors that shape the foundation of trust.

As we delve into the social lens, we aim to understand the nuanced and multifaceted nature of human trust in AI, recognizing its dynamic interplay with the human experience and the broader societal context. After all, studying trust through a social lens is crucial for constructing appropriate trust in human-AI interactions because it provides a foundational understanding of the dynamics that govern these interactions. We can then move on to studying, appropriate trust, which per definition builds upon the concept of trust.

We see human trust as a multi-dimensional concept as suggested by Roff and Danks [315]. On the one hand, trust corresponds to reliability and/or predictability and on the other hand trust depends upon people's values, preferences, expectations, constraints, and beliefs. Prior studies have examined how trust is attributed according to the ability of the system [59, 318], but to the best of our knowledge the link between integrity of systems and explanations has not been explored so far. Understanding the integrity dimension enables designing these agents to align their behaviour to the values of the humans they interact with. Note, that different people prioritize different values, which in turn guides how people behave and judge the behaviour of others [121].

This alignment of values is pivotal in domains where AI agents are entrusted with significant decision-making responsibilities. In cases such as autonomous vehicles, healthcare recommendations, and legal advisory systems, the stakes are high, and the level of trust required is substantial. The assurance that AI systems operate in harmony with ethical

values and individual values helps assure users that their interests and ethical principles are safeguarded. It also strengthens trust in AI systems by reinforcing the belief that these systems prioritize their individual user's interests and adhere to ethical principles. This understanding is a central aspect of the social lens because it deals with the human experience, perceptions, and the ethical implications of trust in AI systems. Therefore, we explore the following research question (Chapter 4):

**RQ3: How does human and AI agent value similarity influences a human's trust in that AI agent?**

***Integrity of AI Systems & Explanations*** Integrity is a fundamental ethical principle that encompasses honesty, transparency, and consistency in one's actions. In the realm of AI systems, integrity can be viewed as a value that AI systems should uphold. When AI systems demonstrate integrity, they not only follow ethical guidelines but also exhibit a sense of moral responsibility, which is essential for fostering trust. Users tend to trust AI systems that display integrity because they perceive these systems as fair and principled.

By studying RQ2, we found that integrity holds a significant place in the matrix of trust [254]. In this social lens, understanding integrity is paramount, as it is one of the core factors that influence the perceived trustworthiness of AI systems. Thus, integrity critically contributes to the development of appropriate trust in AI systems, emphasizing its importance in the complex landscape of human-AI interaction.

One of the methods for AI systems to display integrity is the use of Explainable AI. Explainable AI (XAI) is meant to give insight into the AI's internal model and decision-making [395] and has been shown to help users understand how the system works [52, 287]. Efforts to ensure that AI is trusted appropriately are often in the form of explanations [24, 225, 414]. Furthermore, integrity has linked it to conventional standards of morality - especially those of honesty and fairness [159, 246]. XAI can be regarded as a way to enhance system integrity i.e., the system being honest about making decisions is a form of integrity. Therefore, the question arises what the effect would be of explicitly mentioning principles related to integrity into XAI on appropriate trust of a user in the system.

**RQ4: How does the expression of different principles of integrity through explanations affect the appropriateness of human's trust in the AI agent?**

#### **APPLICATION LENS: BUILDING APPROPRIATE TRUST IN AI-BASED PREDICTIVE POLICING**

Transparency through explanations can help to achieve appropriate integrity and thus trust [252]. However, it is not clear yet when to choose what forms of explanations [353]. For example, explanations could be visual in the form of a graph such as a saliency map [4], or textual in the form of words and phrases or analytical, allowing users to explore the data and the model [142, 194]. Despite the considerable interest different XAI presentation methods have received individually, only a few studies have compared these different presentation forms to learn when, why, and for whom they work [277, 295, 313, 353]. To address this research and empirical gap, in RQ5, we investigate how users interact and

perceive different explanations, such as textual, visual, or a combination of text and visual (i.e., hybrid) to foster appropriate trust in an AI-based predictive policing system<sup>1</sup>.

**RQ5: What effect do different types of explanations have on building appropriate trust in AI-based predictive policing systems?**

We choose AI-based predictive policing as our use case because it represents a domain where appropriate trust is pivotal due to the high-stakes nature of decision-making. Furthermore, examining how different explanations impact appropriate trust in this context is practical as it informs the design and deployment of AI systems particularly in sensitive areas like law enforcement, promoting responsible and effective use [338].

As per Ribera et al. [309], the effectiveness of explanations is not solely contingent on the delivery of explanations but is also influenced by the specific end-user who receives these explanations. For instance, developers and AI experts may utilize explanations to validate the system's proper functioning. Previous research indicates that a hybrid form of explanations significantly enhances the understanding of lay users compared to visual explanations [33]. However, there has been limited exploration into the comparative effectiveness of various explanation methods in instilling appropriate trust, particularly between expert users and lay users. In the context of predictive policing, this comparison gains relevance, given that some police officers may have varying levels of professional experience, such as those who recently joined the department. Consequently, this chapter also aims to assess the efficacy of different explanation types with both expert and lay users, considering professional experience as a moderating factor. Thus, we prompted users with a selection of hotspots related to predictive policing, gauging their comprehension of the provided explanation and assessing whether they could place appropriate trust in the system.

To quantify appropriate trust, we employed established measures from existing literature. The sub-question was approached by enlisting and contrasting participants with varying expertise, such as police officers and lay users, with a focus on examining the role of explanations in cultivating both appropriate and subjective trust.

Overall, this chapter serves as the application lens in the thesis, combining insights from both the social and formal lenses to address the main research question. From the formal lens perspective, this chapter engages with the structured methodology of employing various types of formal definitions of appropriate trust as studied in RQ3 and established measures from the literature to systematically investigate the impact on appropriate trust in AI-based predictive policing systems. Simultaneously, the application lens draws on social lens by incorporating the use of different explanation presentation forms (no explanation, textual, visual, or hybrid) by introducing a systematic approach to understanding the role of explanations in the context of building trust. Furthermore, the inclusion of participants with different expertise levels, such as police officers and lay users, reflects a social lens, acknowledging the diversity in user backgrounds and experiences. By doing so, it provides a nuanced application lens, offering insights into the practical implications of different

<sup>1</sup>In this RQ, we only focus on presentation form of explanations which is distinct from different type of explanations such as counterfactual, importance-based or factor-based [410]

explanation methods on the establishment of appropriate trust in AI-based predictive policing systems.

## 1.4 SCOPING THIS THESIS

As building appropriate trust in human – AI interaction is a complex topic, it is not possible in a single thesis to fully address how to design for it. Hence, in Figure 1.2 we present how we scoped down our work, in terms of the problem targeted and adopted methodological focus. Table 1.1 summarizes the scoping of our research.

Table 1.1: Research Questions and Our Methodological Focus

| Research Questions   | Our Methodological Focus  |
|--|---|
| <b>RQ1.</b> How can we formally define appropriate human trust in human–AI interactions?   | A formal conceptualization of appropriate trust based on the concept of nested beliefs.   |
| <b>RQ2.</b> What’s state-of-the-art in fostering appropriate trust in AI systems?  | A systematic literature review of Measurement, Tasks, Methods, and results of those methods of building appropriate trust in human-AI interaction.  |
| <b>RQ3.</b> How does human and AI agent value similarity influence a human’s trust in that AI agent?   | Creation of AI agents value profiles based on Schwartz Portrait Value Questionnaire [24] and teaming participants with those AI agents to examine value similarity and effect on human trust. |
| <b>RQ4.</b> How does the expression of different principles of integrity through explanations affect the appropriateness of human’s trust in the AI agent? | Design of integrity-laden explanations focusing on three principles of integrity: fairness, honesty, transparency and examining the effect of them on appropriate trust.                      |
| <b>RQ5.</b> What effect do different types of explanations have on building appropriate trust in AI-based predictive policing systems?                     | Design of text-based, visual and hybrid (Text+Visual) explanations to examine the effect of them on appropriate trust in an AI-based predictive policing system.                              |

## 1.5 CONTRIBUTIONS

The posed research questions are explored in depth in the papers presented in Part II of this thesis. This section briefly summarizes the contributions in the Table 1.2.

Table 1.2: Contributions per chapter in this thesis

| Chapter | Adopted Lens | Contribution(s)   |
|---------|--------------|---|
| 2       | Formal       | A formal conceptualization of appropriate trust based on nested beliefs and how trust beliefs are established in a dyad between a human and agent.  |
| 3       | Formal       | 1) A systematic review of current practices in building appropriate trust, different ways to measure it, types of tasks used, and potential challenges associated with it.<br>2) A novel Belief, Intentions, and Actions (BIA) mapping to study commonalities and differences in the concepts related to appropriate trust by (a) describing the existing disagreements on defining appropriate trust, and (b) providing an overview of the concepts and definitions related to appropriate trust in AI from the existing literature. |
| 4       | Social       | An understanding of how human and agent Value Similarity (VS) influences a human's trust in that agent.   |
| 5       | Social       | 1) A measurement method for appropriate trust based on a specific task in human-AI interaction.<br>2) An understanding of how expressing integrity through explanations can help in building appropriate trust in AI systems.   |
| 6       | Application  | An understanding of how user expertise and different types of explanations affect user's appropriate and subjective trust in predictive policing.   |

## 1.6 THESIS ORGANIZATION

The remainder of this thesis consists of the chapters listed below and schematized in Figure 1.2 with the underlying concepts and relationships.

**Chapter 2:** Building Appropriate Trust in AI systems – A Formal Perspective (*A combined chapter of published extended abstract [114], doctoral consortium [251] and a co-authored journal paper [369]*)

**Chapter 3:** Systematic Literature Review – Building Appropriate Trust in AI Systems (*to appear in ACM Journal of Responsible Computing*)

**Chapter 4:** Effect of Value Similarity on Trust in Human-Agent Interaction (*Published in AAAI/ACM AIES Conference 2021 - [254]*)

**Chapter 5:** Integrity Based Explanations for Fostering Appropriate Trust in AI Agents (*Published in ACM Transactions on Interactive Intelligence 2023 - [252]*)

**Chapter 6:** Fostering Appropriate Trust in Predictive Policing AI Systems (*In review at ACM AIES conference 2024*)

**Chapter 7:** Discussion & Conclusion

This thesis is aimed at understanding how we can design for appropriate trust in human-AI interaction. We begin with understanding the topic of appropriate trust in Chapter 1 and adopt a three lenses approach: Formal, Social and Application to study appropriate trust. We first adopted the formal lens, where we funnelled down 1.2 to conceptualizing appropriate trust through nested beliefs and dyadic trust, as our specific focus was on human-AI interaction (Chapter 2). After conceptualizing appropriate trust and its definition, we broadened our focus by studying how other researchers from different disciplines have defined it and what methods and measures have been adopted to achieve appropriate trust in human-AI interaction (Chapter 3).

Our findings from Chapter 3 provided insights to funnel down our approach of the use of Explainable AI (XAI) to build appropriate trust. To further understand the use of XAI, we adopted a social lens to study how integrity-laden explanations can help build trust (Chapter 4) and then appropriate trust in human-AI interaction (Chapter 5). We specifically looked at integrity-laden explanations as they appeared to be a research gap in exploring their effectiveness in fostering appropriate trust (Chapter 3), and they also provided us with a way to design ethical AI agents that possess internal integrity.

Finally, in Chapter 6, utilizing the application lens, we investigated the effect of integrity-laden explanations in a specific domain (predictive policing) using an iterative and incremental design cycle informed by the previous Chapters 4 and 5.

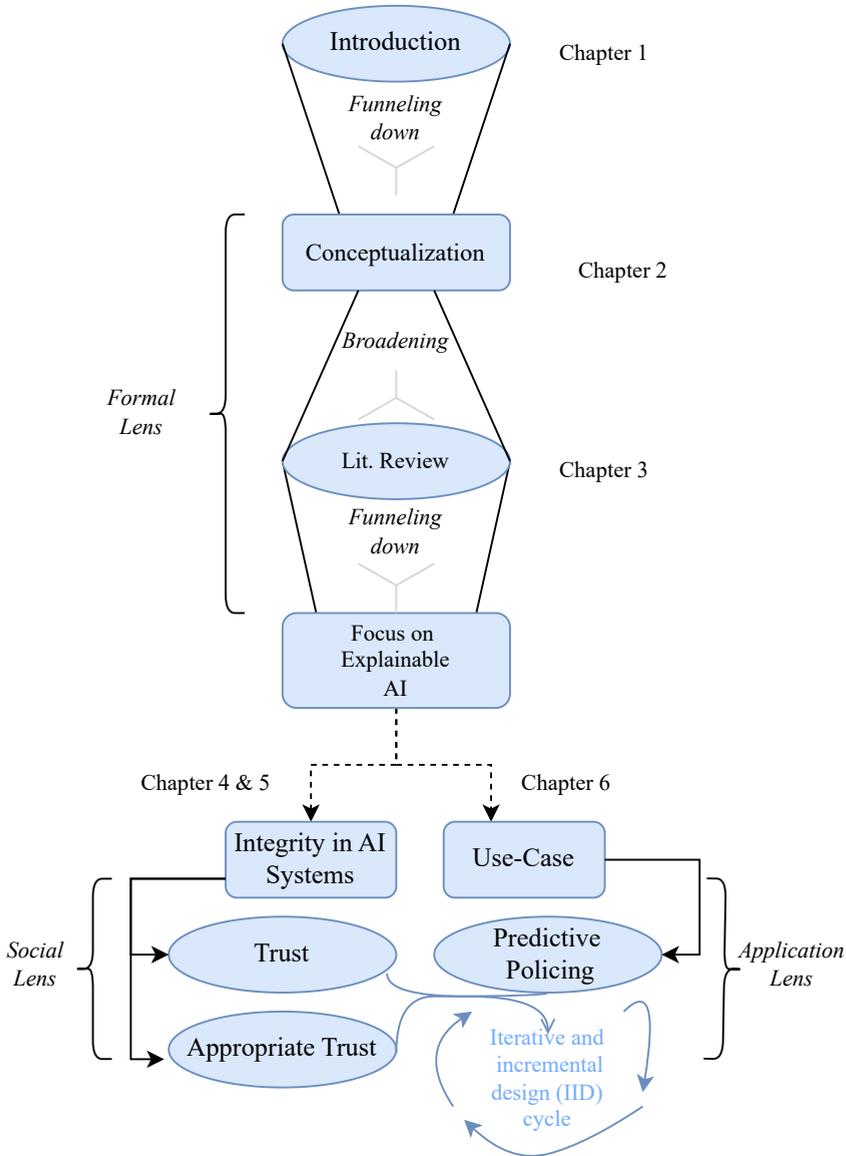


Figure 1.2: Structure of the thesis – outline of each chapter presented along with its main concept.



# I

## FORMAL LENS



## 2

# BUILDING APPROPRIATE TRUST IN AI SYSTEMS – A FORMAL PERSPECTIVE

---

[>] This chapter comprises of the following published articles where I have contributed specific sections.

[📄] **Siddharth Mehrotra**, Modelling Trust in Human-AI Interaction: Doctoral Consortium Track. in *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '21*, Online, IFAAMAS.

[📄] Carolina Centeio Jorge, **Siddharth Mehrotra**, Myrthe L. Tielman, and Catholijn M. Jonker. "Trust should correspond to trustworthiness: a formalization of appropriate mutual trust in human-agent teams." in *Proceedings of the 22nd International Workshop on Trust in Agent Societies (TRUST 2021)* co-located with the *20th International Conferences on Autonomous Agents and Multiagent Systems (AAMAS 2021)*.

[📄] Anna-Sophie Ulfert, Eleni Georganta, Carolina Centeio Jorge, **Siddharth Mehrotra**, and Myrthe Tielman. Shaping a multidisciplinary understanding of team trust in human-AI teams: a theoretical framework in *European Journal of Work and Organizational Psychology (EJWOP)* (2023): 1-14.

## 2.1 INTRODUCTION

Artificial Intelligence (AI) agents are becoming more intelligent and able to execute relevant tasks for our daily lives. For some of these tasks, humans and AI agents should learn to cooperate, coordinate and collaborate for optimal outcomes. The idea is to capitalize on the strengths of both humans and AI agents. In the context of collaboration between humans and AI, an optimal outcome refers to the achievement of the best possible result or performance [52]. This could involve a combination of factors such as increased efficiency, enhanced productivity, improved decision-making, and successful resolution of tasks or challenges.

A key element in achieving effective decision-making is mutual appropriate trust [182], i.e., both trustor and trustee should trust each other appropriately. We take appropriate to mean that a human's trust in an agent should correspond to that agent's trustworthiness, and an agent's trust in a human should correspond to the human's trustworthiness. On an intuitive level, we argue that appropriate trust happens when an entity's trust towards another entity corresponds to the latter's actual trustworthiness. Consequently, when there is appropriate trust, there is no under-trust (leading to under reliance) or over-trust (leading to over-reliance), which minimizes negative performance outcomes. For example, if an agent X trusts another agent Y to execute a task (e.g., driving a car) which requires skills that Y does not have, agent X over-trusted agent Y and the consequences can be negative and even disastrous (e.g., car accident). On the other hand, if agent X does not trust agent Y to execute a task (e.g., driving a car) and agent Y would execute the task perfectly well, agent X is under-trusting agent Y which can also negatively affect the outcome (e.g., walking instead). In particular, when X is a human and Y is an AI agent, and trust is not appropriate, this will lead to disuse or misuse of technology [294]. However, such intuitive notions leave ambiguities. Therefore, we propose to formally define what appropriate trust means in human-AI interaction.

In the design and development of AI agents, formalizing the concept of appropriate trust is paramount [400]. Formalism serves as a key foundation, providing essential clarity and precision in the creation of AI systems [79], and the use of formalized theory to study complex and dynamic systems [359]. Furthermore, formalism is crucial for guiding the development process, allowing for the creation of measurable metrics and evaluation criteria [199]. This, in turn, facilitates the assessment of AI systems in fostering mutual appropriate trust in human-AI interactions. Finally, a formal definition contributes to the theoretical underpinning of the concept, aiding in the creation of models and frameworks that enhance the reliability of AI systems. Therefore, our main research question is:

How can we formally define appropriate human trust in human-AI interactions?

To achieve a formal definition of appropriate trust in human-AI interaction, we first need to understand trust itself and how it forms and dissipates in human-AI dyads. This understanding can be gained by a nuanced examination of how both trustor and trustee perceive each other's competence, benevolence, and ethical considerations [47]. By examining aforesaid considerations, we can gain insights into the multifaceted factors influencing trust formation, including performance, transparency, and alignment with human values (Section 4). Additionally, in human-AI interaction insights from dyadic trust can inform

the development of adaptive models and algorithms, enabling AI systems to respond to real-time feedback, align with user preferences, and uphold transparent decision-making processes.

After formally defining trust based on human and AI agent beliefs (Section 2.2), we will establish a formal notation for appropriate trust based on nested beliefs (Section 2.3). Finally, to establish the grounding of dyadic trust we will conceptualize how trust beliefs (i.e., belief in trustworthiness) can be established in human-AI interaction (Section 2.4) based on the Ability, Benevolence and Integrity (ABI) model [242].

## 2.2 TRUST AS A BELIEF OF TRUSTWORTHINESS

Humans base their daily behaviour on trust; every time we interact with, delegate to or rely on the intention of another individual, group or thing [371]. In an interaction between two cognitive agents [369] (artificial or human), trust involves two parties, the trustor and the trustee, and an action (trusted by the trustor to the trustee) that affects a goal (of the trustor). For this discussion, it is important to acknowledge that trust and trustworthiness are two related, but distinct concepts. While trustworthiness, the characteristic that someone is to be trusted can be conceptualized as a property of the trustee (e.g. following the work of Castelfranchi and Falcone [54]), trust is a directional attitude of the trustor towards a trustee, which involves the perceived trustworthiness of this trustee. This implies that the trustor must have a “theory of the mind” of the trustee, which may include personality, shared values, morality, goodwill, etc. [56].

Trust is an aspect of relationships and, as such, can only be viewed in the context of individuals and their relationships [330]. As an example, let’s imagine that a cognitive agent Y (artificial or human) drives well and is trustworthy regarding driving tasks. For another cognitive agent X to trust agent Y for a driving task, agent X has to believe that agent Y is trustworthy for this task. The assertion in the example implies a nuanced relationship, recognizing the interplay between objective trustworthiness and subjective trust beliefs. This corresponds to the concept that any changeable notion that an agent has about the world is a belief that agent has. In this, we follow the Beliefs concept from the Belief-Desire-Intention (BDI) architecture for agents [306]. Formalising this, in terms of the trust  $T$  of agent  $x$  in agent  $y$ , is a belief of  $x$  (trustor),  $B_x$ , about  $y$ ’s (trustee’s) trustworthiness,  $TW_y$ , meaning that:

$$T(x, y) = B_x(TW_y) \quad (1)$$

Accordingly, in order to understand trust, we first need to understand trustworthiness, and secondly how beliefs about trustworthiness are formed. Trustworthiness is a complex concept, and following the literature it can consist of a set of dimensions that range from the trustee’s competence to its intentions [133]. How an entity can be considered trustworthy is not a trivial question, and is context-dependent, as well as dependent on the nature of the trustee [146, 325]. When considering human trustworthiness in organizational behaviour, *Ability, Benevolence and Integrity (ABI)* model [241] is often employed. Similarly, other dimensions of trust (perceived trustworthiness) in technology as being *Performance, Process and Purpose* [214], which are linked with the ABI model according to Lee & See [216]. In this work, we use the ABI model from Lee & See’s work [216].

When talking of artificial agents and societies, for example, we can consider beliefs such as *Willingness*, *Competence* and *Dependence* to estimate the trustworthiness of another cognitive agent [57]. Finally, trustworthiness or its dimensions can be affected by external factors, which are contextual conditions determining the situation in which the task is executed [107], such as environmental configuration, emotional state, workload, etc.

As mentioned previously in this section, trust varies with the person and across relationships. We will illustrate this using the *ABI* model of trust [241], which has been widely used to study human trust. The authors define trust as “*the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party*”. In this model of trust, trustworthiness is defined as “the extent to which an actor has the ability to execute relevant tasks, demonstrates integrity, and is benevolent towards fellow team members” [382]. Furthermore, Lee & See [216] define *Ability*, *Benevolence* and *Integrity* as follows:

1. **Ability:** “Ability is that group of skills, competencies, and characteristics that enable a party to have influence within some specific domain”.
2. **Benevolence:** “Benevolence is the extent to which the intents and motivations of the trustee are aligned with those of the trustor. Benevolence suggests that the trustee has some attachment to the trustor”.
3. **Integrity:** “Integrity is the degree to which the trustee adheres to a set of principles the trustor finds acceptable”.

We can see that although Ability depends only on the trustee, both Benevolence and Integrity depend on both the trustor and the trustee. Even though trustworthiness is a characteristic of a trustee, this characteristic will differ per trustor. For example, if Alice trusts Bob to babysit her children, her trust is based on her belief in Bob’s ability to care for the children (his ability), his goodwill towards her and her children (his benevolence), and his honesty about his qualifications and intentions (his integrity). However, Charlie, who has had a negative experience with Bob in the past, might not trust Bob to babysit his children, even if Alice does. This is because Charlie’s perception of Bob’s trustworthiness is different from Alice’s, potentially stemming from situations like Bob questioning Charlie’s honesty, compromising integrity in Charlie’s eyes. Additionally, differences in shared values, like preferred sleep schedules (8 pm for Alice and Bob, 6 pm for Charlie), might create a clash of styles, further influencing Charlie’s perception of Bob’s suitability and ultimately, his trust. These examples highlight how a trustee’s trustworthiness is not a fixed characteristic but rather a dynamic concept shaped by the individual’s past experiences, shared values, and perceptions.

Following, both trust and trustworthiness depend on the two cognitive agents (artificial or human), trustor and trustee ( $x$  and  $y$ ), that compose the dyadic relationship. Thus, we stipulate that we need to define the trust of an agent  $x$  in agent  $y$  as the belief  $\mathcal{B}$  of agent  $x$  regarding the trustworthiness of  $y$  with respect to  $x$ , adapting Expression 1 to:

$$T(x, y) = \mathcal{B}_x(\mathcal{T}\mathcal{W}_y(x)) \quad (2)$$

where  $\mathcal{T}\mathcal{W}_y(x)$  means trustworthiness of  $y$  with respect to  $x$ .

## 2.3 APPROPRIATE TRUST

As mentioned before, it is important that human's trust in the AI agent is appropriate. The challenge is to make the AI agent behave so that humans tune their trust towards the AI agent, since we do not have control over the human. However, leveraging the idea that agents reflect about their own trustworthiness, we may be able to build agents which actively influence humans to appropriately fine-tune their trust in them. With the information regarding the agent's trustworthiness, human teammates can adapt to the qualities and limitations of the agent and, consequently, adjust the utilization of the agent accordingly. Without this knowledge, however, it would prove difficult to coordinate within a task-environment given an unpredictable agent teammate.

We posit that how trustworthy an agent,  $a$ , is for a human,  $h$ , and how a human trusts the agent (human's belief in agent's trustworthiness) should be similar to get appropriate trust. If the belief of an agent in their own trustworthiness towards human is different from their belief of human's trustworthiness towards them then we come closer to under-trust  $T(a, h) \downarrow$  or over-trust  $T(a, h) \uparrow$  *i.e.*

$$\mathcal{B}_h(\mathcal{T}\mathcal{W}_a(h)) > \mathcal{T}\mathcal{W}_a(h) \rightarrow T(a, h) \uparrow \quad (3)$$

$$\mathcal{B}_h(\mathcal{T}\mathcal{W}_a(h)) < \mathcal{T}\mathcal{W}_a(h) \rightarrow T(a, h) \downarrow \quad (4)$$

Therefore, to avoid such situations, the agent's belief in their own trustworthiness should match with their belief about the belief of human's trust in them, hopefully (assuming the beliefs are correct) leading to appropriate trust by a human in that agent. However, shifting our focus to human-AI interaction, we encounter a more complex scenario. In this setting, not only is it crucial for the agent to appropriately trust the human (1) and for the human to trust the agent (2), but also the agent has a belief about whether the human appropriately trusts the agent, and that (3) the human believes the agent appropriately trusts the human, and why. In this chapter, we will focus on the cases 1 and 3, since we are addressing trust from the agent's perspective as we can only directly manipulate the artificial agent's beliefs. Fig.2.1 schematizes a dyadic human-agent relationship.

Following figure 2.1, we have the trust of the AI agent in the human, meaning the agent's belief on human's trustworthiness,  $T(a, h) = \mathcal{B}_a(\mathcal{T}\mathcal{W}_h(a))$  (from Expression 2), and the trust of the human in the agent, which is the human's belief on agent's trustworthiness  $T(h, a) = \mathcal{B}_h(\mathcal{T}\mathcal{W}_a(h))$ .

An agent, to be able to promote and elicit appropriate trust (from the human towards the agent), does not only need to reason with beliefs about human's trustworthiness, but also with beliefs about human's trust (estimating whether the human trusts the agent). What's more, we identify that we may also need beliefs about trust when appropriately estimating human's trustworthiness. This being said, in the dual-mode vehicle example, can the agent trust the human to follow an instruction, if the human does not trust that agent? Considering again the *ABI* trustworthiness model mentioned in the previous section, we believe that if a trustee trusts a trustor, this is a sign of their benevolence towards the trustor, which in turn would increase the trustee's trustworthiness to that trustor. Thus, in order to trust the human teammate, the agent should estimate the human's trust in the agent. In order to estimate  $\mathcal{B}_a(\mathcal{T}\mathcal{W}_h(a))$ , we may also need the agent's belief in human's trust in the agent, *i.e.*,  $\mathcal{B}_a(\mathcal{B}_h(\mathcal{T}\mathcal{W}_a(h)))$ . Following the example, for the agent to trust the

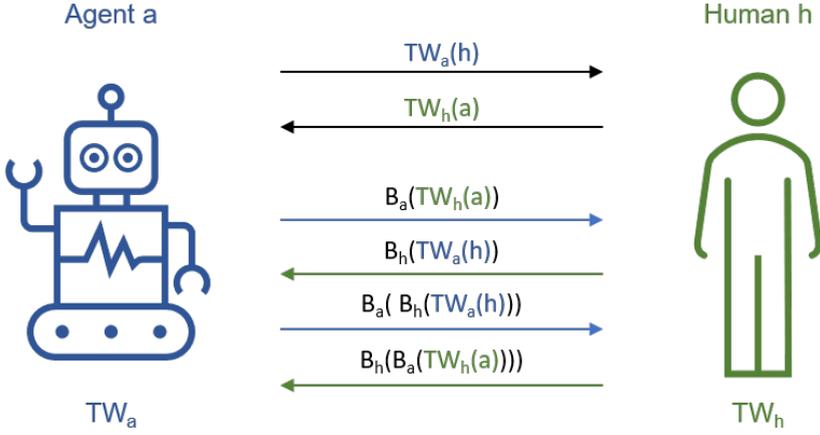


Figure 2.1: Human-Agent dyadic trust based on the beliefs of trustworthiness. Here  $\mathcal{T}\mathcal{W}_a(h)$  means trustworthiness of agent  $a$  with respect to human  $h$ .

human to follow an instruction, the agent needs to believe that the human trusts the agent (e.g., the human relies on this particular agent's knowledge/intelligence).

As mentioned before it is important for the agent to appropriately trust the human. So, when estimating whether it can trust its human teammate to follow an instruction, the agent's trust in the human should correspond to the actual human's trustworthiness (e.g., to what actually the human can and/or wants to do), *i.e.*,

$$T(a, h) \equiv B_a(\mathcal{T}\mathcal{W}_h(a)) \equiv \mathcal{T}\mathcal{W}_h(a) \quad (5)$$

which requires that the agent also accurately estimates the human's trust in the agent. The *human's* trust in the agent, on the other hand, is the belief of the human in the agent's trustworthiness,  $B_h(\mathcal{T}\mathcal{W}_a(h))$ , and should correspond to the agent's actual trustworthiness ( $\mathcal{T}\mathcal{W}_a(h)$ ), *i.e.*,

$$T(h, a) \equiv B_a(B_h(\mathcal{T}\mathcal{W}_a(h))) \equiv B_h(\mathcal{T}\mathcal{W}_a(h)) \equiv \mathcal{T}\mathcal{W}_a(h) \quad (6)$$

Lastly, since the nested concepts presented on Expression 5 are based on  $\mathcal{T}\mathcal{W}_a(h)$ , this means that we may be able to calibrate human's trust in the agent ( $T(h, a)$ ), by manipulating  $\mathcal{T}\mathcal{W}_a(h)$  through the accurate belief of the agent's own trustworthiness. This means that if the agent is aware of its own trustworthiness, meaning that if the agent's belief in agent's trustworthiness corresponds to actual agent's trustworthiness, *i.e.*,

$$B_a(\mathcal{T}\mathcal{W}_a(h)) \equiv \mathcal{T}\mathcal{W}_a(h) \quad (7)$$

the agent may be able to alter its own trustworthiness (or simply how it lets the human perceive it).

## 2.4 HOW ARE TRUST BELIEFS ESTABLISHED?

In previous sections, we formalized that if the agent is aware of its own trustworthiness, meaning that if the agent's belief in their trustworthiness corresponds to their actual trustworthiness, the agent may be able to change the perception of the human and, consequently can help in calibrating human's trust. However, how these beliefs can be established still remains an open question.

To study how beliefs can be established, we will again utilize the ABI model proposed by Mayer et al [241]. A common outline in studying trustworthiness of AI agents is often linked with measuring the performance of the agent based on the capabilities [256]. Following this, much of the previous research has looked at 'ability' as the core factor of estimating trust. However, we propose to view trustworthiness as more than ability.

We identify a gap in the literature which focuses upon modelling integrity and benevolence of an artificial agent towards a human [372]. We argue that this research gap is pivotal in our understanding of trust beliefs by the agents which focuses on aspects other than their capabilities. Therefore, in the following sub-section, we will refer to our readers prior work on ability and then we propose a first attempt on how integrity can be modelled. Finally, we will discuss an existing model for computational benevolence [372].

### 2.4.1 ABILITY

We can infer an agent's ability from the aggregation of all functionalities and capabilities it has. Understanding the agent's capabilities embedded by the developer can help both the agent and the human understand whether an agent has very low ability or very high ability. For this purpose, several existing trust-related functionality aggregators which focus on the system capabilities may be used, such as the ones described in [36, 62, 370]. In these functionality aggregators the trust is based on the task success, number of errors, skills and knowledge to accomplish a task. For example, competence beliefs about an agent [57], automation capability [190] and agent's core functionality [165] are prominent inclusions for functionality aggregators. Therefore, in this paper we propose to rely on such existing aggregators to understand an agent's ability in terms of how it can form belief about its own trustworthiness.

### 2.4.2 INTEGRITY

We assume that an agent's integrity for a specific task is its integrity towards a human in accomplishing that specific task. We define integrity as the similarity of the human and agent values<sup>1</sup>, meaning, having similar priorities over those values (which can be related to the actual definition of integrity focusing upon principles). In this section, we propose a basis for formally defining how integrity beliefs are formed. In particular, we derive two cases where an agent either possesses information regarding the values of the other or does not.

**Case 1** We start by defining this relationship in the case where an agent has some belief about the values of the human. We state that the belief an agent  $a$  has about the

<sup>1</sup>Values are abstract motivations that justify opinions and actions, and are intrinsically linked to moral judgement [331].

integrity of a human  $h$  directly follows from the how similar agent  $a$  believes the priority ranking of their values to be to that of  $h$ :

$$\text{if } \{\mathcal{B}_a^{\text{sim}}(\mathcal{B}_a(\text{priority}_h(V_h)), \text{priority}_a(V_a))\} \uparrow \text{ then } \{\mathcal{B}_a(I(h))\} \uparrow \quad (8)$$

$$\text{if } \{\mathcal{B}_a^{\text{sim}}(\mathcal{B}_a(\text{priority}_h(V_h)), \text{priority}_a(V_a))\} \downarrow \text{ then } \{\mathcal{B}_a(I(h))\} \downarrow \quad (9)$$

where  $\mathcal{B}_a^{\text{sim}}(a, h)$  represents the belief of  $a$  about the similarity of  $a$  and  $h$ ,  $V_a$  is the value set of  $a$ ,  $\text{priority}_a(V)$  represents a priority ranking of agent  $a$  over this value set, and therefore  $\mathcal{B}_a(\text{priority}_h(V_h))$  is  $a$ 's belief about  $h$ 's value and priority thereof. In other words, integrity beliefs of  $a$  about  $h$  are formed by  $a$  comparing their belief about  $h$  with what they know about themselves. We stipulate that if the belief about the similarity is higher, so will the belief in the integrity and *vice-versa*.

**Case 2:** In general it is important for an agent to rely on knowledge about human values. However, a situation could arise in which an agent has no information regarding a certain human teammate. In such a case we focus on the integrity *reputation* of the agent as per [419].

We state that the integrity reputation of an cognitive agent (artificial or human)  $a_0$  according to an artificial agent  $a$ ,  $IR_a(a_0)$ , is the average sum of beliefs about  $a_0$ 's integrity that are communicated ( $CB_a$ ) by other autonomous agents ( $a', a'', \dots$ ) to  $a$ . The equation 12 represents the communicated beliefs of  $a_0$  according to  $a$  as sum of the communicated beliefs by other autonomous agents. Finally, the equation 13 represents the average sum of those beliefs forming the integrity reputation of  $a_0$  according to  $a$ .

$$CB_a(a_0, \{a', a'' \dots\}) = \sum_{i=a'}^{a^n} \mathcal{B}_{a'}(I_{a'}(a_0)), \mathcal{B}_{a''}(I_{a''}(a_0)) \dots \mathcal{B}_{a^n}(I_{a^n}(a_0)) \quad (10)$$

$$IR_a(a_0) = \frac{CB_a}{n} \quad (11)$$

### 2.4.3 BENEVOLENCE

Mayer et al. proposed that the effect of integrity on trust will be most salient early in the relationship prior to the development of meaningful benevolence [241]. Therefore, we believe it is firstly important to understand how integrity can be understood and modeled as effect of perceived benevolence on trust will increase over time as the relationship between the parties develops. Moreover, as benevolence is about interpersonal relationships, it might not develop in agent-human relationships in the way it does for human-human ones. There are a number of steps taken in the social science research community to understand benevolence [171, 213]. However, we could only find one example of modelling benevolence in computer science community by Urbano et al. who classify benevolence as a Social Tuner in Human-AI interaction [372]. According to them, Social Tuner measures the trustee's specific attachment toward the truster. This attachment is captured by the coefficient of benevolent actions. They estimate the value of the benevolence of the trustee toward the truster,  $ben_{x,y}$  from the coefficient of benevolent actions. The coefficient of benevolent actions  $\rho_{ba} \in [0, 1]$  measures the historical pattern of favorable or beneficial

actions taken by the trustee towards the truster.

$$ben_{x,y} = \frac{1}{2}\rho_{ba} + \frac{1}{2} \frac{cumValAgreem}{n} \quad (12)$$

where  $\rho_{ba}$  is the result of the correlation between the number of agreements established between truster and trustee in the past and cumulative value of past agreements (*cumValAgreem*).

It is worth noting that the estimation of benevolence is only possible when there are, at least, two past interactions between the truster and the trustee under evaluation. The manner in which [371] studies benevolence can also fit in our formalism focusing upon beliefs which forms specific relationship between trustee and trustor. Also, the value of benevolence must be updated at every new trustworthiness estimation, as benevolence may evolve due to the mutual (dis)satisfaction of the trustee with the relationship, which may change with time and context.

## 2.5 SUMMARY OF CONTRIBUTIONS

This chapter set out to address the RQ: How can we formally define appropriate human trust in human – AI interactions? The initial phase of our exploration involved studying trust as a belief of trustworthiness. This phase on inquiry mentioned that to understand trust, we first need to understand trustworthiness, and secondly how beliefs of trustworthiness are formed. By synthesizing theoretical constructs, we adopted the ABI model [242] to define the trust of an agent x in agent y as the belief B of agent x regarding the trustworthiness of y with respect to x. Furthermore, we proposed that for appropriate trust to develop, an agent's perception of its trustworthiness towards a human and the human's actual trust in the agent should be aligned. When these beliefs diverge, it can lead to either under-trust or over-trust situations, refer Equation 3 & 4.

Based on the theoretical constructs in Section 1.2 and 1.3, we formalized that in human-AI interaction it is important that 1) the agent appropriately trusts the human and 2) the human appropriately trusts the agent, but it is also important that 3) the agent has a belief about whether the human appropriately trusts the agent, and that 4) the human believes the agent appropriately trusts the human, and why [381]. We schematized these notions through formalisms of nested beliefs and proposed to view them in a human-AI agent dyadic relationship. In Section 1.4, we studied the salient role of mutual trust in a dyadic relationship, including belief formations shaping human trust. Subsequently, our focus shifted to establishing these beliefs, wherein we sought to define an AI agent's integrity and benevolence in Section 1.5. By establishing a theoretical foundation concerning the ethical principles and their congruence with human values, this exploration laid the essential groundwork for understanding appropriate trust in AI systems.

## 2.6 LIMITATIONS

To appreciate the contribution of this work, we also need to understand its limitations. In our definitions for trustworthiness and trust in agents or humans, we rely on the belief of one in another. However, we explicitly do not focus on how those beliefs are generated and how we can evaluate them. Although these are crucial for achieving mutual appropriate trust, we argue that we can only start to generate and evaluate beliefs once we understand

what they are about. This work takes this first crucial step. Moreover, we note that according to many models, including the trust model by Rahman & Hailes [3], trust beliefs will involve beliefs of risk, utility and motivations along with trustworthiness. This means that a study into how beliefs about trustworthiness are formed should also take these aspects into account.

Nilsson states it is important to expose beliefs to the reasoned criticism of others [275]. Yet, we do not explicitly incorporate other researchers' criticism for the formation of beliefs regarding trust or trustworthiness. However, we believe this can be formally extended in future work when, for example, agent  $a_1$  is forming the belief regarding human  $h_1$ 's trust in itself,  $a_1$  can take into account another agent's criticism (if this other agent has interacted with  $h_1$  before).

Going forward, we aim to refine the formalization and implement a method to evaluate it. In particular, we would like to conduct user studies to evaluate our notions regarding beliefs of beliefs in an experimental setting. This can both help us understand how trust beliefs are formed in humans, and how agents can appropriately use these beliefs.

## 2.7 FOLLOW-UP DISCUSSIONS AND THESIS POSITIONING

This chapter laid the groundwork for defining and understanding appropriate trust in human-AI interactions with a formal lens. We have proposed a formal definition based on the belief of trustworthiness and emphasized the importance of considering the dyadic nature of the relationship between a human and an AI agent.

While this chapter defines trust as a belief in trustworthiness, it emphasizes the limitations of a simple definition. We conceptualize trust based on the Lee & See [216] work which links trust to a belief of trustworthiness. Then, we argue that an agent should aim for appropriate trust, which means it should aim for a human's belief in their trustworthiness to be equal to their actual trustworthiness. Given the diversity in terminology related to 'appropriate trust', it would be relevant to compare this conceptualization of appropriate trust to existing definitions in the literature, and adjacent concepts such as calibrated trust and warranted trust [167].

Furthermore, following our definition of appropriate trust, we would need to compare trust directly with the trustworthiness. However, it is a complex process to compare beliefs of trustworthiness with actual trustworthiness because beliefs about trustworthiness are subjective and limited by future actions and hidden motives, making it a challenge to compare them with someone's actual trustworthiness. Therefore, we propose to first research how prior works attempted to investigate this complex process.

Following the ABI model [242] and our directional definition of trust (Section 2), benevolence and integrity depend on both trustor and trustee. Moreover, ability for one task isn't the same as ability for another, and one could also question if the same integrity values are relevant in all contexts. Hence, the need to study in what contexts appropriate trust has been investigated so far.

Finally, several researchers have made attempts to study how to achieve appropriate trust in human-AI interaction (See Chapter 3). In future work, it would be relevant to first study what other methods have been successfully used so far to achieve this goal by other researchers. In synopsis, the next proposed steps can be:

1. **validation and refinement** of the proposed definition through comparison with existing understandings,
2. **identification of knowledge gaps** where the focus on nested beliefs and the dyadic nature can offer new insights,
3. **building upon existing research** by utilizing established knowledge, and
4. **informing future research directions** by highlighting areas like specific contexts and robust measurement development that require further exploration.



## 3

## 3

# A SYSTEMATIC REVIEW ON FOSTERING APPROPRIATE TRUST IN HUMAN-AI INTERACTION: TRENDS, OPPORTUNITIES AND CHALLENGES

*Appropriate Trust in Artificial Intelligence (AI) systems has rapidly become an important area of focus for both researchers and practitioners. Various approaches have been used to achieve it, such as confidence scores, explanations, trustworthiness cues, or uncertainty communication. However, a comprehensive understanding of the field is lacking due to the diversity of perspectives arising from various backgrounds that influence it and the lack of a single definition for appropriate trust. To investigate this topic, this paper presents a systematic review to identify current practices in building appropriate trust, different ways to measure it, types of tasks used, and potential challenges associated with it. We also propose a Belief, Intentions, and Actions (BIA) mapping to study commonalities and differences in the concepts related to appropriate trust by (a) describing the existing disagreements on defining appropriate trust, and (b) providing an overview of the concepts and definitions related to appropriate trust in AI from the existing literature. Finally, the challenges identified in studying appropriate trust are discussed, and observations are summarized as current trends, potential gaps, and research opportunities for future work. Overall, the paper provides insights into the complex concept of appropriate trust in human-AI interaction and presents research opportunities to advance our understanding on this topic.*

### 3.1 INTRODUCTION

Artificial Intelligence (AI) has become an increasingly ubiquitous technology in recent years, with applications in a wide range of industries and areas of life. The ability of AI to process and analyze large amounts of data quickly and accurately makes it particularly valuable for domains with high-stake decision-making such as finance, healthcare, and transportation [266, 335]. While AI-embedded systems are powerful, they can still fail or behave unpredictably, leading to inappropriate trust, and introducing the corresponding risk of *misuse* and *disuse* [294].

Both *disuse* [238] and *misuse* [243] of AI-embedded systems by humans have led to severe issues, such as Amazon's AI recruiting tool being biased against women [83], a railroad accident in which the crew neglected speed constraints [347], and the use of facial recognition technology in law enforcement to target Black and Latino communities [179]. One of the major reasons of *disuse* and *misuse* of AI is people's over- or under-trust in it, or in other words, lack of appropriate trust in AI [282]. Appropriate trust is often linked to the alignment between the perceived and actual performance of the system [406]. We argue that human trust in the AI system must be appropriate because, with appropriate trust in AI, people may be simultaneously aware of the potential and the limitations of AI. This should lead to reducing the harms and negative consequences of *misuse* and *disuse* of AI [292].

People have long been aware of the importance of appropriate trust in interpersonal relationships [101]. Taking an example from the Indian scripture "*Bhagavad Gita*", dated 400 BCE [116], the deity *Krishna* advises that humans should be careful in trusting others and develop trust in degrees so that their trust is often appropriate [46]. Furthermore, he suggests by cultivating appropriate trust, humans gradually move forward in interpersonal relationships. This highlights for how long this concept has played a role and is vital and helpful in understanding how people can develop appropriate trust in interpersonal relationships and AI systems.

To achieve appropriate trust in AI systems, different approaches have been taken

such as use of confidence scores [24, 185, 230, 301, 416], explanations [205, 207, 344, 394], cues (alarms [64, 408], warning signals [279] or uncertainty communication [365]). Many studies aim to adjust the trust bestowed in a system to reflect the trustworthiness of said system [225, 328, 407, 416]. Despite these efforts, a comprehensive understanding of the field is currently lacking, and consensus on the definition of appropriate trust remains elusive. Different perspectives and varying definitions of trust, trustworthiness, and reliance contribute to this lack of clarity, as pointed out by Gille [128].

According to Jacovi et al.'s overview [167], there are numerous types of trust that need to be more precisely defined and differentiated. For example, the confusion between two similar, yet different concepts, appropriate trust and appropriate reliance, which often stems from a lack of clear understanding of these terms' definitions. Various strategies have been employed to establish an appropriate level of trust in human-AI interaction. Researchers from diverse scientific fields have conducted empirical studies and developed theoretical models to explore different methodologies for building such trust. However, despite the crucial role of appropriate trust in ensuring the successful use of AI systems, there is currently a fragmented overview of its understanding [256].

To highlight and better understand appropriate trust in human-AI interaction, our paper aims to present a comprehensive overview of the current state of research on Human-AI trust by emphasizing definitions, measures, and methods of fostering *appropriate trust* in AI systems. Furthermore, we make an attempt to map different terms associated with appropriate trust and provide a comprehensive summary of current trends, challenges and recommendations.

In this work, we study the state-of-the-art in building appropriate trust by examining its evolution, definitions, related concepts, measures, and methods. Our research questions are:

1. What's the history of appropriate trust in automation before AI systems?
2. How does current research define appropriate trust and what related concepts exist?
3. How can we structurally make sense of these concepts related to appropriate trust?
4. What's state-of-the-art in fostering appropriate trust in AI systems? *which includes*
  - (a) How do studies measure whether the trust is appropriate or not?
  - (b) What kind of tasks do researchers employ in their studies to understand appropriate trust?
  - (c) What different types of methods for building appropriate trust exist?
  - (d) What are the results of the methods aimed at building appropriate trust?

To investigate the questions above, we provide a history of appropriate trust development and present a systematic review to identify current practices in the theoretical and experimental approaches. Furthermore, we identify potential challenges and open questions, allowing us to draw research opportunities to understand appropriate trust. First, we provide an overview of the history of understanding appropriate trust in automated systems. Next, we describe our systematic review methodology and the corpus, summarize the current understanding of appropriate trust and propose a Belief, Intentions, and Actions

(BIA) mapping to highlight commonalities and differences between concepts. Following this mapping, we present the results of the systematic review, discussing different ways to measure appropriate trust, types of tasks used, approaches to building it, and results of the appropriate trust interventions. Finally, we discuss the challenges identified in studying appropriate trust and summarize our observations as current trends, potential gaps, and research opportunities for future work.

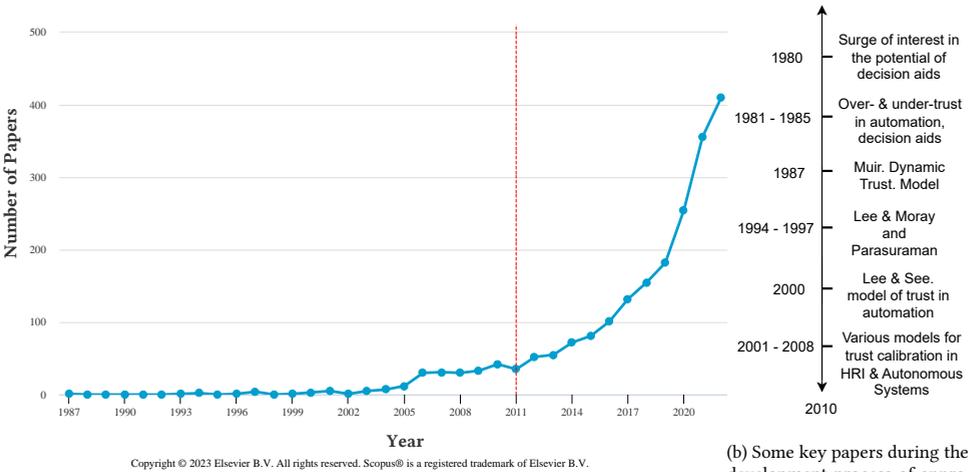
Our main contributions are:

1. A Belief, Intentions, and Actions (BIA) based mapping of appropriate trust and related concepts.  
*Our mapping is result of analyzing how authors define and quantify the abstract notion of appropriate trust and related concepts such as warranted trust, justified trust and meaningful trust;*
2. An exhaustive presentation of different definitions used, measures of appropriate trust, tasks adopted by authors and various methods for building appropriate trust.  
*Our presentation is based on similarities and differences in the approaches that authors have used to define, measure and build appropriate trust in variety of tasks.;*
3. A set of future research opportunities highlighting current trends, challenges and recommendations for future work.  
*Our set of future research directions results from a structured summary of our analysis based on the implications of the approaches (definitions, methods, tasks and measures) adopted by the authors to foster appropriate trust in human-AI interaction.*

## 3.2 BACKGROUND AND HISTORY OF APPROPRIATE TRUST

The topic of appropriate trust has been maturing for years. As technology evolved from automated machines to decision aids, virtual avatars, robots, and finally, AI teammates, appropriate trust has been studied in both depth and breadth across a variety of domains. As discussions of the failures of under- and over-trust in automation begin to appear, researchers started to study how they could calibrate human trust in automation. One of these early studies defined trust calibration as the relation between user reliance and system reliability [28]. Trust calibration was studied by looking at how usage of a system over time changed trust levels, *calibrating* it to the demonstrated reliability of the system. The coining of this term marked the beginning of appropriate trust research within computer science communities, influenced by, but distinct from, previous trust research in e.g. psychology and philosophy.

Understanding the historical context and evolution of appropriate trust allows us to position this work within the broader context of the field. Therefore, in this section, we chronologically describe past efforts to study appropriate trust until the starting point of our systematic search. The background and history of appropriate trust can provide insights about its conceptualization and how technological and social factors have influenced the research field. Moreover, historical analysis can highlight the various theoretical frameworks that have been used to study trust calibration and their limitations. By examining the historical development of this topic, we can better understand its current conceptualization



(a) A time-series chart shows the number of studies from 1987 to 2022

(b) Some key papers during the development process of appropriate trust research.

Figure 3.1: (a) A timeline for the development of appropriate trust as a topic of research from 1987 to 2022 based on the hits from the SCOPUS database. The red dotted line indicates rapid rise of research interest in appropriate trust research. (b) Some key papers during the early stage development of the topic. It's important to note that these are just some key developments and trends in the study of appropriate trust during this time period, and the field has continued to evolve and expand in the years since 2010.

and identify gaps in the literature. In Figure 3.1a & 3.1b, we illustrate the timeline for these developments.

### 3.2.1 1980-1990s: OVER- AND UNDER-TRUST IN AUTOMATION

The question of how and when to trust automation easily pre-dates the modern computer era. In the early 1980s, there was a surge of interest in the potential of computer-based decision aids to support decision-making in various fields [200, 350, 357]. As automation gained further computing power and was able to solve tasks with high complexity, people started relying on the advice provided by these systems. However, early studies found that users tended to over-trust this advice, even in cases where it was clearly incorrect or irrelevant [161, 336]. This phenomenon was referred to as "automation bias" or "automation-induced complacency" [398]. This concern populated further in the late 1980s, where researchers were concerned about the reliability and safety of nuclear power plants.

Over-trust in automation is only one side of the coin, while under-trust is the other. One of the factors identified as contributing to various accidents such as the Baltimore train incident [346] or misuse of anti-ballistic missiles [110] was the tendency of operators to under-trust the information provided by the control systems and not to rely on them. This problem led to the development of various training and simulation programs aimed at improving operator trust in the automation [22].

In this era, researchers were interested in understanding how humans interact with automated systems and errors that can occur when trust is misplaced. Studying the operator role and human-system integration, Knee and Schryver found that over- and under-trust stem partially from consistent, reliable performance by the Intelligent Machines (IM) within

tasks, problems, etc. that the human operator may not fully understand (due to the lack of training, experience, or even the ability to be actively involved in system operation) [196]. According to them, such cases may support "blind reliance" on the part of the human operator, i.e., acceptance of IM control actions without question of its intent or motives. In conclusion, the study of trust in automation from the 1980s to the 1990s sheds light on the pitfalls of misuse, disuse, and overuse of automated systems, highlighting the importance of understanding how humans trust automated systems.

## 3

### **3.2.2 1990s: INTRODUCTION OF HCI AS A FIELD AND FOCUS ON APPROPRIATE TRUST**

In 1987, Muir presented a model based on dynamics of trust between humans and machines for calibrating user's trust in decision aids [267]. At this point in time, extensive research began in the Human-Computer Interaction (HCI) community to examine the factors that influence a human's trust in automation [148]. One of the themes of this research was calibrated trust.

In the CHI '94 conference, Bauhs and Cooke showcased the effect of system information on trust calibration [28]. The authors reported that the system information aided in calibrating users' confidence in system reliability, but it had little effect on users' willingness to take expert system advice. In the same year, Lee & Moray showed how trust and self-confidence relate to the use of automation and refer trust calibration as correspondence between a person's trust in automation and the automation's capability [215]. Following Lee & Moray's work, a seminal article by Parasuraman and Riley [294] in 1997 on the use, disuse, abuse, and misuse of automation indicated the issue of over- and under-reliance on machines due to lack of trust.

Many articles followed Lee & Moray and Parasuraman & Riley research. Ostrom in 1998 [283] showcased that effectively studying trust in automation can help alleviate the uncertainty in gauging the responses of others, thereby guiding appropriate reliance. Tangentially, Kaber, and Endsley introduced the concept of situational awareness to tackle the issue of mistrust in automated systems in the same year [103]. Thus, the emergence of trust calibration studies signalled and ushered in a greater focus on user-centered design as a means of minimizing automation dis- and mis-use.

### **3.2.3 2000s: EMERGENCE OF APPROPRIATE TRUST AS A KEY TOPIC OF RESEARCH**

A seminal article by Lee & See in 2004 provided the first conceptual model of the relationship among calibration, resolution, and automation capability in defining appropriate trust in automation [216]. This work by Lee & See was built on the key work by Cohen et al. in 1998 [71]. The Lee & See model was based on purpose, process, and performance dimensions of information that describe the goal-oriented characteristics of the agent to maintain an appropriate level of trust.

In 2006, Duez et al. [98] followed Lee & See's model to study information requirements for appropriate trust in automation, while Dongen and Maanen [374] investigated whether calibration improves after practice and whether calibration of own reliability differs from calibration of the aid's reliability. Thus, researchers were able to develop models of information communication [98] and asymmetrical reliability attribution [374] in automated

systems, which improved understanding of how users calibrated their trust over time. Following the mentioned works and literature on calibrated trust, the Human-Robot Interaction community developed an of understanding appropriate trust in robot capabilities, such as Freedy's et al. measures of trust in human-robot interaction for detecting over- and under-trust in 2007 or Hancock *et al.*'s 2011 meta-analysis of factors affecting trust in Human-Robot Interaction [138]. Their results indicated that improper trust calibration could be mitigated by manipulating robot design, focusing on quantitative estimates of human, robot, and environmental factors. Similarly, Sanders et al. [323] provided a model of human-robot trust targeting performance, compliance, collaboration, and individual human differences to study how human trusts can be calibrated in situations of over- and under-reliance.

The topic of appropriate trust also started to pick up in industrial settings during the 2000s. For example, in 2008, Wang and their colleague from a defense R&D studied the effectiveness of providing aid reliability information to support participants' appropriate trust in and reliance on a combat identification aid [389]. Their results showed that participants who needed to be made aware of the aid's reliability trusted in and relied on the aid feedback less than those who were aware of its reliability, highlighting appropriate reliance on the aid.

Thus, the emergence of appropriate trust as a prominent topic in the 2000s was marked by the increasing prevalence of automation and innovative steps taken by researchers to study the role of this topic. Notably, Lee & See's 2004 article [216] introduced a conceptual model that interconnected calibration, resolution, and automation capability to define appropriate trust in automation. This work which was built on Cohen's et al. work [71] was followed by many authors such as [98, 138, 323, 389] where fresh insights were seen considering purpose, process, and performance dimensions of information, offering a deeper understanding for trust calibration. Furthermore, the relevance of appropriate trust extended to industrial settings, as demonstrated by studies on combat identification aids and defense technology [390].

### 3.2.4 PARALLEL DEVELOPMENTS: INFLUENTIAL DOMAINS

While research in automation has made significant contributions to our understanding of how people develop and calibrate their trust in computer systems, appropriate trust is also studied in a variety of other fields, including psychology and philosophy. In many cases, our current understanding of appropriate trust have in fact stemmed from the paradigms established in these domains [119, 216, 293].

Different disciplines study appropriate trust differently, however they all seek the capacity for accurate trust assessment, with the goal of establishing a robust basis for augmented decision-making. Appropriate trust has been studied extensively in *psychology*. It is understood as trust that is based on a rational assessment of the risks and benefits of trusting another person or source of information [222, 326]. For example, Evans et al. showcased that older children (9-10 years) are more sensitive to changes in trustee's characteristics, suggesting that they are not only more trusting, but more discerning in their decisions of when to trust [105]. Similarly, Barnard showcased that how medical professionals change their attitudes and behaviors to gain trust of their patients and

proposed which conditions would win justified trust<sup>1</sup> of patients in them [27]. Therefore, in *human-human interaction* the ability to accurately calibrate trust is essential for building and maintaining strong relationships, as it helps individuals to avoid betrayals and to cultivate mutual respect and understanding. Overall, appropriate trust is an important aspect of social functioning and well-being.

In *Philosophy*, the concept of appropriate trust is closely related to the idea of epistemic responsibility, which emphasizes the need for individuals to take responsibility for their beliefs and to use appropriate methods for evaluating evidence and making judgments [30, 302, 316]. In particular, according to Onora O'Neill, appropriate trust involves a "reasonable reliance on another's goodwill, competence, and reliability" [280]. Other philosophers, such as Karen Jones [180] and Katherine Hawley [141], have also explored the concept of appropriate trust and the importance of carefully calibrating one's trust based on various factors, such as past experiences, social norms, and situational factors.

The research interest in trust calibration evolved slowly compared to promoting trust in automation [67]. This can be partly due to a higher interest in understanding multidimensional aspects of trust, and partly due to the complex nature of automation systems. However, in the last ten years (2012-2022), interest in appropriate trust research has grown drastically, see Figure: 3.1a. This trend is likely driven by the increasing cognitive complexity of AI and ubiquity of interpersonal interactions, as well as organizational interest. Therefore, it has become timely to provide an in-depth literature overview of the state-of-the-art for building appropriate trust in AI. We follow the methodology outlined in this section to provide a comprehensive overview of studies from 2012 till June 2022 in the following section with our systematic review methodology.

### 3.3 SYSTEMATIC REVIEW METHODOLOGY

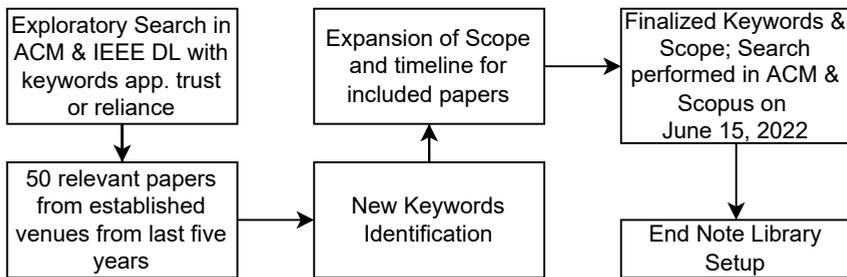


Figure 3.2: Search process for preparing the corpus of the systematic review

We conducted a systematic review to understand (a) current understanding about building appropriate trust in AI, (b) how appropriate trust and its related concepts have been defined and conceptualized, and (b) what measures and methods have been made to achieve appropriate trust in AI. We adapted the procedure by Calvaresi et al. [53] by developing the research protocol following inclusion and exclusion criteria. For search and identification

<sup>1</sup>Different terms have been used in the literature which are related to appropriate trust such as "Justified Trust", "Optimal Trust", etc., refer Section 3.1 for details.

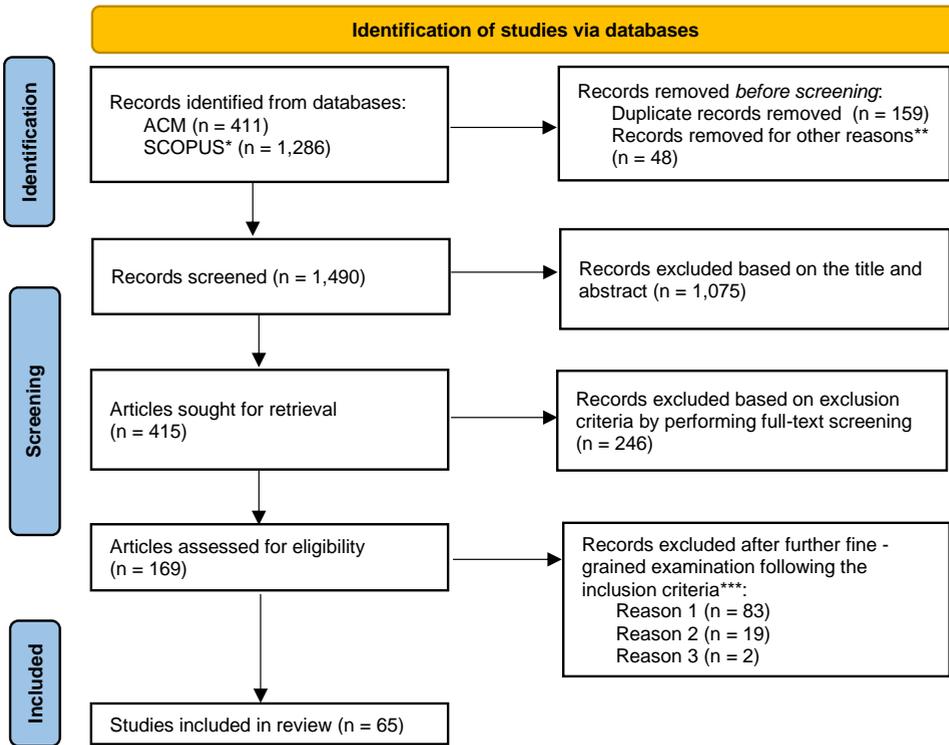


Figure 3.3: \*SCOPUS data includes all retrieved databases such as IEEE, Springer, SAGE, etc. ACM data was excluded from the search as it was taken from ACM DL.

\*\* Other reasons include: records not retrieved, broken URL, and blank pages in the published record.

\*\*\* Reason 1: The article’s focus is NOT on appropriate trust derived from the primary or secondary research question. Reason 2: The article does not use a method or a measurement technique to measure or calculate trust. Reason 3: The article is published as a short version of a long paper (In this case, we included the longer version of the article).

of the relevant articles, we followed the PRISMA guidelines [288]. The specifications of these guidelines is illustrated in Figure 3.3.

### 3.3.1 SEARCH STRING

Appropriate trust is a complex concept and the term ‘appropriate’ is often interchangeably used with terms for similar concepts (such as *appropriate reliance*, *justified trust*, etc.) [167, 364]. Therefore, we first conducted an exploratory search to determine which terms for similar concepts are used. In the ACM and IEEE Digital Libraries, we searched for articles with the keywords “appropriate trust” or “calibrated trust” from the last five years<sup>2</sup>. This exploratory search produced 186 results. Among these 186 results, we focused on articles from four of the most most reoccurring and relevant computer science venues,

<sup>2</sup>This phase was conducted in May 2022. We decided for the last five years as it coincides with the recent rise of interest in appropriate trust research.

FAccT, CHI, IUI, and HRI. We selected 50 articles (FAccT: 6, CHI: 20, IUI: 12, and HRI: 12) with the highest use of keywords and similar concepts throughout the articles.

We manually reviewed every title, keyword, and abstract to find new keywords to be included in our final search string (e.g., “optimal trust”, “justified trust”). We iterated different combinations of the keywords until all papers deemed relevant in the exploratory step appeared among the ACM & IEEE Digital Libraries search results. Analyzing the text of the relevant articles and their references, we noticed that scholars from the Computer Science community often cite scholars from other disciplines who also study appropriate trust. These disciplines include engineering, social sciences, psychology, mathematics, and decision sciences. Therefore, we decided to include these subjects in our search criteria. Furthermore, we decided to broaden our timeline to include articles published in the last decade<sup>3</sup> (2012-2022) after examining the references of the articles. Figure 3.2 visualizes our search process and string finalization. The final search string used in ACM and SCOPUS search is:

(( *"appropriate trust"* ) OR ( *"calibrated trust"* ) OR ( *"warranted trust"* ) OR ( *"justified trust"* ) OR ( *"optimal trust"* ) OR ( *"responsible trust"* ) OR ( *"trust calibration"* ) OR ( *"over trust"* ) OR ( *"under trust"* ) OR ( *"over-trust"* ) OR ( *"under-trust"* ) OR ( *"meaningful trust"* )) AND PUBYEAR > 2011 AND PUBYEAR < JUL 2022 AND ( LIMIT-TO ( SUBJAREA , "COMP" ) OR LIMIT-TO ( SUBJAREA , "ENGI" ) OR LIMIT-TO ( SUBJAREA , "SOCI" ) OR LIMIT-TO ( SUBJAREA , "PSYC" ) OR LIMIT-TO ( SUBJAREA , "MATH" ) OR LIMIT-TO ( SUBJAREA , "DECI" )) AND ( LIMIT-TO ( LANGUAGE , "English" ) )

### 3.3.2 SELECTION CRITERIA

Our search string generated 1,697 articles from the ACM and SCOPUS databases. This phase of generating the final list of articles based on the search string was conducted on June 15, 2022. The screening of articles was carried out manually in three stages: (A) title and abstract screening based on the inclusion criteria, (B) full-text screening based on the exclusion criteria, and then (C) full-text screening with a fine-grained examination based on the inclusion criteria. Our inclusion criteria were:

1. **Language:** The article should be in English.
2. **Peer-Reviewed:** The article should have been peer-reviewed. For example, articles from arxiv, OSF, magazine articles, etc., were excluded.
3. **Format:** Only full and short articles were included so that all the reviewed articles could contain similar details about a study. Therefore, posters, dissertations, workshop papers, workshop calls, etc., were excluded.
4. **Publication Singularity:** Only the complete version of the article is included.
5. **Human-Centered:** Studies needed to have some form of human involvement to be included. For instance, full simulated multi-agent models were excluded.

<sup>3</sup>Since the aim is to identify the current trends and understand recent works in appropriate trust research, we chose to restrain this work to papers published in the last decade

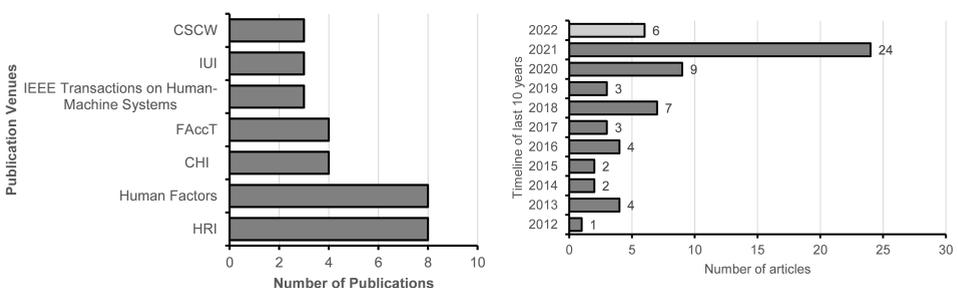
6. **Inclusion of a Definition:** For a paper to be included, it should have an explicit definition or implicit definition through either referencing previous work or describing measures of appropriate trust or the similar concept (calibrated trust, warranted trust, etc.).
7. **Conceptualization of Appropriate Trust:** The articles should conceptualize appropriate trust with measurable constructs or a similar concept. For example, the article uses measurable constructs for appropriate trust.

After applying the inclusion criteria in a two-step abstract and full-text screening process, 169 articles remained. On these 169 articles, we performed further fine-grained examination based on the following criteria:

1. **Contribution Scope:** Articles whose primary contribution was unrelated to appropriate trust were excluded. Articles discussing the need for appropriate trust without any direct contribution to define, measure, or model it were also excluded.
2. **Contribution Type:** Surveys, Scoping Reviews, and Literature Reviews were excluded.

The research team registered the protocol of the review with Open Science Foundation (OSF)<sup>4</sup> to make the selection of reviews a transparent process. Once the registration was completed, the first and second author independently examined the full text of 169 articles. Both authors used the Rayyan web app [286] to organize their decisions. When there were discrepancies between their decisions, the two researchers involved the senior author in discussing it. This discussion process resulted in the final list of 65 articles for the systematic review.

### 3.3.3 CORPUS OVERVIEW AND ANALYSIS



(a) Number of the selected papers per top six publishing venues (b) Number of papers per year from 2012 to 2022

Figure 3.4: Distribution of selected articles over last ten years and top six of their publication venues. Please note that the data for 2022 is incomplete since the data collection for this literature review was conducted in the mid of June 2022.

The final corpus consists of 65 papers, on which we first performed a metadata analysis. We were interested in the publication venues, timeline of publications, and application

<sup>4</sup>[https://osf.io/c78tw/?view\\_only=16c398038f474b9b8922277a3fd94c87](https://osf.io/c78tw/?view_only=16c398038f474b9b8922277a3fd94c87)

scenarios. The top six publication venues and chronological distribution of articles over the last ten years are shown in Figure 3.4a and 3.4b.

In Figure 3.4a, we can observe that the most popular venues, among others are Human Factors and Ergonomics Society (Human Factors) and HRI ( $n = 8$  each) and CHI & FAccT ( $n = 4$  each, idem) which account for 47.3% of the final corpus. Also, the last five years have experienced a growth in the number of publications related to the appropriate trust, see Figure 3.1. This trend reflects a growing interest in human-centered AI and the importance of studies focusing on appropriate trust, distinct from enhancing trust in AI.

## 3

### 3.4 DEFINITIONS AND RELATED CONCEPTS

Appropriate Trust in AI systems is growing rapidly as a research field for both researchers and practitioners. To understand how to achieve appropriate human trust in AI (Human-AI trust), it is important to understand how we define it and its related concepts. The increasing interest in Trustworthy AI research [358] has brought to light a growing need for clarity among the community regarding the different concepts and definitions related to appropriate trust in AI.

Terms like “appropriate trust”, “calibrated trust” and “appropriate reliance” are often used interchangeably in prior research [327]. There have been debates in the community about what appropriate trust is and how different concepts related to appropriate trust are different or similar, for instance during the CHI TRAIT workshop in 2022 [25] and the CSCW '23 workshop on “Building Appropriate Trust in Human-AI Interactions” [10]. These debates are a result of the complex nature of trust in AI systems, which can be difficult to understand and evaluate. In this systematic review, we identified different terms related to appropriate trust in the literature, the most common ones being calibrated trust (number of articles ( $n$ ) =16), appropriate reliance ( $n=8$ ) and warranted trust ( $n=6$ ). The full list of terms is available in Table 3.1, with the corresponding definitions as given by the papers. We can see from Table 3.1 that there is often more than one definition of appropriate trust or its related concepts. This discord and diversity among different concepts motivates us to establish links between them and present a unified mapping.

#### 3.4.1 A BELIEF-INTENTIONS-ACTIONS (BIA) MAPPING

Given the number of terms and slightly different definitions that exist, our first aim is to achieve a clearer understanding of the different concepts surrounding appropriate trust. To this end, we grouped all the presented concepts at different levels of human perception in a conceptual mapping, following Michael Bratman’s theory of human-practical reasoning [38]. In this subsection, we will first discuss the relationships among appropriate trust and its related concepts following this mapping. Following, we relate the concepts to the definitions presented by the authors of the included papers.

We illustrate our categorization of the concepts associated with appropriate trust in Figure 3.5. Figure 3.5 presents a Belief, Intentions, and Actions (BIA) mapping of appropriate trust and related concepts. These levels allow us to separate the different perspectives on trust as a belief, intention, or action. More specifically, *Beliefs* describe a perception of the world and the other agents in it, including beliefs about other agents’ intentions and actions. Beliefs may or may not be justified based on current information about the world

Table 3.1: Definitions of Appropriate Trust and its related concepts, A \* represents articles before the year 2012 or after the end of search date. The \* articles were not included in the review process.

| Keyword   | Definition   |
|---|--|
| Appropriate Trust - Based on System Performance or Reliability: | <ol style="list-style-type: none"> <li>1. Appropriate trust is the <b>alignment between the perceived and actual performance</b> of the system. Appropriate trust is to [not] follow an [in]correct recommendation." [406].</li> <li>2. If the <b>reliability of the agent</b> matches with user's trust in the agent then trust is appropriately calibrated [279].</li> <li>3. In human-robot teaming, appropriate trust is maintained when the human uses the robot for tasks or subtasks the robot <b>performs better</b> or safer while reserving those aspects of the task the robot <b>performs poorly</b> to the human operator [282].</li> </ol>   |
| Based on TW and beliefs:  | <ol style="list-style-type: none"> <li>1. Appropriate trust in teams happens when one teammate's trust towards another teammate corresponds to the latter's <b>actual trustworthiness</b> [182].</li> <li>2. We can understand 'appropriate trust' as obtaining when the trustor has <b>justified beliefs</b> that the trustee has suitable dispositions [80].</li> </ol>  |
| Based on the Calculations:                                      | <ol style="list-style-type: none"> <li>1. "Appropriate trust is the <b>fraction of tasks</b> where participants used the model's prediction when the model was correct and did not use the model's prediction when the model was wrong; this is effectively participants' final decision accuracy" [394].</li> <li>2. <b>Trust appropriateness was calculated</b> by subtracting a_ideal from a participant's allocation for a given round. Thus, a positive value indicates trust that is too high, a negative value indicates trust that is too low, and 0 indicates calibrated, appropriate trust [173].</li> <li>3. The level of trust a human has in an agent with respect to a contract is appropriate if the likelihood the human associates with the system satisfying the contract is equal to the <b>likelihood</b> of the agent satisfying that contract* [412].</li> <li>4. The term appropriate trust then is the <b>sum of appropriate agreement and appropriate disagreement</b> of humans with the AI prediction[225]</li> </ol> |
| Warranted Trust   | <ol style="list-style-type: none"> <li>1. A match between the actual <b>system capabilities</b> and those perceived by the user" [328].</li> <li>2. "Human's trust in a AI model (to Contract - C) is warranted if it is caused by <b>trustworthiness in the AI model</b>" [167]</li> </ol>  |
| Justified Trust   | <ol style="list-style-type: none"> <li>1. "Justified Trust is computed by evaluating the human's understanding of the model's decision-making process. In other words, given an image, justified trust means users could <b>reliably predict</b> the model's output decision." [6]</li> </ol>  |
| Contractual Trust   | <ol style="list-style-type: none"> <li>1. "When a trustor has a belief that the trustee will stick to a <b>specific contract</b>". [167]</li> <li>2. "<b>Belief in the trustworthiness</b> (with respect to a contract) of an AI." [113]</li> </ol>  |
| Calibrated Trust  | <ol style="list-style-type: none"> <li>1. "Trust calibration is the <b>process by which a human adjusts their expectations</b> of the automation's reliability and trustworthiness". [212]</li> <li>2. Calibrating trust is if explanations could help the annotator <b>appropriately increase their trust</b> and confidence as the model learns [126]</li> <li>3. Trust calibration refers to the correspondence between a person's trust in the automation and the automation's capabilities* (based on Lee &amp; Moray [215] and Muir [267]) [216].</li> </ol>   |
| Well-placed Trust*  | "[T]he only trust that is <b>well placed</b> (intention) is given by those who understand <b>what is proposed</b> , and who are in a position to refuse or choose in the light of that understanding [280].  |
| Responsible Trust   | "The area for responsible trust in AI is to explore means to <b>empower end users to make more accurate trust judgments</b> ". [223].  |
| Reasonable Trust*   | "Reasonable trust requires <b>good grounds for confidence in another's good will</b> , or at least the absence of good grounds for expecting their ill will or indifference.". [21]  |

and past agent behavior [113]. Second, *Intentions* represent the deliberative state of the human – what the human has chosen to do. Intentions are desires to which the human has to some extent committed [123]. Finally, *Actions* describe events as they actually occur in the interaction [11], such as a doctor actually offering a patient an in-person consultation. In essence, this mapping provides a mechanism for separating each interaction into three parts; the informational state (beliefs), motivational and deliberative states (intention), and reactive activity (actions).

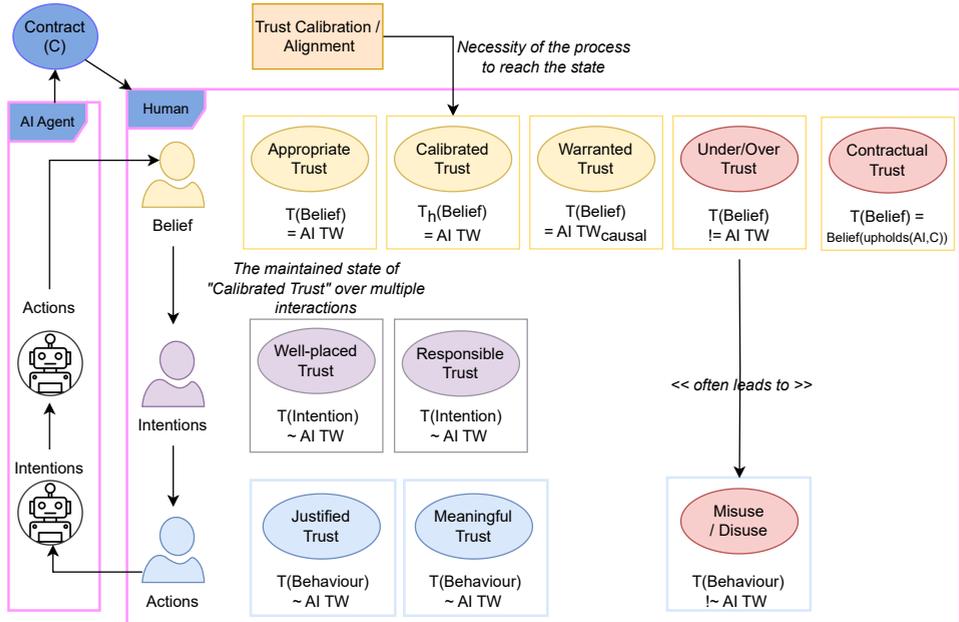


Figure 3.5: In this figure, we present a Belief, Intentions and Actions (BIA) mapping of Appropriate Trust and related concepts. The pink outline represent the elements linked with the human ( $h$ ) and the AI agent entity. The black coloured circle with a robot icon represents the AI agent. For brevity, when writing  $T(\text{Belief})$ , we mean  $\text{Trust}(\text{human}, \text{agent})(\text{Belief})$  and for  $\text{TW}(\text{agent})$  we use AI trustworthiness. Also, Under/Over trust and Contractual trust are represented in different colors as these types of trust aren't (necessarily) appropriate.

Our mapping identifies two actors: the Human and the AI agent. The human actor is illustrated with a 'user' icon and the AI agent is represented by a robot icon. To help distinguish between the different concepts, we formally define them. In our definitions, we use the following variables:  $human \in H$  for a human,  $agent \in A$  for an AI agent, and  $T_{(x,y)}$  to denote the trust that trustor  $x$  has in trustee  $y$ . We then use the following notations:

$T_{(human,agent)}(\text{Belief})$  = Trust of the human in AI agent is human's belief in AI agent's  $\text{TW}$

$\text{TW}_{agent}$  = Actual trustworthiness of the agent

$T_{(human,agent)}(\text{Intention})$  = Human trust in agent is about the human's intentions towards agent

$T_{(human,agent)}(\text{Behaviour})$  = Human trust in agent is about the human's behaviour towards agent

We have divided the definitions of appropriate trust in the Table 3.1, see keyword "Appropriate Trust" based on the similarities such as system performance or reliability, beliefs, and calculations. Based on this division, we can formulate our conceptualization of appropriate trust. We define trust to be appropriate when the human's trust formed by beliefs about the AI agent's trustworthiness (denoted as  $TW_{(human,agent)}(Belief)$ ) is equal to the AI agent's actual trustworthiness  $TW_{agent}(Actual)$ , refer eq. 3.1.

$$\text{Appropriate Trust} \iff T_{(human,agent)}(Belief) = TW_{Agent} \quad (3.1)$$

From this foundation, we go on to differentiate between the many related concepts for appropriate trust which we have encountered. As a note to our readers, the definitions and terms presented in the Table 3.1 don't always match one-to-one with our conceptualization in Figure 3.5, because sometimes different authors define the same term in different ways.

First, we consider **calibrated trust**, the most common term in the reviewed corpus, which introduces notions of dynamic trust and trust variations to the Human-AI interaction [243, 324]. We define calibrated trust as similar to the appropriate trust in that a human's trust belief about the agent corresponds to their actual trustworthiness. However, calibrated trust necessarily involves a process of *trust calibration* or *trust alignment* that corrects for over- and under-trust over the course of time and repeated interactions. We postulate that appropriate trust is the maintained state of "calibrated trust" over multiple interactions. Nevertheless, human trust in an AI system may be appropriate even without calibration.

Distinct from appropriate or calibrated trust is **over-trust** *i.e.*, the human's trust beliefs in an AI agent's is greater than the AI agent's actual trustworthiness  $TW_{Agent}$ . Similarly, when the human's trust belief in an AI agent is less than the AI agent's actual trustworthiness  $TW_{Agent}$  then a state of **under-trust** is reached.

Next, **warranted trust** is defined as trust caused by the trustworthiness of the AI agent. More specifically, we talk about warranted trust if there is a causal relationship between trustworthiness of the trustee and the trust of the trustor [113]. Though we expect warranted trust to mostly be appropriate, not all appropriate trust needs to be warranted. In other words, while trust that is well-supported by evidence and reasoning is probably appropriate, there may be situations where trust is appropriate even if there is no clear evidence or justification to support it. For instance, if an e-commerce website has a polished and visually appealing design, it may create an initial positive impression in the user's mind. This positive impression, in turn, may lead the user to trust the website's content to some degree, even though they lack in-depth evidence about the product's quality. Finally, **contractual trust** in the AI agent is based on the belief that the AI agent will uphold an explicit contract (upholds(AI,C)) which specifies what the AI agent is expected to do [140, 354]. Here, the contract may refer to any functionality of the AI agent that is deemed useful, even if it is not concrete performance at the end-task that the AI agent was trained for [167]. It is important to highlight that contractual trust differs from many other definitions in that it does not directly imply appropriateness, as the human's beliefs about the agent might not be related to its actual trustworthiness.

Unlike the three above-mentioned concepts, **well-placed and responsible trust** are built around intentions. Here, both well-placed and responsible trust are defined as intentions about how to act towards the agent (denoted as  $TW_{human}(Intent)$ ). Meaning, if a human has well-placed trust, that means its intentions are correct given the trustee's

trustworthiness. Here, intentions are being referred as human's intentions to rely on, cooperate with, or be vulnerable to the agent (trustee) in some way. These intentions reflect the human's willingness to trust the agent.

Trust is justified when a human's behavior is appropriate given the agent's trustworthiness. In this case, the human actor is acting trustingly towards an agent. The ability of a user to evaluate trust does not make the AI agent more accurate, robust, and reliable in itself; it can only, at best, make the use of the AI by the human more accurate, robust, and reliable leading to **justified trust**. In contrast, when human's trust is not justified based on the AI's agent trustworthiness, it is plausible that the user can lean towards misuse or disuse of AI.

3

So far, we have described our mapping related to the human actor. Now, we shift our focus to the AI agent. As mentioned earlier, we consider the AI agent attributed with intent. Here, the AI agent can form an intention based on the human behavior which can be associated with the action(s) it can take resulting in human to form beliefs about the AI agent's trustworthiness, making it a closed-loop process. This process helps in arriving at the belief that a contract, as outlined by Jacovi et al [167], has been established between the pair and will be adhered to in the future. In other words, an AI agent can form an intention based on human behavior or actions by observing the human behavior. Based on its intention it can decide how to respond by taking an action. By doing so, the AI agent can ensure that it acts in accordance with the contractual agreement (actions the AI agent is authorized to take and the expectations for how the AI agent should behave) and maintains the trust of the human.

In addition to the aforesaid concepts, we found a few further, more minor and less defined terms which are not completely covered in Figure 3.5 and these are the ones which we haven't defined explicitly, namely meaningful trust (closely linked to *Justified Trust*), optimal trust (*Justified Trust*), moral trust (*Responsible Trust*), capacity trust (*Perceived Capability*), and well-deserved trust (*Justified Trust, Warranted Trust*). Our list of related concepts is not exhaustive, and there could be further concepts that appear outside the domain of our review that we haven't included in our search criteria.

In summary, we have described the distinction between appropriate trust and related concepts stemming from human beliefs, intentions, and actions. We believe this is one of the first conceptualization in human-AI interaction research to describe, associate, and categorize various concepts in a single framework, which could help reduce the discord among the community on approaching the concept of appropriate trust.

### 3.5 RESULTS OF THE SYSTEMATIC REVIEW

In this section, we review how authors of the included papers define and measure appropriate trust<sup>5</sup>, what different domains, settings, and tasks they employ, methods for building appropriate trust and the results achieved.

---

<sup>5</sup>We follow the same terminology (appropriate trust/ calibrated trust/ warranted trust etc.) as the authors of the reviewed papers to maintain the consistency.

### 3.5.1 MEASURES (HOW TO MEASURE APPROPRIATE TRUST?)

Human trust is studied differently based on whether it's conceptualized as a mental attitude [55, 106], a belief [182, 188, 415] or a behaviour [54, 279, 405]. These approaches are typically linked to specific measures which either focus on subjectively measuring attitudes or beliefs, or which look at behavior which demonstrates human trust. As measuring the appropriateness of trust naturally includes measuring trust, we draw the same distinction and divide this subsection into three parts in accordance with Wischniewski et al. [402]: a) Perceived trust, b) Demonstrated trust, and c) Mixed approach. Simply put, we say *perceived trust* is about measuring a person's subjective beliefs, while *demonstrated trust* focuses on their behavior [261]. While measuring perceived trust is typically done via questionnaires, surveys, interviews, focus groups, and similar reporting tools, demonstrated trust is usually about measuring trust-related behaviors (for instance, in the form of reliance). In demonstrated trust, participants are given the option to use or rely on the system. The underlying assumption is that the more often people use or rely on a system, the more they trust it.

#### PERCEIVED TRUST

Among the papers in our corpus, ~40% measure appropriate trust or related concepts by examining a match between the system's capabilities and the user's trust as a belief. The most common strategy to measure appropriate trust was manipulating a system's trustworthiness and using self-report scales to compare how self-reported trust adapts to the trustworthiness' levels. For example, Chen et al. [63] presented participants with either 60%, 70%, 80%, or 90% reliable systems and measured trust through subjective self-report. Similarly, with a within-subjects experimental design, de Visser et al. [88] had participants interact with a system which trustworthiness' levels were manipulated through its reliability from 100% to 67%, 50%, and, finally, 0%. Then, the authors used a self-reported trust scale to measure trust which then through comparison with system's trustworthiness provided appropriateness of trust. In the prior examples, manipulating trustworthiness helped the authors to do a before/after comparison. According to Miller [261] this comparison is crucial for measuring appropriate trust. The authors highlight that without manipulating the trustworthiness of the machine, we cannot establish whether the intervention has correctly calibrated trust.

We found some authors measured perceived trust by performing a match between the trust ratings and the *static* reliability of the robot or the AI system [7, 35, 86, 172, 202, 227, 282]. However, the match between trust ratings and static reliability may not be perfect. There may be other factors such as appearance or behavior that influence how people rate the trustworthiness of a robot or AI system, even if the system is perfectly reliable. Furthermore, this method does not take into account the dynamic nature of trust. Therefore, we argue that it's difficult to match performance levels with subjective scale ratings.

#### DEMONSTRATED TRUST

We found that only ~26% of studies used behavioral measures for appropriate trust, and we identified three approaches to do so. The most common is **agreement percentage**, that is the percentage of trials in which the participant's final prediction agreed with the AI's correct prediction and cases where participants didn't agree with the AI's wrong

prediction [24, 44, 225, 270, 394, 416]. Usually, appropriate trust is seen as a sum of appropriate agreement ratio (human agreement with correct AI predictions) and appropriate disagreement ratio (disagreement with incorrect AI predictions) [76, 225, 279]. Another measure of appropriate trust is related to **switch percentage** [416], that is the percentage of trials in which the participant decided to use the AI's prediction as their final prediction instead of their own initial prediction. However, it is usually not a standalone measure of appropriate trust and is coupled with other measures. For example, Zhang et al. used a statistically significant interaction between switch and agreement percentages and the AI's confidence level [416]. When AI's confidence level was high and the switch and agreement percentage was high (and vice-versa), then trust was deemed appropriate.

A final method is to measure **ideal trusting behavior** during the task beforehand and compare to which extent the actual users' behavior matches it [145, 173]. For example, for an experiment where users have to delegate a number of tasks to AI, it is possible to calculate the most optimal number of tasks to delegate to AI in order to achieve the best speed and performance at a given AI's reliability [145]. The closer users are to this number, the more appropriate their trust in AI is.

### MIXED APPROACH

A total of ~20%<sup>6</sup> studies from our corpus used a combination of both self-report measures and behavioral measures to understand appropriate trust. These measures can be categorized into two different subgroups.

The first subgroup includes measures that focus on participants' decisions and compliance with the system's recommendations along with self-reported scales. For example, Wang and Pynadath measured appropriate trust by letting users decide when and when not to trust a low-reliability robot [391]. They measured self-reported trust by modifying Mayer's scale [242] and used behavioral measure of compliance as dividing the number of participant decisions that matched the robot's recommendation by the total number of participant decisions. Accordingly, when both measures matched the reliability of the robot, the trust was considered appropriate. Similarly, Kaniarasu et al. conducted a study where participants rated trust at trust prompts and used buttons to indicate trust changes [185]. Appropriate trust was measured by examining the degree of alignment of user's trust with the robot's current reliability (high or low). Finally, Zhang et al. measured participants' reliance on AI using two behavioral indicators, agreement frequency and switch-to-agree frequency, as well as via subjective trust ratings [413]. Their diverging reliance and subjective trust ratings results highlight the difference between these two types of measures.

The second subgroup includes measures that examine how participants calibrate their trust over time as they become more familiar with the system's capabilities and policies. For example, Albayram et al. measured how participants calibrated their trust as they grew familiar with the system's capabilities by using subjective responses and number of images allocated to the automation for pothole inspection by varying automation reliability [7]. Similarly, de Visser varied the anthropomorphism of the automation to understand trust calibration and appropriate compliance [88]. By using both subjective ratings and

<sup>6</sup>The remaining 14% of the reviewed papers presented frameworks or theoretical models where no user-study was conducted in which a measure appropriate trust was used.

a compliance measure, they measured appropriate trust as the match of a user's trust with the actual reliability of the aid. In both of the previous examples, the researchers manipulated the trustworthiness of the system to measure an appropriate level of trust. This approach is in line with Miller et al. [262] who states that '*there must be some known or estimated 'level' of trustworthiness that is manipulated as part of the evaluation.*'

### SYNOPSIS

In summary, measures of appropriate trust typically involve either a comparison of two different measures: trust of the human and trustworthiness of the system, or they involve some form of agreement metric. The first type naturally involves knowing the trustworthiness of the system. Trustworthiness can be defined as absolute (e.g. the system is correct or not) or relative (the system gets better/worse over time). Although the first might give more insight into how good the system is, it does mean the AI needs to be either wrong or right, which needs to be known. The relative measure allows for an easier comparison, as appropriateness is just about whether trust moves up or down in the same direction as trustworthiness. However, if trust is low for a nearly perfect system and slightly higher but still low for a perfect system, it is still inappropriate despite moving in the correct direction.

Comparing trustworthiness with trust naturally also involves measuring trust. In this also, two methods can be distinguished. The first is subjective and behavioural measures based on questionnaires, and the second is on actions. The main disadvantage of questionnaires is that outcomes can be difficult to directly compare with trustworthiness, while it is easier to establish if reliability is correct. On the other hand, questionnaires better capture the concept of trust as a nuanced belief, as reliance behaviour could be caused by more than just high trust. This is also reflected in the differences between behavioral and subjective scales that can occur when both are used [413]. This highlights the disadvantage of seeing appropriate trust in terms of an agreement metric; this is, by definition, about reliance behaviour and often imposes constraints on the type of human-AI collaboration. Given the limitations of most current measures, the option to use different methods simultaneously has the opportunity to offer a more nuanced result. Which mix is the best might depend highly on the collaboration between the human and AI.

An example of simultaneous use of different methods is (a) the use of validated questionnaires to measure perceived trust combined with (b) behavioural measures to measure demonstrated trust could offer a more insightful measurement than use of one alone [402]. The underlying assumption is that these measures provide an accurate understanding of human's trust. However, as human trust is a multi-dimensional concept its measurement based on scales or behaviour might not provide its complete understanding [348]. For example, behavioral measures are context-specific and may not generalize well across different situations and subjective measures may involve participants' individual biases or the willingness to disclose their true feelings. Therefore, we propose the next steps in determining how to measure appropriate trust should be to examine combination of measures other than perceived or demonstrated trust. These measures can include personality traits [120], past experiences [131], social norms [360], and cultural values [420], and how these measures can differ across different contexts and populations. The importance of the context of the task and domains of the study for measurements highlights a need to explicitly define and describe these for studies in appropriate trust.

### 3.5.2 TASKS

In this section, we describe the tasks and domains observed in the corpus of this review. We cluster these tasks around distinguishing characteristics which emerged.

We grouped all studies into different application domains to get an overview of the tasks. In enumerating the domains seen within our corpus of papers (See Table 3.2), we observe that military operations, transport, and domain agnostic applications are the most common in appropriate trust research. On a more granular level, we see that tasks such as automated driving ( $n = 14$ ), prediction and classification ( $n = 14$ ), and reconnaissance ( $n = 8$ ) are most commonly given to users. Human-AI collaborative tasks such as working in a military environment with humans (e.g., [394]) and teaming for military missions [279, 364] are the particularly preferred cases of the reviewed articles. The popularity of military and transport application fields within the study of appropriate trust could follow from the more severe risks associated with the incorrect use of technology in those settings.

3

| <i>Domain</i>   | <i>Tasks</i>  |
|-----------------|---|
| Military        | <ul style="list-style-type: none"> <li>Object recognition [68, 172, 227], Prediction [35], Reconnaissance [85, 136, 171, 173, 274, 279, 391, 408], Remote operation [87, 176, 364], Search and Rescue [185], Non-Experimental [282, 365]</li> </ul> |
| Transport       | <ul style="list-style-type: none"> <li>Automated Driving [5, 18, 19, 102, 132, 144, 191, 201, 202, 229, 264, 373, 386, 401], Non-Experimental [321]</li> </ul>  |
| Domain agnostic | <ul style="list-style-type: none"> <li>Classification [272, 406], Multi-arm trust game [74], Object Recognition [413], Non-Experimental [66, 149, 164, 182, 328, 333, 337]</li> </ul>   |
| Healthcare      | <ul style="list-style-type: none"> <li>Classification [145, 269, 270, 272], Meal Design [44], Non-Experimental [113, 167, 223, 245]</li> </ul>  |
| IT              | <ul style="list-style-type: none"> <li>Prediction [126], Classification [24], Question Answering [24], Non-Experimental [174]</li> </ul>  |
| Justice         | <ul style="list-style-type: none"> <li>Prediction [225, 394], Classification [24], Question Answering [24]</li> </ul>   |

Table 3.2: Domains and Associated Tasks Across Our Corpus

When analyzing the breadth of user studies included in this review ( $n = 46$ , 45.6% between, 32.6% within subject, 15.2% mixed design), we see a number of patterns emerge in the characteristics of the tasks assigned to participants. We group those characteristics along the dimensions of risk, dynamism, and users' expertise. Interestingly, only three studies [270, 274, 373] preform a non-controlled experiment, relying on think-aloud sessions, co-design, and interview sessions. They targeted medical, mobility, and military experts for interaction design. To some extent, this does suggest a lack of space within appropriate trust research for the voices of users and stakeholders, and little input on its design processes on their part.

### **RISK**

We highlight risk as an integral part of experimental set-ups, as vulnerability is a key element of trust [216, 242]. Yet, it can be overlooked in studies of human-computer trust. We differentiate between explicit and implicit risk using the criteria proposed by Miller [261]. In these criteria, trust is characterized by the presence of vulnerability and stakes, which introduce a downside to inappropriate trust. The user must be aware of these stakes throughout the experiment, so that actions can be adjusted to accommodate risk. We see that 78.3% of studies include an element of risk in their design. This element is largely implemented in one of two ways; simulated through points gained and lost (45.7%); or incentivized through performance-based pay bonuses (21.7%). Only one study [8] used a task which was risky in the experimental setting itself, namely disassembling traction batteries in a recycling context.

The remaining papers rely on the understood risk of a given task (automated driving and remote operation) in the real world to assume users would engage realistically with their experiment [87, 176, 225, 364, 401], or do not discuss risk in their methodology [76, 406]. Given the importance of risk to trust, it is difficult to argue that users in such studies demonstrated trust at all, with no consequences attached to over- and under-trust, users may rely on, and positively perceive a system regardless of its actual trustworthiness.

### **DYNAMISM**

The next element of task design we analyzed is dynamism, that is, changes in Human-AI trust over time informed by the history of interaction [148]. Specifically, we investigate whether studies measure trust levels at multiple points, thus accounting for this dynamic aspect of trust. Across all studies, we find that 63% measure trust more than once throughout the task. In cases of automated driving tasks, this can sometimes even be a continuous measure of trust derived from driver behaviour [5, 191, 401]. Meanwhile, a third of studies measured trust only once throughout the experiment, reducing the complexity of the trust relationship to one snapshot.

Moreover, most of the studies reviewed were either laboratory-based which used simple tasks or theoretical models, which further fails to reflect real-world scenarios. Thus, generalizability of these findings to more complex and dynamic real-world situations is uncertain.

### **PARTICIPANT EXPERTISE**

Overall, 67.4% of studies recruited non-expert participants, because often researchers design tasks so that the participant pool felt equally qualified to complete them without any

specific training [19, 24, 44, 74, 394, 416]. Recruitment of non-experts also occurred for the tasks that could require more specialized knowledge, such as military-related tasks [185]. The main reason could be that candidates with required expertise are not available and/or are not easily found. This claim can be supported by the fact that all automated driving studies recruited licensed drivers to their experiments, while only three non-automated-driving user experiment studies recruited expert participants [126, 269, 364]. Given that a users' perception of their expertise can affect the extent to which they trust and rely on the automated system [413], participant expertise should more closely align with the expected expertise of the end user, for more realistic results.

### 3.5.3 METHODS FOR BUILDING APPROPRIATE TRUST (HOW TO ACHIEVE IT?)

In this section, we describe what different approaches were taken towards achieving appropriate trust in the reviewed corpus. A categorization of the methods revealed four broad categories: (1) Improving system transparency, (2) Cognition and perception, (3) Models, guidelines, theories and frameworks, and (4) Relational framing and continuum of trust. These are further shown in Figure 3.6.

#### IMPROVING SYSTEM TRANSPARENCY

The first category of methods aims to achieve appropriate trust by adding transparency to systems. About 52% of articles in our corpus target improving transparency of the system to build appropriate trust, *i.e.*, informing users about the specific capabilities and limitations of AI. This indicates that there is a common assumption that improving system transparency can help the human user to better decide when to trust or distrust the AI system.

One way transparency is improved is through providing **Explanations**. Explanations focus on the inner-workings of the AI systems ( $n = 16$ ), appearing either for every AI recommendation [126, 364, 413, 416] or under specific circumstances. For example, Adaptive Explanations by Bansal et al. appear only for the predictions where the AI is quite confident and are absent for the low confidence predictions as a way to avoid human over-trust in the latter case [24]. This explanation method was found to be effective in trust calibration, as here the AI system adjusts to the user's attitude and behaviour following the signs of over- and under-trust. To further mitigate over-trust, Lakkaraju et al. call for designing explanations as an Interactive Dialogue where end users can query or explore different explanations for building appropriate trust [208].

Another way to instill transparency is through **Confidence Scores** of the AI models to align the user's trustworthiness perception of the system with the actual trustworthiness ( $n = 12$ ). These scores reflect the chances that the AI is correct, thus relating to its competence and capability. According to Zhang et al., confidence scores are a simple yet effective method for trust calibration [416]. However, it does not necessarily improve AI-assisted decision-making [24]. Furthermore, confidence scores are not always well calibrated in ML classifiers [273] which can lead to inappropriate trust.

A combination of explanations and confidence scores has been used for appropriate trust as well under the term of **Informed Safety & Knowledge** in relation to autonomous vehicles [191]. The confidence scores informed the drivers of the vehicle's safety. At the

same time, the explanations were provided to demonstrate the vehicle's knowledge of any maneuver, enabling the drivers to adjust their level of trust in the system appropriately.

Similar to confidence scores, **Uncertainty Communication** (n=3), *i.e.*, emphasizing the instances when AI is "unsure" of a prediction or does not have a definite answer, can also calibrate trust. For example, an AI agent can yield back the full control to humans and explicitly indicate that it does not "know" the solution [365]. The results of this method demonstrate that it helps users to spot flows in the reasoning behind the AI predictions and when AI is "unsure" about them, and consequently rapidly calibrate their trust.

While confidence scores and uncertainty communication come mostly in a form of a text message, their more anthropomorphized counterpart is **verbal assurances**. Within this method of transparency, the system verbally indicates to the users what it can and cannot do in a form of promises [7, 8] or intent [229]. For example, Albayram et al.'s results show that participants calibrated their trust based on the system's observed reliability following the promise messages [7]. Besides written or verbal indicators, odors, presented as **Olfactory Reliability Displays** [401], can also serve to communicate a change in reliability levels of AI for users to calibrate their trust. The authors communicated a change in reliability levels of an automated vehicle simulator using two odors *i.e.*, lemon for a change to low and lavender for a change to high reliability. Their results indicate that olfactory notifications are useful for trust calibration.

Providing more information about not only the AI capability, but also about the task and the context, or in other words, **Situational Awareness Communication**, can provide transparency to achieve appropriate trust [19, 176]. For example, Azevedo et al. showed that with activation of different communication styles to encourage or warn the driver about Situational Awareness (SA) when deemed necessary helps in calibration of trust [19]. Similarly, Johnson et al.'s results show that warning drivers about SA is effective at increasing (decreasing) trust of under-trusting (over-trusting) drivers, and reducing the average trust miscalibration time periods by approximately 40%.

Studying various methods of improving system transparency for building appropriate trust in AI systems can provide valuable insights. Overall, these works show the value of understanding system transparency and how it can be increased in multiple ways. Some of the most common examples are confidence scores and explanations. However, we also see some unique solutions, such as using olfactory displays or verbal assurances. All these solutions seem promising for improving system transparency, but that communicating system uncertainty or providing real-time situational awareness helps is only sometimes a given. This shows that there is still much to gain, especially in understanding why an AI system is uncertain or what can help to improve its situational awareness for improving system transparency.

### COGNITION AND PERCEPTION

Another group of methods to achieve appropriate trust is related to human factors, and accounted for 21% of the reviewed papers. Several of them focus on the **users' mental model** of AI [282]. The more correct the mental model is, the more likely it is that trust will be calibrated appropriately, which links back to our previous method of increasing transparency. One of the ways to achieve this is through training users how to perform the task and how to collaborate with an AI-embedded system [176, 270]. The results show that

training that emphasized the shortcomings of the system appeared to calibrate expectations and trust [176]. Another way to build a more correct mental model of AI is to let users observe the system's performance over time [24]. By observing the system performance over time in Bansal et al. study [24], participants developed mental models of the AI's confidence score to determine when to trust the AI.

Other human factors are related to **nudging and cognitive forcing functions**. For example, adding friction in the decision-making process of AI to purposefully slow down its recommendation and providing users a nudge gives them an opportunity to better reflect on the final decision [269]. Naiseh et al.'s results show that with a nudging based XAI approach such as, ("*You are spending less time than expected in reading the explanation.*"), users can calibrate their trust in AI. Similarly, introducing cognitive forcing interventions, *i.e.*, not automatically showing AI recommendations but on-demand or with forced wait can significantly reduce over-reliance compared to the simple explainable AI approaches [44].

Another potential method to calibrate trust through understanding human factors was proposed by Johnson et al. [176]. The authors gave participants trust calibration **training** about task-work and teamwork before the task. Their results show that training that emphasized the shortcomings of the autonomous agent appeared to calibrate expectations and trust. Lastly, the characteristics of an AI-embedded system, notably its degree of **anthropomorphism** contributes to appropriate trust [86]. The results showed that increasing the humanness of the automation increased trust calibration *i.e.*, compliance rates matched with the actual reliability of the aid on increasing humanness.

In synopsis, studying cognition and perception can help us to better understand how people interact with AI systems and how they form impressions of AI systems. Also, studying the mental processes involved in perception, learning, reasoning, and decision-making can help us in designing for appropriate trust in AI systems.

### MODELS AND GUIDELINES

Theoretical foundations can provide insights into how to establish appropriate trust in human-AI interaction (n=12) [87, 164, 176, 264, 328, 337]. One example is using models and frameworks to understand how the actual and perceived trustworthiness of AI systems relate to each other. Several papers use different models to explain this relationship and suggest ways to improve it. For instance, Schlicker et al. use two models from organizational psychology to identify factors that influence the match between how trustworthy the system is and how trustworthy the user thinks it is [328]. Similarly, Israelsen proposes a three-level model that compares the user's and AI's abilities, analyzes the user's past experiences with similar systems, and measures the user's willingness to depend on the system [164]. Some similarities between the theoretical models we reviewed are that they often try to explain how the user's perception of the AI system's trustworthiness is influenced by various factors, such as the system's performance, reliability, transparency, feedback, and context. These factors can help us understand how people interact with AI systems. By understanding these factors that influence trustworthiness, we can design AI systems that can be appropriately trusted [256].

Another type of model focuses on the **communication of trustworthiness cues** in AI systems. For example, Liao and Sundar [223] proposed the MATCH model for responsible

trust, which describes how trustworthiness should be communicated in AI systems through trustworthiness cues. With their model, they highlight transparency and interaction as AI systems' affordances for designing trustworthiness cues. Apart from communicating trustworthiness cues, some authors studied building appropriate trust by allowing for real-time trust calibration [5, 136, 333]. For example, Shafi [333] provided a parametric model of machine competence that allowed generating different machine competence behaviors based on task difficulty to study **trust dynamics** for real-time trust calibration. Furthermore, Guo and Yang modeled trust dynamics using **Bayesian inference** when a human interacts with a robotic agent over time [136]. Here, based on the real-time trust values, a human can calibrate its trust in the robot.

Unlike theoretical models, **guidelines** offer practical design solutions to achieve calibrated trust in AI. For instance, a *calibrated trust toolkit* [373] aids transparent design of autonomous vehicles, analogous to methods in section 3.3.1. These guidelines address post-design implementation, offering a road-map for human factors in industrial robots and trust calibration for robotic teammates.

In synopsis, various theoretical models and guidelines have been proposed to understand the mechanisms around achieving appropriate trust in AI. Theoretical foundations, such as the models and frameworks, provide valuable insights into the factors influencing trustworthiness perception. By examining factors like system performance, reliability, transparency, feedback, and context, we understand how users interact with AI systems, ultimately aiding in designing AI systems that can be appropriately trusted.

Furthermore, models like the MATCH model by Liao and Sundar focus on communicating trustworthiness cues, emphasizing transparency and interaction as essential elements in designing trustworthiness cues in AI systems. Like the one proposed by Shafi, real-time trust calibration models offer insights into how trust dynamics can be managed during human-AI interactions, allowing for adjustments based on task difficulty and performance. In addition to theoretical models, practical guidelines play a vital role in achieving calibrated trust in AI. These guidelines offer actionable recommendations for designers and developers, ensuring that AI systems align with their original design intent. It is worth noting that industrial organizations also contribute to this field, offering guidelines that often focus on building users' trust but increasingly recognize the importance of achieving appropriate trust through effective communication and transparency, as emphasized by the Google PAIR guidebook [290].

A nuanced approach is crucial in designing trust models for AI systems, considering the intricate interplay of various factors influencing trustworthiness. Likewise, when confronted with many guidelines on trust in AI, tailored selection and adaptation are crucial to ensuring that the chosen guidelines align closely with the unique context, objectives, and stakeholders of the AI system under consideration. Therefore, designing a comprehensive model that addresses all aspects is a complex challenge. Similarly, navigating the many guidelines for building appropriate trust in AI systems can be overwhelming. Therefore, it is essential to consider the specific context, domain, and stakeholders involved. Different guidelines may have varying focuses, such as ethics, explainability, or fairness, so selecting the most relevant ones based on the specific requirements and goals of the AI system can help guide the implementation of appropriate trust measures.

### CONTINUUM OF TRUST

In order to achieve appropriate trust, one has to be able to recognize when it is not there to fix this. Therefore, studying the entire continuum of trust beyond its appropriate level, *i.e.* over-, under-, mis-, and dis-trust, is helpful in achieving it. For example, it can be possible to achieve calibrated trust through fostering both trust and distrust in AI at the same time [264]. Sensibly placed distrust makes users not agree with the opinion of others automatically, but rather increases their cognitive flexibility to trust appropriately [284]. Yet, only 14% of the reviewed papers look into this. The literature proposes terms like **calibration points** [245] or **critical states** [157] to classify the situations when the intervention for calibrating trust is needed. The former term is characterized as a way to classify situations in which the automation excels or situations in which the automation is degraded [245]. The later is characterized by the situations in which it is very important to take a certain action such as an autonomous vehicle detects a pedestrian [157]. In both of these situations, a mismatch can occur between levels of performances and expectations, which would allow users to reflect whether their trust levels are appropriate or not.

Generally, we find that the reviewed papers mostly rely on analyzing human behaviors to determine whether trust needs to be calibrated. For example, states of over- and under-trust are inferred from monitoring the user's reliance behavior rather than subjective trust measures [279]. Collins and Juvina propose to watch out for any behaviors that can be considered as exception out of principle of trust calibration (appropriately calibrated trust) to understand better long-term trust calibration in dynamic environments [74]. In their study with a multi-arm trust game, during critical states, users unexpectedly changed their trust strategy, tending to ignore the advice of the previously trusted AI advisors and leaning more towards the previously non-trusted ones. One of the unique findings from this work was that a) trust decays in the absence of evidence of trustworthiness or un-trustworthiness and b) perceived trust necessity and cognitive ability are important antecedents on the trustor's side to detect cues of trustworthiness.

The previous example teaches us that trust calibration is a complex process that requires a nuanced understanding of the context and user behavior, and that the ability to adapt and change trust strategies in response to changing situations is an important aspect of successful trust calibration. Similar to Collins and Juvina, Tang et al. [355] explicitly used distrust behaviors by leveraging data mining and machine learning techniques to model distrust with social media data. Distrust was conceptualized such that it can be a predictor of trust and of the extent to which it is mis-calibrated. Lastly, one paper relied on physiological markers such as gauge behaviour from a eye tracker coupled with the rate of reliance on AI and compared it with the system's capability to identify if trust is mis-calibrated [19].

In conclusion, there are various approaches adopted by the authors ranging from examining behavior and performance to studying distrust and trust mis-calibration for building appropriate trust. Authors have proposed over- and under-trust detection, calibration points, and critical states to study appropriate trust through the continuum. Furthermore, studies on distrust have shown that it can play a critical role in trust calibration, and trust mis-calibration can be used to understand long-term trust calibration in dynamic environments.

### 3.5.4 RESULTS OF CALIBRATION INTERVENTIONS

In this subsection, we provide a general overview of the findings of the reviewed papers. In particular, we focus on the results of applying the methods for building appropriate trust described in Section 6.3.

From the categories of methods described in this section, improving system transparency was the most common. Most papers supported the hypothesis that transparency facilitates appropriate trust in a system. For example, it was found that uncertainty ratings [365], confidence scores [416], providing explanations [24, 208, 272], and reliability and situational awareness updates [19] improved appropriate trust in a system. However, other papers add some nuance to this conclusion. For instance, Bansal et al. found that explanations increased the human's acceptance of an AI's recommendation, regardless of its correctness [24]. Furthermore, Wang & Yin found that only some of their tested explanations improved trust calibration, indicating that not all explanations are equal [394]. Lastly, though confidence scores can help calibrate people's trust in an AI model, Zhang et al. (2020) found that this largely depends on whether the human can bring in enough unique knowledge to complement the AI's errors [416]. These results highlight that further research is necessary to study exactly what methods of increasing transparency are useful to facilitate appropriate trust, given the context of the interaction. We believe opportunities lie in exploring how diverse factors such as user expertise, task complexity, and the type of explanation influence trust calibration. This could involve controlled experiments that manipulate different transparency elements to pinpoint their individual and combined effects on trust.

Improving system transparency had mixed results for building appropriate trust, and leveraging human cognition and perception for trust calibration yielded the similar results. For example, Riegelsberger et al. found that changes in how a system interacts with the user impacted users' perception of trustworthiness. [310]. Similarly changing the interaction with the system, Bućinca et al. found that cognitive forcing functions<sup>7</sup> reduced over-reliance on AI. However, the performance of human+AI teams was worse than the AI alone with these functions [44]. Other than the use of cognitive forcing functions to compel people to engage more thoughtfully with AI systems, Naiseh et al. found that nudging can also help users become more receptive and reflective of their decision possibly leading to appropriately trusting the AI system [269]. As nudging and cognitive forcing functions target cognitive and perceptual mechanisms for building appropriate trust, the effectiveness of training is also intricately linked to these mechanisms. For example, two studies showed that teams receiving the calibration training reported that their overall trust in the agent was more robust over time [176, 270]. Based on these findings, it is crucial to focus on developing interventions that promote analytical cognitive thinking to foster appropriate trust in AI systems.

The appearance of a system plays a significant role in shaping how humans perceive and mentally process its attributes, which in turn impacts their levels of trust in the system. For example, Jensen et al. discovered that a system with a more human-like appearance was perceived as more benevolent, but this did not lead to differences in trust in behavior leading to unsupported trust calibration [172]. Similarly, both Christoforakos et al. [69]

<sup>7</sup>Interventions implemented during decision-making to disrupt heuristic reasoning and prompt analytical thinking such as on-demand explanation or forced waiting for output [210].

and de Visser et al. [88] found that more human-like systems were considered more trustworthy, but didn't help in trust calibration. These results highlight that the human-likeness strategies for building appropriate trust have been challenging so far. Although it seems clear there is some effect of appearance on trust, how to use this properly to ensure the appropriateness of trust remains an open question.

So far we have looked at results of the trust calibration interventions related to improving system transparency and understanding human cognition and perception including human-likeness. Distinct from these methods, understanding the continuum of trust was also helpful in certain cases for building appropriate trust. For instance, calibration points and critical states prompted users to adjust their trust in the system by facilitating specific moments of engagements [157, 245]. Furthermore, detecting over- and under-trust was critical in providing trustworthiness cues to the user in calibrating their trust levels. However, the use of these cues was found to not necessarily improve the performance of the human-AI teams [279]. Finally, miscalibrated (i.e., over- or under-) trust [74] and distrust [201] were also promising to calibrate human trust in the system in certain situations such as under conditions of increased trust necessity. Miscalibration affected interactions with new trustors, as a reputation for past trustors preceded the entity, causing potential new trustors to approach with caution [201]. Therefore, understanding continuum of trust through user studies can help in building appropriate trust which can improve the human-AI team performance and helpful in trust repair. In particular, opportunities lie in conducting more empirical studies investigating trust development over time with different contexts and how this impacts human decision-making.

In summary, the methods applied in the selected papers yielded mixed results. On the one hand where improving system transparency and understanding human perception and cognition had an impact on appropriateness of trust but on the other hand it did not improve the human+AI joint performance. Similarly, studying the continuum of trust helped in fostering appropriate trust but it also failed to improve human-AI team performance as well as in repairing trust. Overall, it remains complicated to find one-size fits all solution for building appropriate trust in AI systems. Therefore, we recommend that future researchers give careful consideration to a) how they define appropriate trust, b) specify what do they mean by it, c) how they conceptualize their measures and d) avoid using related concepts in particular.



Figure 3.6: Overview of the different appropriate trust building methods adopted in the articles from our corpus.

## 3.6 DISCUSSION

In this systematic review, we have discussed the (a) history of appropriate trust, (b) difference and similarities in concepts related to appropriate trust, (c) a BIA mapping to understand commonalities and differences of related concepts, (d) different methods of developing appropriate trust, as well as (e) results of those methods. In this section, we reflect on our findings by providing critical insights on elaborating key challenges and open questions. Furthermore, we provide some novel perspectives on understanding appropriate trust and finally acknowledge the limitations of this work.

3

### 3.6.1 KEY CHALLENGES

With appropriate trust constituting a central variable to the appropriate adoption of AI systems, different approaches have been taken to understand it. Our aim with this study was to provide an overview of the field's current state. In doing so, we reflected on our findings and found some challenges that exist in our way of understanding this research area. In this sub-section, we elaborate on the aforementioned key challenges, how to overcome possible limitations and summarize critical points with research opportunities for future work. Our identified key challenges are:

1. Discord and diversity in concepts related to appropriate trust such as calibrated trust, justified trust, responsible trust etc.
2. A strong focus on appropriate trust in capability, leaving out other aspects of trust such as benevolence and integrity [182].
3. The issues involved in adequately measuring appropriate trust.

#### DISCORD AND DIVERSITY IN UNDERSTANDING APPROPRIATE TRUST

From the analysis of the reviewed definitions of appropriate trust, we identify 3 major challenges for the current theoretical discourse on the topic. Firstly, as seen in Section 5, there is no uniform understanding on what appropriate trust is: some papers define appropriate trust based on system performance or reliability [274, 279, 282, 386, 406], some relate it to trustworthiness and beliefs [80, 182] and some base it on calculations [76, 173, 394]. Such a variety of the appropriate trust definitions stems from different understanding of what "the right amount of trust" implies. The common denominators of having various definitions of appropriate trust can be linked to: (a) the context in which it is studied often differs from one study to another, (b) the multidimensional nature of trust, often associated with attitude or subjective beliefs, adds complexity to understanding appropriate trust, and (c) different academic fields approach the study of trust in unique ways, leading to divergent interpretations of appropriate trust. For example, in HRI domain trust is often linked to robot's performance [185] whereas in Psychology it is commonly linked to understanding social and interpersonal aspects [308].

In addition to the variety of definitions of appropriate trust, we also found that the literature proposes various related concepts<sup>8</sup> (See Table 3.1), sometimes used interchangeably in the discourse about appropriate trust [24, 279, 401]. For example, we would like

<sup>8</sup>From our understanding, a "concept" is a general idea representing a category, while a "definition" is a precise statement that clarifies the meaning of a term or concept.

to especially stress the difference between appropriate trust and another most used related construct - calibrated trust. Although the logical formulation of the two concepts is similar as shown in the BIA mapping in Figure 3.5, trust calibration requires a process. In contrast, appropriate trust is the maintained state of the calibrated trust over a series of interactions. This conceptual overlap raises questions about the precise boundaries and distinctions between these concepts and highlights the need for a more refined and standardized conceptual framework.

These challenges surrounding understanding appropriate trust emphasize the significance of shaping our research agenda in this domain. To address the need for consensus among researchers, in this work we proposed a framework that explicitly defines appropriate trust and its boundaries. Our framework consider multiple dimensions, such as system capability, trustworthiness, beliefs, and task requirements while accounting for contextual variations. Moreover, we made an attempt to clarify the relationships between appropriate trust and related concepts, establishing clear definitions and boundaries to facilitate meaningful discussions and avoid conceptual confusion. By addressing these challenges and shaping a coherent research agenda, we can advance our understanding of appropriate trust and its implications for various domains.

#### **PROMINENT FOCUS ON SYSTEM'S CAPABILITIES IN DEFINITIONS**

The majority of appropriate trust definitions or its related concepts focus on the capability or ability of an agent. Here, appropriate trust is the alignment between the perceived and actual capabilities of the agent by the human [225, 279, 406]. Much of previous research has looked at 'ability' as the core factor of establishing trust [182, 251], which bring the focus upon the engineering aspect of trustworthiness. However, we view trustworthiness as more than just ability. Our interpretation of trustworthiness can be enhanced when we not only focus upon agent capabilities but also on understanding other factors such as integrity and benevolence [242, 293] or process and purpose [216].

Hoffman et al. state that "a thorough understanding of both the psychological and engineering aspects of trust is necessary to develop an appropriate trust model" [149]. Our examination of the psychological aspects of trust in human-AI interaction has revealed a need for improvement in the existing literature regarding modeling the integrity and benevolence of an AI agent toward a human as highlighted by Ulfert et al. [369], Mehrotra et al. [256] and Jorge et al. [182]. Mayer et al. [242] propose that the effect of integrity on human trust will be most salient early in the relationship, before the development of meaningful benevolence *i.e.*, X has disposition to do good for Y [92]. Therefore, we pose that it is important to first investigate how humans perceive AI system's integrity and how to model this relationship for fostering appropriate trust in AI system. Then it becomes vital to study the effect of perceived benevolence on trust as it increases over time as the relationship between the parties develops [251]. Throughout, the perceived ability of the system remains important. However, we pose it is crucial to not forget these other factors in research on appropriate trust.

#### **ADEQUATELY MEASURING APPROPRIATE TRUST**

While analyzing our corpus, we encountered common issues with appropriate trust measurements identified by Miller [261]. These issues include the absence of risk and vulnerability elements in user studies; overlooking instances of under-trust; uncertainty

regarding the extent to which behavioral experiments can capture trust; the robustness of single/multiple-item questionnaires in capturing changes in trust levels over time; reliance on agreement/disagreement with model predictions without considering discrepancies in human goals; and the use of appropriate situational awareness as a proxy for trust.

First, we found some papers in our corpus [8, 76, 185, 270, 328] which had little or no element of risk in the task design. We posit that in a questionnaire, survey, or field study it is crucial that participants have experienced or currently experience vulnerability to the possibility of the AI system failing. Trust cannot exist without the element of risk, and participants must have a personal stake in the situation. Including risk and vulnerability factors allows researchers to evaluate the trustworthiness of systems or services accurately.

Second, we observed some articles focused on capturing over-trust in AI [44, 164, 174, 391, 401], however under-trust was often overlooked. We posit that calibrated trust requires equal consideration of both scenarios. Appropriate trust necessitates equal consideration of both over-trust and under-trust scenarios because a skewed focus on one aspect can lead to sub-optimal outcomes.

Third, it is not clear to what extent behavioral experiments which account for 70% of experiment designs, especially physiological & empirical measures, can be used as a proxy to capture trust. While behavioral experiments can offer valuable insights into trust-related behaviors, their ability to fully capture the complexity of trust can be unclear due to simplified environments, artificial motivations, lack of context, limited generalizability, and the subjective nature of trust [104].

Fourth, it is difficult to establish whether single/multiple-item questionnaires are robust enough to capture changes in trust levels over time [8, 74, 176, 191, 201]. Also, in almost 40% of studies trust is measured before and after the user study, though it is not always appropriate to reflect on users' attitude at such a high level of granularity. A focus on trust dynamics over time as indicated by some studies [8, 18, 136, 201] could be a better approach.

Fifth, measures of trust related to whether humans agree or disagree with a model prediction are employed in some studies [44, 225, 413, 416], however what happens when the model targets differ from human goals? Sixth, reliance was often used as a proxy for trust, or even treated as the same thing. As Tolmeijer et al. [364] highlighted trust in an agent as the belief that "an agent will help achieve an individual's goal in a situation characterized by uncertainty and vulnerability" [216], while reliance on AI is defined as "a discrete process of engaging or disengaging" [216] with the AI system. Finally, some authors [19, 176] acknowledge the ambiguity of using appropriate situational awareness as a proxy for measuring appropriate trust in their approach.

In this sub-section, we explored the discord and diversity in concepts related to appropriate trust, including calibrated trust, justified trust, and responsible trust. We also highlighted a prevalent focus on trust in capability, neglecting other important aspects like benevolence and integrity. We found a lack of clear understanding of appropriate trust and identified issues in assessing it. Finally, we have yet to completely characterize how to measure appropriate trust adequately. For example, there is more work to do to fully understand the element of risk or vulnerability, to have a clear distinction between reliance and trust, and an uneven focus on both over- and under-trust.

### 3.6.2 OPEN QUESTIONS

While analyzing the text from our corpus, we discovered some open questions on determining whether appropriate trust in AI systems is achieved. First, what to take into account when deciding whether human's trust in the AI system is over-trust or under-trust? From the reviewed articles, this distinction seems to be primarily based on the AI accuracy, *i.e.*, correct or incorrect AI recommendations[406, 413, 416].

We argue the process of determining where the threshold lies in deciding over- or under-trust can not be solely about making a right or wrong decision; instead it should consider multiple aspects. For example, while accuracy indicates human reliance on the AI system's outputs, it does not capture the nuanced nature of trust. Trust involves more than mere reliance; it encompasses perceived reliability, multiple interactions, transparency, and the belief that the AI system has the user's best interests. For instance, a user may rely on an AI-based navigation system when using it for the first time to reach their destination, leading to 100% reliance. However, trusting the system 100% may require interacting with it multiple times in different contexts. Hence, we argue that a comprehensive evaluation of trust should consider a multidimensional approach that incorporates both accuracy and factors related to transparency, interpretability, adaptability, longitudinal interactions, user feedback, and the cognitive and emotional aspects of trust. This broader perspective will enable researchers to understand better when human trust in an AI system gears towards over-trust or under-trust [276].

Second, how to calculate appropriate trust for a task with non-binary decision-making? *i.e.*, when the decisions are non-binary (e.g., price estimation) it is relatively difficult to identify over- and under-trust at regular time intervals. This could be because it involves a continuous scale of possibilities, making it challenging to define clear boundaries for what constitutes over-trust or under-trust. However, when the decisions are binary it is easier to assess trust since one can directly compare the outcomes to the binary decisions (e.g., correct or incorrect). In our analysis, we could not find any articles from the reviewed corpus that clarify how to calculate appropriate trust if the decisions are non-binary. We believe in such cases, it is essential to consider a more nuanced approach that takes into account the specific characteristics of the task and the decision-making process such as by assigning probabilities to different outcomes or decision options.

Third, and relating to the previous point, as AI systems can change over time, so how can we measure appropriate trust, or even reliance, as they becoming moving targets? Consider the automated vehicle which is highly reliable in dry, clean, weather but whose performance degrades in rainy conditions, forcing the driver to dynamically adjust their trust. We only find mention of this limitation in five of the articles we reviewed. Further, we could not find reviewed articles addressing how periodicity in the trust gain and loss is affected by the task *i.e.*, frequency or regularity with which trust is gained or lost in a task, thus we have limited understanding of trust dynamics in real-world long-term interactions.

We postulate that a common reason why we couldn't find articles relating to periodicity of trust is because dynamics of trust development and erosion is itself a complex topic which can impact task performance and efficiency. Hence, we need further research on generating empirical evidence, insights, and theoretical frameworks to address the gap in knowledge regarding the influence of task frequency and regularity on the periodicity of trust gain and loss.

### 3.6.3 NOVEL PERSPECTIVES

We found some distinct perspectives on understanding appropriate trust in AI while analyzing our corpus. First, Chiou and Lee argue that the current approach to studying trust calibration neglects relational aspects of increasingly capable automation and system-level outcomes, such as cooperation and resilience [66]. They adopt a relational framing of trust to address these limitations based on the decision situation, semiotics, interaction sequence, and strategy. They stress that the goal is not to maximize or even calibrate trust, but to support a process of trusting through automation responsiveness. We resonate with the perspective put forward by the authors; however, to achieve a higher degree of automation responsiveness, human values, societal norms, and conflicts are to be studied and implied in the AI systems.

Second, Toreini et al. suggest that we need to study the locus of trust to understand appropriate trust in the systems [366]. They raise the questions such as whether we trust the people who developed the system or the system itself. What purpose are the broader organizations serving? Furthermore, the authors acknowledge the limitations of individuals' capabilities concerning assessing ability and benevolence and propose that individuals accomplish this indirectly by assessing the ability and benevolence of the entity developing the AI. Finally, among the enormous amount of methods and approaches presented in the review, the work by Collins and Juvina highlights the importance of trust mis-calibrations to study appropriate trust [74]. According to the authors, when the need for trust becomes stronger, individuals may stop trusting their previous trusted partners and instead try to establish trust with those they previously distrusted. Studying these exceptions to the principle of trust calibration might be critical for understanding long-term trust calibration in dynamic environments. We believe that this change in trust tactics which is known in human-human interaction is missing in the human-AI interaction studies. Furthermore, we couldn't find any studies in which humans interact with several AI systems in real life, so this aspect of trust strategies needs to be studied if we wish to learn about how trust mis-calibration can be a useful tool to understand appropriate trust in AI systems.

### 3.6.4 LIMITATIONS

Despite the systematic review's comprehensive analysis of the state of the art in fostering appropriate trust, there are several limitations to this study that need to be acknowledged.

First, while we included studies from a number of relevant disciplines (refer to our search string in section 3.3.1), it is possible that some relevant studies were missed. Additionally, we only focused on studies published in English, which may have led to language bias. Future reviews should consider including studies in other languages to increase the generalizability of the findings.

Second, our mapping to concepts related to appropriate trust based on beliefs, desires, and intention is only one of many possible ways to organize such concepts under an umbrella. As such, future research can focus on the development of a clear and concise mapping of these definitions from a multidisciplinary perspective.

Third, our search period only included papers from 2012 till June 2022 and the research on appropriate trust is growing at a faster pace. Therefore, papers which were published from June 2022 are missing from this review. Finally, the current review only focused on

the current state of the art in fostering appropriate trust in AI systems. While the review identified potential research gaps and opportunities, additional research is necessary to develop new approaches and design techniques to better understand the topic.

### 3.6.5 SUMMARY

This sub-section aims to summarize the current trends, challenges, and recommendations concerning the definitions, conceptualizations, measures, implications of measures, and results for establishing appropriate trust in AI systems. By addressing the evolving trends, inherent challenges, and potential solutions, we aim to enrich the overall understanding of the topic, enabling readers to grasp the broader context and implications associated with building appropriate trust in AI systems.

Our aim with this summary is to provide a well-structured gateway for both experts and newcomers to understand the trends and challenges with an actionable set of recommendations. With these recommendations we make an attempt to connect all the sections of this chapter to provide broader context and implications of building appropriate trust in AI.

| Section            | Current Trends   | Challenges  | Recommendations   |
|--------------------|--|---|---|
| <b>Definitions</b> | (1) 75.3% (n = 312) of articles from our corpus which were sought for retrieval did not provide a definition of appropriate trust or a related concept <sup>9</sup> .        | (1) A lack of clear definition creates a confusion among readers from different backgrounds.  | (1) Provide a clear definition of appropriate trust or a related concept.   |
|                    | (2) Of the articles, which provided a definition in our final corpus, 25% (n = 16) of them provided new definitions which were often not related prior works, see Table 3.1. | (2) A variety of definitions inherent to multidisciplinary fields without relating it to other fields can cause misunderstanding to the reader. | (2) We need to converge in the future to establish common ground to define <i>what appropriate trust means in human-AI interaction?</i> |

<sup>9</sup>Italics is for supplementing the information.

### Conceptual-ization

3

(1) Many types of appropriate trust concepts are only sometimes explicitly distinguished.

*For example, the differences between optimal trust, well-placed trust, meaningful trust, justified trust etc., are often unclear and used interchangeably.*

(2) Interchangeable use of Appropriate Trust with Appropriate Reliance

(3) 38% of articles in our final corpus conceptualize appropriate trust or related concepts as the measure of alignment between the perceived and actual ability of the system.

(1) A plethora of concepts related to appropriate trust is causing the HCI community to diverge in multiple ways.

*This unclear connotation of similar concepts often creates confusion among researchers, especially new graduate students.*

(2) A core distinction in philosophy, which is often neglected in the empirical HCI literature, regards trust and reliance as distinct concepts.

(3) To explore the extent and magnitude of how the trustworthiness properties of machines, beyond their ability, impact trust. For example, what do we mean by integrity of a machine, and how can we measure it?

(1) Related concepts which are distinct from the goal of appropriate trust should be defined, distinguished, treated and measured as independent concepts. *For example, warranted trust and contractual trust have different goals than appropriate trust.*

(2) We propose Hoff & Bashir distinction [148], where trust is the belief that “*an agent will help achieve an individual’s goal in a situation characterized by uncertainty and vulnerability*” and Lee and See’s reliance distinction [216] “*a discrete process of engaging or disengaging*”.

(3) We must focus on measuring less studied dimensions of trustworthiness, i.e., integrity and benevolence, to understand human trust in AI systems.

|                                    |   |  |  |
|------------------------------------|---|--|--|
| <b>Measures</b>                    | <p>(1) 40% (n=26) of articles in our final corpus study appropriate trust in binary decision making tasks <i>i.e.</i> to [not] follow an [in]correct AI recommendation.</p> | <p>(1) To develop strategies for building appropriate trust in AI systems that continuously make decisions, such as in price estimation. Also, the potential issues that arise when the AI model targets diverge from human goals.</p> | <p>(1) We need to investigate new measures to assess dynamic trust in practice. For example, we can use situational reference points to keep aligning the goal [60].</p>                       |
| <b>Results</b>                     | <p>(1) Around 37% of reviewed articles report the effect of improving system transparency for establishing appropriate trust in human-AI interaction.</p>                   | <p>(1) A disadvantage of single focus on improving system transparency requires ground truth, which is often not available or there is no really ‘ground’ at all.</p>  | <p>(1) Include post-experiment surveys or interviews where the participants can give their impressions on the trustworthiness of the AI Agents.</p>  |
|                                    | <p>(2) In 43% of the included articles, the objective of the designed task had direct influence on the results of appropriate trust in human-AI interaction.</p>            | <p>(2) If the objective of the task to foster appropriate trust in the AI agent is built around improving the fairness of the AI agent then the results will be different compared to objective of improving the accuracy.</p>         | <p>(2) Ensure to control initial participant’s expectations about the AI system and report results with scientific rigor about how the design of the task may have influenced human trust.</p> |
| <b>Implication of the measures</b> | <p>(1) 45% of articles involving a user-study focused on detecting over trust in AI, under trust in AI systems is often overlooked.</p>                                     | <p>(1) Under-trust in AI systems is a common challenge.</p>  | <p>(1) Investigate and adopt methodologies from social sciences and psychology to study under trust in AI [184].</p>   |

|   |   |  |
|---|---|--|
| <p>(2) Around 10% of articles in our corpus follow some already established guidelines to design for fostering appropriate trust.</p> | <p>(2) There are multiple guidelines from academia and industrial organizations outlining trust calibration principles that AI-based systems should adopt. However, there is less effort that has been put in translating those principles into practice.</p> | <p>(2) Adopt established guidelines while designing an user study and report if those guidelines did not scaled for the user study.</p>  |
| <p>(3) Locus of trust in the AI systems: are we trusting the people who developed the system is unexplored.</p>                       | <p>(3) Identify and explore the fundamental correlations between appropriate trust in AI systems and the manufacturers of AI.</p>   | <p>(3) Adopt Toreini et al. [366] recommendation on analyzing factors such as the transparency of the AI development process, the track record of the manufacturer in delivering trustworthy AI, and the level of accountability and responsibility taken by the manufacturer for the AI's outcomes.</p> |

Table 3.3: A detailed summary of current trends, challenges and recommendations based on the results of the systematic review.

### 3.7 CONCLUSION

Appropriate trust in AI systems is crucial for effective collaboration between humans and AI systems. Various approaches have been taken to build and assess appropriate trust in AI systems in the past. This chapter provides a comprehensive understanding of the field with a systematic review outlining different definitions of appropriate trust, methods to achieve it, results of those methods, and a detailed discussion on challenges and future considerations. Through this review of current practices in building appropriate trust, we have identified the challenge for a single definition of appropriate trust and the ambiguity surrounding related concepts such as warranted trust, appropriate reliance, or justified trust.

Our review has proposed a Belief, Intentions, and Actions (BIA) mapping to study commonalities and differences among different concepts related to appropriate trust. We found three common measurement techniques to measure appropriate trust as Perceived,

Demonstrated and Mixed. In addition, multiple domains and associated tasks have been used to study appropriate trust. Furthermore, our analysis of articles revealed four common methods for building appropriate trust such as transparency, perception, guidelines and studying the continuum of trust.

In synopsis, the review highlights what approaches exist to build appropriate trust and how successful they seem to be. We have discussed the challenges and potential gaps in studying appropriate trust, which presents opportunities for future research such as discord & diversity in defining appropriate trust or a strong focus on capability. Overall, this chapter provides (a) a comprehensive overview of the current state of research on appropriate trust in AI by studying measures, tasks, methods, and results of those methods, (b) a BIA mapping of appropriate trust and its related concepts, and (c) a set of recommendations for fostering appropriate trust in AI based on current trends and challenges. With these contributions, we can advance our understanding of designing for appropriate trust in Human-AI interaction taking a step closer towards Responsible AI [163].



# II

3

## SOCIAL LENS



## 4

# EFFECT OF VALUE SIMILARITY ON TRUST IN HUMAN-AGENT INTERACTION

4

*As AI systems are increasingly involved in decision making, it also becomes important that they elicit appropriate levels of trust from their users. To achieve this, it is first important to understand which factors influence trust in AI. We identify that a research gap exists regarding the role of personal values in trust in AI. Therefore, this chapter studies how human and agent Value Similarity (VS) influences a human's trust in that agent. To explore this, 89 participants teamed up with five different agents, which were designed with varying levels of value similarity to that of the participants. In a within-subjects, scenario-based experiment, agents gave suggestions on what to do when entering the building to save a hostage. We analyzed the agent's scores on subjective value similarity, trust and qualitative data from open-ended questions. Our results show that agents rated as having more similar values also scored higher on trust, indicating a positive effect between the two. With this result, we add to the existing understanding of human-agent trust by providing insight into the role of value-similarity.*

## 4

## 4.1 INTRODUCTION

In the Indian epic Mahabharata, Arjun and Bhima are important characters. They go through common struggles and trust in each other's abilities. They challenged Jarasandha and Chitrasena's (*two kings*) armies and fought for Kampilya together (*a capital kingdom*). What made them have so much trust in each other? According to Rajagopalachari [305], the most compelling reason was that they shared similar values. In this chapter, we explore how we can take inspiration from this story when trying to understand trust in AI.

As AI systems gain complexity and become more pervasive, it becomes crucial for them to elicit appropriate trust from humans. We should avoid under-trust, as it would mean not making optimal use of AI. Yet we should also avoid over trust, as relying on AI systems too much could have serious consequences [294]. As a first step towards eliciting appropriate trust, we need to understand what factors influence trust in AI agents. Despite the growing attention in research on trust in AI agents, a lot is still unknown about people's perceptions of trust in AI agents [129]. Therefore, we wish to know what it is that makes people trust or distrust AI? In this chapter, we see trust as multi-dimensional as suggested by Roff and Danks [315]. On the one hand, trust corresponds to reliability and/or predictability and on the other hand trust depends upon people's values, preferences, expectations, constraints, and beliefs. Various studies have examined how trust is attributed according to the first dimension [59, 318], but fewer have investigated the second dimension, where the focus is on people's shared values [77]. The implication of the latter dimension for the design of agents is on how to design these agents with respect to values as different people prioritize different values, which in turn guides how people behave and judge the behavior of others [121].

We argue that there is a research gap in understanding the role of values on the trust a human has in that agent. Siegrist et al. state [341]:

*"people base their trust judgments on whether they feel that the agency shares similar goals, thoughts, values, and opinions"*

For example, if you value *cost-efficiency* over *aesthetics* when it comes to buildings,

you would probably trust an architect more if they have shown that *cost-efficiency* is also important to them. Regarding trust in AI systems, we resonate with Tolmeijer et al. [361] in observing the potential for overlap and contrast with the psychology, ethics, and pragmatics of trust between humans. Based on this, we hypothesize that the trust of humans in AI agents is positively correlated to the similarity of the values of those agents and humans. Taking this approach forward in AI agent research, we examine the effect of (dis)-similarity (of human & agent's values) on a human's trust in that agent. We design five different agents with varying value profiles so that for any human, some of these are more similar and some less similar to the value profile of that human. The agents team up with participants for a risk-taking task scenario for which they have to interact and decide on the appropriate action to take. Participants evaluate the agents based on how much they trust each agent and their perceived Value Similarity (VS).

In the remainder of this chapter, we first review related work on value similarity and give an overview of existing literature on the use of values to promote trust. We then describe the design of the agents we use in the experiments, and the setup of our user study. We discuss our results and conclude with potential applications and limitations of our work.

## 4.2 RELATED WORK

Trust within the AI domain has been explored mostly in contexts such as decision making [339], examining/assessing user's trust [278], and improving the system performance [297]. We argue that it is important to also consider the similarity of personal values when researching trust. But can an AI agent have personal values? Increasingly, researchers are trying to incorporate values in AI systems, especially systems which are in some way involved in (helping humans with) decision making.

Winikoff explains value-based reasoning to be an desirable property for having appropriate human trust in autonomous systems [400]. This thought echoes with prior work by Banavar [137], van Riemsdijk et al. [375] and Mercur et al. [259]. More recently, Cohen et al. acknowledge [73]:

*“Human users will be disappointed if the AI system makes no effort to represent or reason about inherent social values that users would like to see reflected.”*

Most practical work on implementing human values in AI system focuses on plan selection [77], user-agent value alignment [334] and studying agent's value driven behaviour [90]. One of the earlier attempts to look at the effect of similarity of values on trust was made within social science research by Siegrist et al. [341]. They showed similar values, and trust depends upon each other in human-human interaction. Their findings resonated with Sitkin and Roth [343] who report that interpersonal trust is based on shared values. On these lines, Vaske et al. showed that as salient value similarity increases, social trust in the agency increases [378]. Their findings showed how understanding the value similarity between Colorado residents and United States department of agriculture, resulted in social trust and attitudes towards wildland fire management.

Recently, researchers have been interested in using this concept of value similarity

for AI systems as well. Cruciani et al. designed an agent based model showing how similarity in values can be a successful driver for cooperation in the regulation and design of public policies [78]. They analyze their simulation experiment by looking at how and, how much agents cooperate with similar others. The key takeaway message is the introduction of value similarity for investigating what ultimately motivates trust-building processes. However, their work used predetermined memory coefficient for simulation agents to study coordination and was not validated with human participants. Additionally, Chhogyal and colleagues designed a formal trust assessment model [65]. In their work, they developed value-based trust assessment functions and showed how they lead to trust sequences. However, they did not consider value preferences and neither validated the model with human participants. Building on these works, our research is looking for a deeper understanding regarding the effect of value similarity on trust in a risk taking scenario accounting for the perception of human participants instead of providing simulation based results.

### 4.3 METHOD

The primary goal of our study is to understand how (perceived) value similarity affects trust. We focused on exploring how users' trust is affected by interaction with different agents with varying value similarity. More specifically, we have the following hypothesis:

**Value similarity** between the user and the agent **positively** affects the trust a user has in that agent.

#### 4.3.1 CREATION OF VALUE PROFILES

We used the Schwartz Portrait Value Questionnaire (PVQ) [331] to draw each participant's user profiles which consist of ten value dimensions. There are statements about each value dimension in the PVQ. Participants were asked to read carefully and respond to how each statement resonate with them as a person on a scale of 1-6, where '1' means '*very much like me*' and '6' implies '*not at all like me*'.

For each '*very much like me*' we assigned a score of 1 and for each '*not at all like me*' a score of 6 to that value. Furthermore, we created an actual value profile for each user based on their rank<sup>1</sup> (refer column 'PVQ Score' in table 1). We combined the first two values according to rank as group one, the second two values as group two, and so on till group five. We grouped ten values into five groups with two values each. Sometimes, a group can have more than two values because multiple values could receive the same final score. To resolve this conflict, we employ Algorithm 1 (see Appendix 1) to get user priority. For example, in table 1, there are three values with a score of 0.9 (refer set C1); and we needed only two values for each group. Therefore, participants were asked to choose one value over another based on the meaning of two values (refer Figure 4.1) following algorithm 1. In our user-study, we did not come across a conflict case where there were more than four values with the same PVQ score.

<sup>1</sup>We define rank as a position in the hierarchy of importance of the values.

Table 4.1: An example of generating value profiles of agents based on human value profile. *Rank* represents order of the values, *PVQ* represents the PVQ scores by participants, *Corrected* represents scores after applying the algorithm 1 (See Appendix A). Lower scores corresponds to higher ranks. C1 showcases conflict between three values for group two. Group 1 (G1) - Group 5 (G5) are groups for the first two ranks, the second two ranks, and so on... representing five different agents.

| <i>Rank</i> | <i>PVQ Score</i> | <i>Corrected</i> | <i>Value</i>   |               |               |
|-------------|------------------|------------------|----------------|---------------|---------------|
| 1           | 1                | 1.0              | Security       | } Group 1: G1 |               |
| 2           | 1                | 1.0              | Self Direction |               |               |
| 3           | 2                | } C1 {           | 1.90           | Traditional   | } Group 2: G2 |
| 4           | 2                |                  | 1.95           | Conformity    |               |
| 5           | 2                |                  | 2.0            | Universalism  |               |
| 6           | 3                | 3.0              | Power          | } Group 3: G3 |               |
| 7           | 4                | 4.0              | Benevolence    | } Group 4: G4 |               |
| 8           | 4                | 4.0              | Hedonism       |               |               |
| 9           | 5                | 5.0              | Achievement    | } Group 5: G5 |               |
| 10          | 6                | 6.0              | Stimulation    |               |               |

### 4.3.2 AGENTS AND THE SCENARIO

We designed a “save a hostage game” in which each participant interacts with five different agents that provided tips and suggestions to save the hostage. The task was inspired by prior work from Wang et al. [392]. In our game, agents were featured with varying value profiles.

#### AGENTS AND VALUE SIMILARITY:

For each participant, we created five different agents with descending value similarity profiles from G1 to G5 (*see table 1 for example*). G1 is the agent who promotes the two top ranked values of the participant, G2 agent which promotes the values ranked 3 and 4, G3 promotes the values ranked 5 and 6, etc. (so the values that each agent promotes can differ for each participant depending on their PVQ outcome).

#### SCENARIO AND AGENT EXPLANATION:

We provided the following scenario to our participants in which they need to team up with AI agents to rescue a hostage: “A hostage is being held inside a building in a market place. The objective is to gather intelligence regarding the building. All five different AI agents are equipped with sensors, infrared cameras, and metal detectors. The AI agents can perform the security check of the building and inform you regarding any danger. You need to make a decision for the action to be taken based on the AI agent’s advice before you enter the building.”

We used the agent's names as A, B, C, D, and E mapping to G1-G5 in our user-study. Each agent provides a suggestion to the user based on their prior common knowledge and values that are of utmost importance. A piece of prior common knowledge for all the agents was *"I have searched the overall place and have found traces of the gun powder. I recommend that you take protective gear & armor shield with you"*.

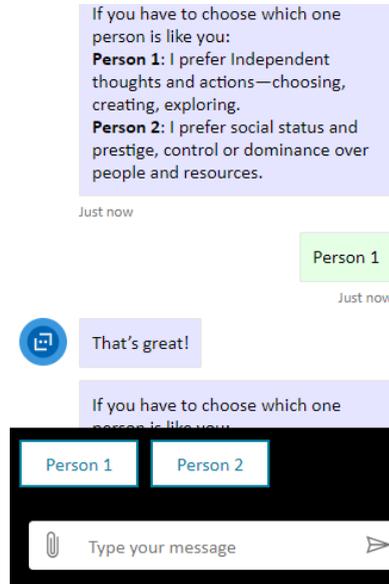


Figure 4.1: Human-AI agent interaction chatbot testbed with HTML front-end.

We designed our suggestions based on the values following the notion of situation vignettes in the work by Strackand and Gennerich [352]. The values were expressed through the suggestions the agent gave. Two researchers from Computer Science background and one from Cognitive Science background brainstormed together and generated sentences that formed suggestions by the agent. Overall, three iterations of each suggestion was performed to reach the final outcome.

For example, an agent provides the following suggestion based on prior knowledge plus their values from group one - security and self-direction: *"I have searched the overall place and have found traces of gun powder. I recommend that you take protective gear & an armor shield with you. For any action you take, do follow social orders & protocols. You should hand over the kidnapper to the police to abide by the national security laws. However, it's up to you what equipment you want to take inside the building & how you wish to deal with the situation."*

### 4.3.3 PARTICIPANTS

We estimated our sample size with the G-Power tool from Faul et al. [108]. Our effect size was 0.30 (*medium*) [122] with linear regression as our choice for modelling variables. G-power calculated our required sample size of 81. We recruited 101 participants from the different universities' mailing list. Twelve participants could not pass our attention

check, leaving 89 participants aged between 22 and 32 years old ( $M = 25.6$ ;  $SD = 0.94$ ). Each participant signed an informed consent form before the user-study. This study was approved by the ethics committee of our institution, ID number 1313.

We asked our participants to provide their cultural backgrounds before starting the user-study. Most of our participants were from the Europe region (34), followed by Asia Pacific (29), Americas (13), Middle East and Africa (9), and Oceania (2). Two participants did not provide their background.

#### 4.3.4 USER STUDY TEST BED

We implemented an online version of our scenario to study the impact of manipulating value similarity on trust. The test bed consists of a chatbot application that can be accessed from a web browser (see figure 1). We used Microsoft Power Apps API <sup>2</sup> to generate suggestions by the agents. These were displayed on the participant's chatbot interface, which sends data back to the test bed server. The user study test bed can be found at <https://edu.nl/buyqj>.

#### 4.3.5 PROCEDURE

Each participant first read an information sheet about the study and then fill out the background survey. Next, participants were asked to complete the PVQ to get their value profiles. After filling the PVQ, the system checked for any conflicts in value groups and asked the participant to choose one over another. Following this, the scenario was introduced to the participant.

All five agents interacted with the participant one by one. The order of appearance of the agents was randomly assigned in such a way that the order was different for each participant. Each agent appears with a small greeting and provides their suggestion. After each agent gave the suggestion, the participant was asked to fill questions from the Value Similarity Questionnaire (VSQ) [341] and questions from the Human-Computer Trust Scale (HCTS) [134]. In HCTS, trust is divided into three attributes, namely: general trust, benevolence, and willingness. The study was designed to be completed in 30 minutes. Participants were given a chance to participate in a raffle worth 5x20 Euro.

## 4.4 RESULTS

We analyzed the results of our study, including both the subjective rating responses to the value similarity, the trust questionnaire and, the explanations provided by the participants for selecting an agent. We were primarily interested in the effect of value similarity on trust. Thus, for this chapter, we focus on understanding the effect on trust by manipulating the value similarity. We call VSQ responses from participants as subjective value similarity. As part of our analysis, we first ran a Shapiro-Wilks test for normality. Since the distribution was not normal, we used non-parametric tests for our analysis.

### 4.4.1 MANIPULATION CHECK

We tried to manipulate value similarity in this study. However, to check whether our most 'similar' to least 'similar' agent were actually perceived as most and least similar,

<sup>2</sup><https://powerapps.microsoft.com/en-us/>

we also measured subjective value similarity. From figure 4.2, we see that the 'G2' agent scored higher than the 'G1' agent,  $\chi_r^2 = 11.725$ ,  $p < .05$ . This was in contradiction with the manipulation that we performed. In an ideal case, we expect the VSQ ratings to follow the order as G1 agent receives the highest VS score and G5 the least. This showcases that our manipulation did not work as expected. Considering this, we now only focus upon value

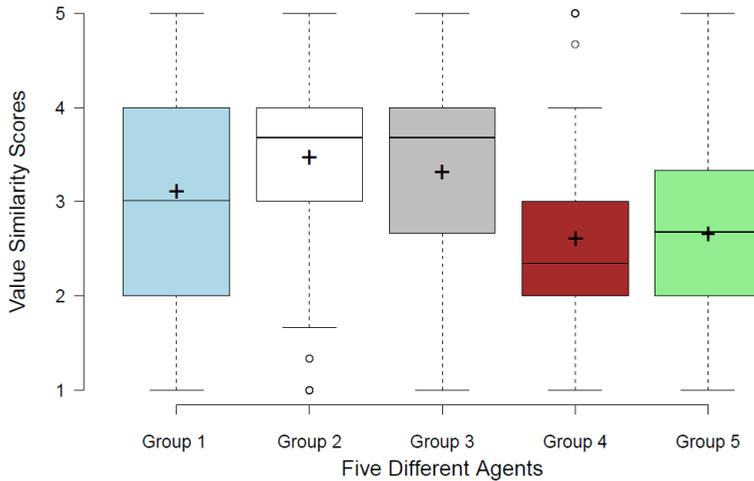


Figure 4.2: Mean subjective VS scores for all VSQ given by participants for the five agents. The horizontal line indicates the median and the plus sign the mean value for VS scores.

similarity as a whole rather than distribution /categorization of five agents. Therefore, in the rest of the paper we disregard our categorization of the agents.

#### 4.4.2 CORRELATION BETWEEN VALUE SIMILARITY AND OVERALL TRUST

We analyzed responses for the VSQ and HCTS to see to what extent subjective value similarity has an affect on trust.

A Kendall rank correlation test revealed that VS and trust are significantly moderately correlated in accordance with Ratner [307] with a correlation coefficient of  $0.46$  and  $p < 0.05$ .

We also applied a simple linear regression model to predict a quantitative outcome of trust based on a single predictor variable *i.e.* value similarity. To check linear model assumptions, we used the 'GVLMA' - Global Validation of Linear Models Assumptions [298] which provides a testing suite for many of the assumptions of general linear models. The four assumptions: normality, heteroscedasticity, linearity and, uncorrelatedness of the model were acceptable by the GVLMA. Linear regression showed that both the p-values for the intercept and the predictor variable were highly significant indicating a significant association between the variables. Our goodness-of-fit measures showcase  $\sigma$

= 0.984 meaning that the observed trust values deviate from the true regression line by approximately 0.984 units on average on a scale from one to five and  $r^2$  was 0.308.

Finally, to seek an answer to the problem: “can we predict trust from VS?”, we need to look at the intercept and residuals of the linear regression. On observing the intercept and residuals we have good reason to believe an overall effect of value similarity on trust. This confirms how closely VS and trust are related. Additionally, we wished to check if differences in cultural background of participants affected the effect of VS on trust. However, because our sample size was very diverse there were not enough participants from any distinct cultural background for a statistical comparison between them. Such an effect is potentially important, but future work would need to be done to test for this.

#### 4.4.3 BENEVOLENCE, AND WILLINGNESS AS ATTRIBUTES OF OVERALL TRUST

We examined the results of HCTQ as attributes of trust namely benevolence, willingness and general trust on value similarity. We already reported the results of the general trust in previous sections. Now, we focus ourselves to Benevolence and Willingness. A Kendall tau correlation was performed to determine the relationship between benevolence, willingness and value similarity. There was a medium, positive correlation between benevolence and value similarity, which was statistically significant ( $r = .47$ ,  $n = 436$ ,  $p = .0002$ ). Similarly, for willingness, correlation was found to be positive ( $r = .37$ ,  $n = 436$ ,  $p = .0002$ ).

#### 4.4.4 QUALITATIVE DATA ANALYSIS

We were interested in understanding which agents participants preferred the most. For this, we asked them to choose an agent to take with them inside the building and were asked to explain their reasons for doing so. We analyzed participants responses for selecting an agent. Our results indicate that the participants pick that agent that shares the most similar values. In figure 4.3, we can observe that more than 72% of participants chose the agent they ranked highest on value similarity and trust. This gives us another impression that subjective value similarity and trust correlate with each other. Now, we classified participants qualitative explanations into four themes found by thematic analysis [40]. This classification provides insight into the reasons for participant’s choices and how it translates to actual behaviour for selecting an agent. The four themes for selecting one agent over others are:

1. Common Values - the selected agent had more values in common with the human than the other agents.
2. Balanced Advice - the selected agent provided more balanced advice to the human participant than the other agents *i.e.*, balance of taking risk v/s following protocol
3. Developed Trust - the selected agent’s advice/suggestion inclined the participant to trust the agent.
4. Participant’s Belief - the agent was selected based on its advice/suggestion; this decision was neither related to values nor developed trust but more as human selection, based on developed belief because of presented explanation.

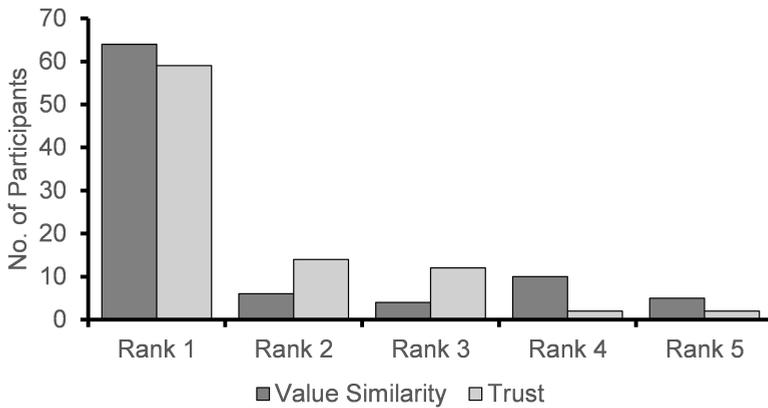


Figure 4.3: This figure represents the number of participants who choose an agent to take inside the building based upon their rank of value similarity and trust.

## 4

Out of a total of 89 participants only 55 provided an explanation for choosing an agent. Three researchers coded the explanations written by the participants. Each researcher performed the coding with three to four iterations before deciding upon final themes. Inter-coder reliability analysis was performed using Cohen's kappa to determine agreement and consistency between all coders. There was a near-perfect agreement among all three coders for three dimensions  $\kappa = .900$ , (95% CI, .643 to .937).

Based on our analysis, we found that 42% of the participants explanations were related to common values between the participant and the agent they chose. This was followed by 23% for balanced advice given by the agent, 16% for developed trust and 16% for belief of the participant. These results shows that in our experiment, VS and balanced advice promoted the intended behaviour of participants to select an agent. For example, P54 said, *'He [Agent A] thinks the same way as me [values] so I think he'd back me in my decisions'* relates to choosing an agent based on the common values. Similarly, P39 said, *'I believe agent B thinks 100% like me and gives me all the trust and responsibility'* relates to developed trust for the agent. We also came across many responses where participants choose the agent because of balanced advice by them. For example, P44 said, *'Agent B shows a balance of risk taking and following protocol to handle a delicate situation'*. Finally, few participants stick to their beliefs for their decision. This can be seen with what P27 reported as *'I believe he [Agent B] would be able to help save the hostages and neutralize the threat with non lethal force if possible and lethal if absolutely necessary'*.

## 4.5 DISCUSSION

In this section, we discuss the results of our study, relating them to prior work and making inferences on how the results can be applied to the design of AI agents. Recall that our main goal was to understand the effect of similarity of human & agent's values on a human's trust in that agent. Based on our study results, the hypothesis (that the VS between the user and the agent positively affects the trust a user has in that agent) can be partially accepted. We showed that there exists an overall significant effect of VS on trust. Even

though our failed manipulations did not interfere with our paper's primary goal, we were intrigued to find out that our manipulations of VS were not successful. In the following section, we discuss possible reasons for our unsuccessful manipulations.

#### 4.5.1 WHY WERE OUR MANIPULATIONS UNSUCCESSFUL?

If we wish to eventually promote appropriate trust, we should also be able to influence trust. To this end we need to know what factors influence trust, and we need to be able to manipulate these factors in the designs of agents. In this chapter we have added to the knowledge on factors that influence trust by showing the relationship with value similarity. However, the manipulation of those factors did not fully succeed in our study. Therefore, it is relevant to examine closer why our manipulations failed and provide some suggestions for how value similarity might be manipulated successfully in the future.

Regarding our specific agent design, a successful manipulation would have led to the observation that the 'G1' agent is rated highest for the perceived VS and the 'G5' agent the least. However, we observed that instead both the 'G2' and the 'G3' agent were rated as having more similar values than the 'G1' agent. To understand why this happened, we examined the actual value profiles of the participants more closely. Consider the case when VS scores of the 'G2' agent were higher than those of the 'G1' agent. Observing the participants' specific value profiles for who this occurred could provide us with potential reasons why manipulations were not successful. Figure 4.4 provides an overview of the values used in the explanations of the 'G1' and the 'G2' agents, and how often those occurred. This figure shows that the values of Self-Direction, Universalism & Achievement were most prominent for the agent 'G1' and Stimulation, Benevolence & Security for the agent 'G2', for those participants where 'G2' scored higher than 'G1' in value similarity. Given that people felt most similar to agents which promoted stimulation, benevolence and

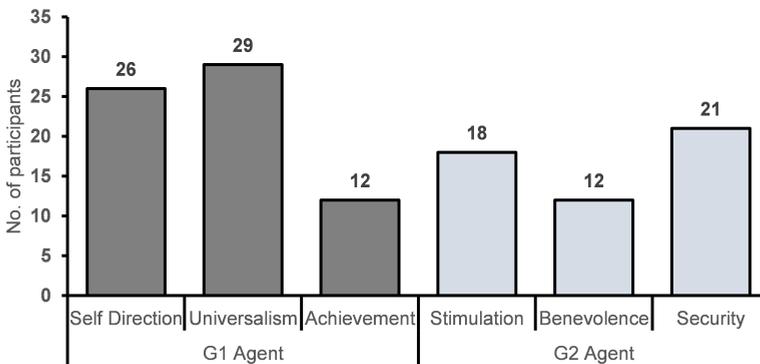


Figure 4.4: Top three most common values in the value profile of the G1 agent (values ranked 1 and 2 of participant) and the G2 agent (values ranked 3 and 4 of the participant). The numbers on the top of the histogram represent how often those values occurred for our participants where the VS scores for the G2 agent were higher than those of the G1 agent.

security (as opposed to the values of self-direction, universalism and achievement which scored higher in their value profile), we speculate that the choice of scenario might have played a role. The major values for agent 'G1' - Self Direction and Universalism, were those

which participants already possessed but were not so relevant in this context of saving a hostage. On the other hand, for agent 'G2' - Security and Stimulation were vital because they relate to safety and motivating the participant to save the hostage. It makes intuitive sense that contextual values are of utmost importance especially in those scenarios where there is a risk associated with trusting someone and not all the values are equally salient. However, the value profile survey is general, and not context-dependent.

Therefore, we speculate that when designing value profiles for artificial agents, one should not just take into account general value profiles, but also note which contextual values are most important as also echoed by Liscio et al. [224].

## 4

Another potential reason for our failed manipulation could be that a discrepancy existed regarding values of the agent in how they were perceived by some of the participants and how they were intended. By perceived values we mean that the value laden explanations that agents provided were sometimes interpreted as promoting different values than for which they were written. As explained in section 'Scenario and agent explanation', it took three iterations for each explanation to be finalized, which indicates how quickly disagreements about underlying values of explanations can occur. We speculate that this discrepancy is a possible reason for our failed manipulation and resound with Wang et al. [388] that designing agent explanations that can be consistently interpreted by humans is still an open research area. Secondly, consistency in value preferences from humans is debated, and people could just show inconsistencies as mentioned by Boyd et al. [37].

#### 4.5.2 TRUST IN AI SYSTEMS

In our user-study, agents provided their suggestions based on value-based reasoning using VS. With the use of value-based reasoning, an agent includes the representation for human values and can provide reasoning using human values to make decisions. Winikoff argue that a computational model of relevant human values can be used to provide higher level, human-centered explanations of decisions by AI agents. This means that agents could use value-based reasoning when trying to influence trust [400]. In summary, given a group of randomly generated agents, humans would trust those that align with their subjective values. The reported correlation can also come from human's consistent value judgment about the suggestions and scenarios.

Value Similarity is not the only thing that influences trust; many other factors can influence trust as well. Three of the main aspects of trust are benevolence, willingness and competence. Value similarity could be seen as a part of benevolence, or even willingness. However, competence is less related to values [403]. We did not focus upon this factor because we provided all our agents a ground truth *i.e.* prior common knowledge. Instead, we focused upon benevolence and willingness as other two factors of trust affected by VS in accordance with Gulati et al. [134]. Based on our results, both these factors were moderately positively correlated with VS. This implies that if we wish to understand how humans trust systems we need to look beyond trust as being influenced only by the system's reliability as reliability and values can be different things with different effects on trust. Rather, we also need to consider trust in benevolence and willingness and understand how these are influenced by aspects such as value similarity [73].

### 4.5.3 LIMITATIONS & FUTURE WORK

We investigated the effect of VS on trust with a risk-taking scenario of saving a hostage. We chose this scenario to gain a deeper understanding of how participants trust an agent with most to least VS. However, we believe that further evaluation with more real-life examples would provide additional insights on participant's trust. Additionally, although we cross-examined the participant value profile with their responses to the VS questionnaire, we did not focus on their understanding of value laden explanations. We posit that examining the perception of the values could have provided a more subtle effect of our manipulations. We see this as an opportunity to further extend our work into understanding the beliefs and perceptions of the participants for agents with varying VS. Also, future work could extend the proposed method to multiple scenarios with different context information. Additionally, crowdsourcing could be another way to generate explanations instead of pre-designing by experts or experimenters, especially in translating abstract values to specific descriptions or behaviors. Finally, as explained in the previous section, we would like to study the potential effect of culture on our findings.

## 4.6 CONCLUSION

Our study shows that value similarity between an agent and a human is positively related to how much that human trusts the agent. Based on this finding, we would encourage designers of explanation and feedback-giving agents to create agents that shows human values. An agent with similar values to the human will be trusted more which can be very important in any risk-taking scenario. Although a system without value-based reasoning may be easier to develop, the benefits of including VS are worth it, especially in trust-critical situations.



# 5

## INTEGRITY BASED EXPLANATIONS FOR FOSTERING APPROPRIATE TRUST IN AI

5

*Appropriate trust is an important component of the interaction between people and AI systems, in that ‘inappropriate’ trust can cause misuse, misuse or abuse of AI. To foster appropriate trust in AI, we need to understand how AI systems can elicit appropriate levels of trust from their users. Out of the aspects that influence trust, this chapter focuses on the effect of showing integrity. In particular, this chapter presents a study of how different integrity-based explanations made by an AI agent affect the appropriateness of trust of a human in that agent. To explore this, (1) we provide a formal definition to measure appropriate trust, (2) present a between-subject user study with 160 participants who collaborated with an AI agent in such a task. In the study, the AI agent assisted its human partner in estimating calories on a food plate by expressing its integrity through explanations focusing on either honesty, transparency or fairness. Our results show that (a) an agent who displays its integrity by being explicit about potential biases in data or algorithms achieved appropriate trust more often compared to being honest about capability or transparent about the decision-making process, and (b) subjective trust builds up and recovers better with honesty-like integrity explanations. Our results contribute to the design of agent-based AI systems that guide humans to appropriately trust them, a formal method to measure appropriate trust, and how to support humans in calibrating their trust in AI.*

## 5.1 INTRODUCTION

AI technologies are creating new opportunities to improve people’s lives worldwide, from healthcare to education to business. However, people do over-trust or under-trust these technologies occasionally [294, 314]. Under-trust can lead to under reliance and over-trust can lead to over-compliance which can negatively impact the task. Hence, for AI systems to reach their potential, people need to have *appropriate* levels of trust in these systems, not just trust. Although there are many ways to define appropriate trust [406], in this chapter we take this to mean that the trust a human has in a system needs to align with the actual trustworthiness of the system [114].

It has only been in recent years that have we found research on appropriate trust in AI systems [24, 361, 362, 406]. Appropriate trust is a complex topic as it requires consideration of the influence of context, the goal-related characteristics of the agent, and the cognitive processes that govern the development and erosion of trust [61]. In this work we aim to contribute by studying how explanations given by the AI which highlight different integrity-based principles (e.g. honesty, transparency, fairness) can influence trust and the appropriateness thereof.

Explainable AI (XAI) is meant to give insight into the AI’s internal model and decision-making [395] and has been shown to help users understand how the system works [52, 287]. Efforts to ensure that AI is trusted appropriately are often in the form of explanations [24, 226, 414]. Intuitively, this makes sense as understanding an AI system’s inner workings and decision-making should, in theory, also allow a user to understand better when to trust or not trust a system to perform a task. Many are focused on how the system works: what it can do and can not [226, 393]. This is done in many different ways, from highlighting essential features of a decision [394], contrasting what would have happened if something was different [312] or how confident the system is about its answer [417].

Typically, explanations are focused on giving information about a system’s *ability* to improve appropriate trust. However, literature on how humans trust typically sees trust

as more than a belief about ability. Therefore, it is helpful to expand our perspective on explanations as well. A useful starting point for understanding human trust is the ABI (*Ability, Benevolence, and Integrity*) model from the organizational context by Mayer et al. [242]. This model has been used extensively in modeling trust, such as by Lee & See [217], Hoffman et al. [149], and Wagner et al. [385]. It defines human trust as "A trusts B if A believes B will act in A's best interest and accept vulnerability to B's action" [242]. Moreover, it distinguishes three trustee characteristics that influence a trustor's trust: belief in ability, benevolence, and integrity.

Ability indicates the skills and competencies to do something. Benevolence is about a willingness to do good to a specific trustor. Integrity is defined as the trustor's perception that the trustee adheres to acceptable principles [242]. One of the extensively studied factors in trust research is the ability of the system [36, 58, 102, 157, 247, 370]. However, fewer studies have investigated the integrity and benevolence dimensions of trust [419]. Benevolence is a specific attachment and emotional connection between the trustor and trustee, which builds over time [242]. Human-agent interactions are often short-term, and the extent to which we form emotional connections needs to be clarified. Therefore, more work on long-term social connections between humans and AI might be necessary before fully understanding the role of benevolence in XAI and human-AI trust relationships.

Prior studies on integrity have linked it to conventional standards of morality - especially those of honesty and fairness [159, 246]. XAI can be regarded as a way to enhance system integrity *i.e.*, the system being honest about making decisions is a form of integrity. No matter the exact definition, it is clear that integrity is a concept that can play a role even in short-term interactions. Moreover, we follow Huberts in claiming that integrity is an essential concept for human-AI interaction [159]. By applying Olaf's principle<sup>1</sup>, integrity is a necessity and a mandatory requirement of being true to oneself & others [246]. This aligns with the notion that as AI is increasingly used to make autonomous decisions over time, the principles that underlie these decisions are highly relevant [1]. Furthermore, lack of integrity could cause issues of bias and deception that have already started to impact humankind [208].

Therefore, the question arises what the effect would be of explicitly mentioning principles related to integrity into XAI on appropriate trust of a user in the system. In human-human interactions, principles associated with integrity such as accountability, transparency and honesty have been suggested as important for appropriate trust [204]. The question arises whether XAI could explicitly use references to these principles in explanations, and how this would affect (the appropriateness of) trust in the system? More specifically, we consider three principles related to integrity to express through explanations:

1. Honesty about the system's capabilities and confidence.
2. Transparency about the process of decision making.
3. Fairness in terms of sharing what risks such as biases exist.

---

<sup>1</sup>McFall [246] describes Olaf's principle as "*An attitude essential to the notion of integrity is that there are some things that one is not prepared to do or some things one must do.*"

Honesty, transparency and fairness appear in various studies as common elements of integrity in HCI, HRI or human-AI interaction literature [31, 93, 170, 195, 197] (see section 5.3). Therefore, in this study, we propose to incorporate references to these principles of integrity in explanations, and posit the following research questions:

**RQ1:** How does the expression of different principles of integrity through explanation affect the appropriateness of human's trust in the AI agent?

**RQ2:** How does human trust in the AI agent change given these different expressions of integrity principles?

**RQ3:** How do these different expressions of integrity principles influence the human's decision making, and do people feel these explanations are useful in making a decision?

We conducted a user study with 160 participants where they were asked to estimate the calories of different food dishes based on an image of the food with the help of an AI agent. In our user study, the first research question focuses on how different expressions of principles related to integrity (hereafter referred to as 'conditions') in explanations can affect appropriate trust in human-AI interaction.

In this chapter, we study **RQ1** in the context of making an exclusive choice in the form of a decision to choose oneself or an agent to complete the calories estimation task. Moreover, to allow us to study this question, we formally define what it means for trust to be appropriate in this context. **RQ2** aims at understanding change in human trust in the AI agent over time under different expressions of integrity. Finally, **RQ3** helps in understanding the effect of expressions of integrity on human decision-making and the effectiveness of explanations. Additionally, we were interested in exploring possible effects of covariates such as propensity to trust.

**Contributions** Specially, our research contributes the following:

- 1: We present a measurable construct for appropriate trust in the context of a specific task by providing a formal definition.
- 2: We illustrate an approach for expressing integrity of the AI systems with explanations focusing on honesty, transparency and fairness.
- 3: By conducting an user-study with 160 participants aligned with our research questions, we show that how explanations can help in building appropriate human's trust in the AI system.

We believe our research holds significance for two main reasons. Firstly, before we can investigate methods to establish suitable trust, it is crucial to have a clear understanding of its meaning. Secondly, the potential for conveying integrity-related principles through explanations remains largely unexplored. Through our contributions, we aim to broaden our comprehension of fostering appropriate trust between humans and AI, which is vital for effective human-AI interaction [251].

## 5.2 APPROPRIATE TRUST

### 5.2.1 PRIOR WORK ON APPROPRIATE TRUST

To understand what exactly constitutes appropriate trust in Human-AI interaction, we need to understand how people trust each other *i.e.* interpersonal trust. Mayer et al. define trust as follows: A trusts B if A believes that B will act in A's best interest and accept vulnerability to B's action [242]. Noteworthy in this definition, and what we believe is a key to defining Human-AI trust, are notions of belief and risk. The interpersonal trust reduces this risk by enabling A's ability to anticipate B, where anticipation is A's belief that B will act in A's best interest. Following Hoffman and Lee & See [149, 217], we carry forward this definition of trust in human-AI interaction.

Recently, there has been rapid progress in studies focusing on building appropriate trust in AI [24, 100, 223, 367, 394, 417]. In a recent work by Yang et al. [406], appropriate trust is defined as the alignment between the perceived and actual performance of the system. This definition talks about the user's ability to rely on the system when correct and recognize when it is incorrect. Similarly, Jorritsma et al. relate appropriate trust to appropriate reliance on the system [183]. On a contrary, Tolmeijer et al. informs us that although both trust and reliance are related, they should be treated and measured as independent concepts [363]. The authors define trust as the belief that "an agent will help achieve an individual's goal in a situation characterized by uncertainty and vulnerability", while reliance is defined as "a discrete process of engaging or disengaging" based on Lee & See work [217]. We follow the similar distinction as proposed by Tolmeijer et al. [363] in our work, and see trust as a (subjective) belief, while we see reliance as an (objectively observable) behavior.

Recent works in exploring appropriate trust in human-AI interaction have looked at the role of system trustworthiness and social transparency. For example, Liao and Sundar emphasize the mediating role of information display on trust judgments, and that appropriate trust relies on effective communication of system trustworthiness [223]. On the other hand, Ehsan et al. shows that social transparency could support forming appropriate trust in human-AI interactions by embedding socio-organizational context into explaining AI-mediated decision-making [100]. Additionally, various works in human-robot interaction focus on providing end-users with an accurate mental model of a robot's capabilities for establishing an appropriate level of trust [89, 192, 282, 384]. We believe that in the above-mentioned prior works, the provided constructs of appropriate trust are limited. The majority of these works consider the system's ability or performance for defining appropriate trust. We would argue that there is more to appropriate trust than a correct belief in the ability of the system; such as the psychology of trust focusing on beliefs [221], mutualistic benevolence impacting trust [213], personal integrity requiring truth telling [246] and ethics of trust focusing on fairness [232] or even environment based factors including task & culture [340].

A substantial amount of literature in human-AI interaction focuses on calibrating human trust, which is the process of making trust more appropriate over time. For example, De Visser et al. defined trust calibration based on prior works by Cohen et al. and Lee & See [72, 217] as a process of updating the trust stance by aligning the perception of an actor's trustworthiness with its actual trustworthiness [89]. According to them calibrated trust is a function of perceived trustworthiness which helps in eliciting appropriateness of trust.

Okamura & Yamada proposed a framework for detecting inappropriate trust in a system with a behaviour-based approach [279]. Their framework detects over and under trust in the system by monitoring the user's reliance behavior. In a similar work by McGuirl & Sarter, the AI system provided system confidence information to improve trust calibration [247]. In the above-mentioned studies, the focus of the task was to calibrate human trust. These related works can be helpful to understand the appropriateness of trust as calibration is about the process which incorporates updating trust levels, and appropriate trust can be boolean per situation resulting from that update.

In other works by Mehrotra et al. & Winikoff [253, 400], it has been argued that AI systems' value-based reasoning can help achieve appropriate trust. Mehrotra et al. showed the effect of (dis)-similarity of human and agent's values on a human's trust which forms a part of appropriate trust [253]. According to Winikoff [400], value-based reasoning is a desirable prerequisite for human-AI interaction because (a) an AI system who is able to conduct reasoning using human values in order to make decisions could be used as a basis for providing higher level and more human-oriented support (b) having an explicit model of values can help in verifying AI system's behaviour, for example in system's reasoning and decision-making by taking ethical considerations into account. Building on these works, our research looks for a deeper understanding in evaluating appropriate trust in human-AI interaction by incorporating integrity based explanations where integrity in itself is a part of basic inherent human values.

## 5.2.2 OUR APPROACH ON APPROPRIATE TRUST - A FORMAL PERSPECTIVE

We are interested in the effect of integrity-based explanations on appropriate trust, so we need to first understand what exactly appropriate trust is, and what counts as over or under trust. Over trust is often related to over reliance on the system leading to misuse and under trust is related to under reliance on the system leading to disuse. Also, we define another trust category - inconsistency following Sadiku et al. [319] who states a famous anthropologist, Margaret Mead quote, "*What people say, what people do, and what they say they do are entirely different things*" on understanding psychological notions of human behaviour. Intuitively, inconsistency happens when people choose to rely on those they trust less, or vice versa.

The work described in the previous paragraph provides a conceptual understanding of appropriate trust which we build on. Most notably, we say appropriate trust occurs when a belief about trustworthiness matches with actual trustworthiness. We consider appropriate trust as a state which is either true or false, rather than looking at the whole calibration process. However, for our purposes we also require a practically measurable definition of trust on top of this conceptual understanding. Therefore, we propose a formal definition which tells us exactly in which situations trust is appropriate or not. Specifically, we consider appropriate trust from a specific angle in this chapter. Our definition does not try to give an all-encompassing definition of appropriate trust, but rather does so in the context of a specific type of task. Namely: our task involves an exclusive choice of who will perform the task, the agent or the human. This selection is motivated by prior works on choice behaviour by Israelsen & Ahmed and Okumara et al. [165, 279] and recent work by Miller [262]. During our task a user and an AI agent are working jointly. **The user**

**should select whether for a particular task they want to rely on the AI agent or do it themselves.** In this situation, we define trustworthiness of the agent as how well they perform this task.

In our definitions we use  $TW$  for describing trustworthiness. When discussing the trust of a human  $h$  in an AI agent  $a$  for a task  $t$ , we do not write  $T_h(a, t)$  but  $T_{(human \rightarrow agent)}$  dropping the  $t$  for the ease of reading. We then define:

$T_{(human \rightarrow agent)}$  = Trust of the human in the agent for accomplishing a task

$TW_{human}$  = (actual) Trustworthiness of the human for a task

$B_{human}(TW_{human})$  = Belief of the human regarding its own Trustworthiness for a task

$TW_{agent}$  = (actual) Trustworthiness of the agent for a task

$Selection_{human}$  = Selection by the human for the task. *i.e.* themselves or the agent

We define appropriate trust based on the action and the subjective opinion of the human, as well as the trustworthiness of both human and agent. Now we will describe our concepts with the help of above mentioned parameters:

- **Appropriate Trust:** (a) the human estimates that the AI agent is better at the task than the human, (b) also the actual  $TW$  of the AI agent is equal to or higher than the human's  $TW$  and (c) the human selects the AI agent for the task and *vice-versa* - equation 1 & 2. Here (a) is cognitive trust from the human, (b) is god's eye view of the  $TW$  (described in the next section) and (c) is human selection that could be based on observable behaviour, rationality or simply delegation of the responsibility.

$$[T_{(human \rightarrow agent)} > B_{human}(TW_{human})] \wedge [TW_{human} \leq TW_{agent}] \wedge Selection_{human} = agent \quad (1)$$

$$[T_{(human \rightarrow agent)} < B_{human}(TW_{human})] \wedge [TW_{human} \geq TW_{agent}] \wedge Selection_{human} = human \quad (2)$$

- **Over-trust in the agent:** the human estimates that the AI agent is better at the task than the human and selects the AI agent even though the actual  $TW$  of the AI agent is lower than the human's  $TW$ .

$$[T_{(human \rightarrow agent)} > B_{human}(TW_{human})] \wedge [TW_{human} > TW_{agent}] \wedge Selection_{human} = agent \quad (3)$$

- **Under-trust in the agent:** the human estimates that they are better at the task than the AI agent and select themselves even though the actual  $TW$  of the AI agent is higher than the human's  $TW$ .

$$[T_{(human \rightarrow agent)} < B_{human}(TW_{human})] \wedge [TW_{human} < TW_{agent}] \wedge Selection_{human} = human \quad (4)$$

There could be instances where one can trust someone more than themselves and still choose not to rely on them and vice versa. For example, we rarely doubt the efficacy of automatic shifting mechanisms of today's cars, yet some people still choose to manually shift for the pleasure of it. On the other hand, people might want to avoid responsibility by delegating to the other person even if they have higher trust in themselves. Therefore, we formulate two additional cases based as:

- **Inconsistency with a good outcome:** the human estimates that they are better at the task than the AI agent however, they select the agent for the task, and the actual TW of the AI agent is higher (or equal) than the human's TW and *vice-versa*.

$$[T_{(human \rightarrow agent)} < B_{human}(TW_{human})] \wedge [TW_{human} \leq TW_{agent}] \wedge Selection_{human} = agent \quad (5)$$

$$[T_{(human \rightarrow agent)} > B_{human}(TW_{human})] \wedge [TW_{human} \geq TW_{agent}] \wedge Selection_{human} = human \quad (6)$$

- **Inconsistency with a bad outcome:** the human estimates that the AI agent is better at the task than the human, however they select themselves, but the actual TW of the AI agent is higher than the human's TW and *vice-versa*.

$$[T_{(human \rightarrow agent)} > B_{human}(TW_{human})] \wedge [TW_{human} < TW_{agent}] \wedge Selection_{human} = human \quad (7)$$

$$[T_{(human \rightarrow agent)} < B_{human}(TW_{human})] \wedge [TW_{human} > TW_{agent}] \wedge Selection_{human} = agent \quad (8)$$

## 5

Our definitions are suited for our task requiring exclusive decision-making *i.e.*, tasks where one has to make a decision by either relying on oneself or the other party. We now summarize the cases mentioned above in the following table. In the Table 5.1, equation 1 represents two conditions where  $TW_{human} = TW_{agent}$  and  $TW_{human} < TW_{agent}$  keeping other comparisons same. A similar pattern follows for equations 2, 5 and 6.

Table 5.1: Categorization of the trust categories based on the equations 1 to 8

| Equation | Higher TW | Human trusts who? | Human selects | Trust Category               |
|----------|-----------|-------------------|---------------|------------------------------|
| 1        | AI Agent  | AI Agent          | AI Agent      | Appropriate                  |
| 1        | Equal     | AI Agent          | AI Agent      | Appropriate                  |
| 2        | Human     | Human             | Human         | Appropriate                  |
| 2        | Equal     | Human             | Human         | Appropriate                  |
| 3        | Human     | AI Agent          | AI Agent      | Over trust                   |
| 4        | AI Agent  | Human             | Human         | Under trust                  |
| 5        | AI Agent  | Human             | AI Agent      | Inconsistency (good outcome) |
| 5        | Equal     | Human             | AI Agent      | Inconsistency (good outcome) |
| 6        | Human     | AI Agent          | Human         | Inconsistency (good outcome) |
| 6        | Equal     | AI Agent          | Human         | Inconsistency (good outcome) |
| 7        | AI Agent  | AI Agent          | Human         | Inconsistency (bad outcome)  |
| 8        | Human     | Human             | AI Agent      | Inconsistency (bad outcome)  |

## 5.3 INTEGRITY

### 5.3.1 PRIOR WORK ON INTEGRITY

Mayer et al. states that “the relationship between integrity and trust involves the trustor’s perception that the trustee adheres to a set of principles that the trustor finds acceptable”

[242]. This definition of integrity is rooted in the studies on organisational management. However, the definitions of integrity vary across disciplines, but even within disciplines. For example, again in management science, according to Jeavons [169], integrity has to do with continuity between appearance and reality, intention and action, promise and performance, in every aspect of a person's or an organization's existence. Whereas Hon and J. E. Grunig [153] described integrity in public relations as "the belief that the other party is fair and just".

A literature review by Palanski [291] provides an overview of relevant integrity definitions in philosophy. The review outlines five general categories of integrity: wholeness, consistency of words and actions, consistency in adversity, being true to one-self, and moral/ethical behaviour. Other research in human communication research measures integrity by simply 'being honest' or 'having integrity' [404]. Turning to integrity in human-computer interaction, we see similar concepts taking the form of integrity definitions. For example, McKnight defined integrity as beliefs of honesty and promise-keeping for building trust in e-commerce systems [248]. Jensen et al. measured the integrity of a drone system as being truthful in communication, honest, keeping commitments, being sincere and genuine, and performing as expected [170]. In both the studies mentioned above and in [311, 366], integrity in human-AI interaction is strongly related to honesty and being transparent about the process of decision making.

Kim et al. [195] and Wang & Benbasat [31] explored integrity in terms of fair dealings and unbiased decision-making approaches. Kim et al. found that a robot's integrity is responsible for mediating the relationships between a robot and human trustworthiness. In a recent work by Knowles & Richards, integrity is highlighted in promoting public trust in AI [197]. According to the authors, trust in AI arises in part from a perception of coherence between the human norms as highlighted by Giddens [127]. Giddens talks about human norms in two dimensions - the degree to which agents within the institution are empowered and the use of language by the AI agents. These dimensions resonate with scholarship on trust that emphasises the importance of integrity. In synopsis, we can understand that integrity of an AI agent plays an essential role in building trust. Some approaches link integrity to the sharing of (moral) principles, or keeping to human norms. In other approaches, specific principles are mentioned to constitute integrity. Differences exist, but some common principles related to integrity are honesty, keeping promises/commitments, consistency and fairness. For AI in particular, transparency in decision making is often mentioned as key to integrity as well.

### 5.3.2 OUR APPROACH ON INTEGRITY: INTEGRITY-LADEN EXPLANATIONS

As discussed in the previous section, many definitions of integrity in the AI literature focus on specific principles. In this chapter, we specifically focus on three of them: honesty, transparency and fairness. These are all often used as honesty [170, 220, 404], transparency [17, 93], fairness [50, 197]. Although keeping commitments [356] and consistency [299] are also often used, we choose not to use them in our setting. Keeping commitments and being consistent both imply longer term interaction, and would be most logically related to behaviors more than explanations. Moreover, we could imagine more principles of integrity are used in different settings. We do not argue our list is complete, but rather

make a starting point with three important principles to potentially incorporate in XAI.

Honesty, transparency and fairness are all complex concepts which should be employed in decision making of AI [12]. In this chapter, we choose to express elements of honesty, transparency and fairness in a way that suits XAI. This means we do not claim that our explanations fit the full picture of what it means for an AI system to be e.g. 'honest'. Rather, we designed a specific set of explanations aimed at highlighting: honesty in terms of highlighting uncertainty and confidence; transparency in terms of explaining the process of decision making; and fairness in terms of sharing with users the possible risks and biases that may exist in the advice.

We picked these specifications as they make sense for AI to use in explanations. Uncertainty is often highlighted in confidence explanations [365], transparency is often mentioned as a keystone of AI and the decision making process is something which should be particularly transparent [111], and giving fair advice means not only trying to exclude biases and risks as much as possible, but also being open about this [250]. These specifications also align with the work of Wang and Yin, who provided three desiderata of designing effective AI explanations [395]. These desiderata include (a) designing explanations improve people's understanding of the AI model, (b) helping people recognize the uncertainty underlying an AI prediction, and (c) empowering people to trust the AI appropriately. For brevity's sake, we will use the broader terms 'honesty', 'transparency' and 'fairness about risks' to refer to our specific expressions in the remainder of this chapter.

5

### DESIGN OF EXPLANATIONS

Based on these specifications, two researchers with a Computer Science background and one with a Cognitive Science background brainstormed together and generated sentences that formed explanations expressing the principles of integrity in three different ways. We followed the notion of situation vignettes following the work by Strackand & Gennerich [352] to create text-based explanations.

Each explanation creator was provided with a stack of different expressions of the principles of integrity as identified above. Each note card had one expression printed on it. Each creator read through each other's explanations and decided if they felt it fell within scope or out of scope of the principle to be expressed. For each explanation, creators then described their reasoning for classifying the expression of integrity as within or out of scope. In the end, all creators engaged in similar reconstructive processes to finalize the explanations by controlling the length (word limit [20]) of explanations for the three integrity aspects (honesty, transparency, and fairness). Overall, three iterations were performed for each explanation. The main author followed-up with any necessary questions to determine the researcher's interpretation of each hypothetical situation.

Once the explanations were completed we divided them in a four-part schema. We choose to follow a schema to keep consistency and uniformity in the integrity-based explanations throughout different conditions. Also, keeping a schema supports designing AI agents who can provide forward-reasoning decision support [418] *i.e.*, helping people understand the information in phases and make an informed decision.

The first part of the schema shows an explicit reference to the integrity principle, for example - *I think it is important to be transparent, so I'll tell you how I came to this decision.* This means that the agent explicitly acknowledges that they value a certain

principle. Further, all explanations contain a reference to the source of the data on which the suggestion is based; an estimation of the total calorie count based on the identified ingredients; and the answer the agent picks.

To compare the different expressions of integrity, a baseline explanation was also designed. This type of explanation did not include a specific reference to an integrity principle, but always expressed the source of the data, an estimation of calories without referencing the ingredients, and the final answer chosen by the AI agent.

### EXPRESSIONS OF INTEGRITY IN EXPLANATIONS

Our expressions of integrity are portrayed based on the following schemas. In addition to elements of the baseline explanation (the source of the ingredients and the final answer), all the integrity based explanations included a list of ingredients identified by the AI system of a food plate. Variation was added to avoid mechanical and ‘fake’ looking explanations. Specific examples of the different ways of expressions of integrity through explanations can be found in the Table 5.2.

1. **Honesty** explanations always start with a reference to honesty, followed by an estimation of how sure about the total calories on the plate (e.g. so, I’ll tell you that I’m not entirely sure about identifying the total calories on this plate). Often a confidence % is already added, and usually there are at least 2 statements (e.g. one explaining why this confidence level, one giving options on what the dish could be, or what it could contain, e.g. It could be Caprese salad with 88% of confidence or beet salad with 85% of confidence).
2. **Transparency** explanations always start with a reference to transparency, followed directly by the selected answer and a “I’ll tell you how I came to this decision”. Following (usually directly) is an indication of how sure the system is of what it could possibly be, sometimes in combination (e.g. I’m almost sure this is x, however, I’m not sure about item x”). Sometimes there is a further explanation of why the system is this sure (e.g. because of the low image resolution, My algorithm has failed to recognize the identified portion.”), or some more information about the dish (e.g. Salsa is usually spooned over nachos and are sprinkled with grated mozzarella).
3. **Fairness about risks** explanations always start with a reference to bias, followed by an indication of how sure the system is of what it could possibly be, sometimes in combination (e.g. I’m almost sure this is x, however, I’m not sure about item x). The reasoning explanation can be an explanation of the (lack of) confidence for a choice or of the choice itself. There is always either an explanation of the confidence or an explicit reference to how large the chance of bias in the process or data would be or even both. In some cases, there is a warning with the final answer that bias might be present. The specific explanation is unique for every dish, so no explanation is repeated exactly.

### DIFFERENCES BETWEEN INTEGRITY BASED EXPLANATIONS

1. **Baseline v/s Integrity Conditions:** The baseline lacks a reference to any specific principle, and only refers to the source of the data used, an estimation of the total calorie count and the final answer. The three integrity conditions all include this

data source, estimation and answer as well. In addition, they each explicitly refer to their own principle to start.

2. **Honesty v/s Transparency and Fairness:** Honesty explanations prioritize providing accurate and truthful information about the AI agent's decision-making process and highlighting uncertainty. Also, it is the only one which explains what the confidence intervals mean.
3. **Transparency v/s Honesty and Fairness:** Transparency explanations aim to provide a comprehensive and understandable view of the AI agent's inner workings, without necessarily prioritizing the accuracy or truthfulness of the information provided. Also, it is the only one with a visual representation of ingredients identified, includes more references to what the decision is based on, and mentions the final decision both at the start and end, rather than just the end.
4. **Fairness v/s Honesty and Transparency:** Fairness explanations focus on ensuring that the AI agent's decision-making process does not unfairly discriminate against certain individuals or groups. It also explains why it is certain and where biases might occur more than the others.

## 5

We also designed visual explanations exclusively for the transparency condition of the integrity as this notion deals with the process of decision making. Our classifier provided comparative examples of visual classification. These visualizations categorize confidence values into buckets, such as High / Medium / Low, showing the category rather than the numerical value. The cutoff points for the categories were best match (confidence score  $> 0.8$ ), good match ( $0.5 < \text{confidence score} < 0.79$ ) and unsure match (confidence score  $< 0.49$ ); refer Figure B.4 on page 159. These cutoff points were set in accordance with prior study by Kocielnik et al. [198], and Google's PAIR guidebook [2].

Table 5.2: Different ways of expressing integrity through explanation by an AI agent

| Expression of integrity  | Explanation   |
|--|---|
| <b>Baseline</b> (Average length = 55 words, SD = 6 words)              | <i>The ingredients that I can correctly identify are displayed in the list and their confidence scores. The information I have is based on the UNESCO food nutrition website data. On adding, the total calorie count is 738 calories. Therefore, I would tick option 750 based on the identified ingredients.</i>  |
| <b>Honesty</b> (Average length = 125 words, SD = 23 words)             | <i>I think it is important to be <b>honest</b>, so I'll tell you that I'm not entirely sure about identifying the total calories on this plate. I am not confident about the food item encircled in dark white circle. This is because I have limited training data matching with this encircled food item. The items that I correctly identified are in the table. The information I have is based on the data taken from UNESCO food nutrition website. On adding, the total count is 750 calories which is closer to 738. Therefore, I would tick the option 750 with my overall confidence level as 62.5%. This confidence level means I am moderately sure about my answer.</i>  |
| <b>Transparency</b> (Average length = 128 words, SD = 19 words)        | <i>I have selected 750 calories as the answer to this question. I think it is important to be <b>transparent</b>, so I'll tell you how I came to this decision. I found a similar dish based on my training data from the UNESCO Food &amp; Nutrition website that closely matches the given plate for the calories count. The dish I found is a curry; however, I am not sure about which curry it is. The matching visualization is shown next to the identified ingredients. Based on my training data and similar dish search, the total amount of calories should be 738 calories with 62.5% confidence, similar to the best match example.</i>  |
| <b>Fairness about risk</b> (Average length = 130 words, SD = 25 words) | <i>I think it is important to be <b>fair and unbiased</b>, so I will explain how I combat bias in my answer. I'm not entirely sure about identifying the total calories on this plate. I am not confident about the food item encircled in dark white circle. This is because there is no clear pattern among human annotators of this image. This image is labelled as an Indian Madras curry from UNESCO food nutrition website but I can find annotators for its ingredients only from the western population out of which no one has a profession tag of chef. They have classified the encircled item as bay leaf, fish, meat, chicken or beef. The items that I correctly identified are in the table which gives an estimate of 738 calories. Combining all the existing knowledge with uncertainty regarding the encircled item I will select the option 750 with my overall confidence level as 62.5%.</i> |

## 5.4 METHOD

### 5.4.1 PARTICIPANTS

One hundred eighty two participants (89 female, 93 male) were recruited to participate in the study, via the online crowdsourcing platform Prolific (mean age = 24.8 years, SD = 4.4 years) and the student university mailing list (mean age = 22.1 years, SD = 2.3 years). We recruited through two different methods because we had less turnout of students from the mailing list due to long study completion time. There were no differences among the two samples of participants for the responses we received.

A total of 121 participants participated through the crowdsourcing platform and 61 through the university mailing list. We choose Prolific platform because it is an effective and reliable choice for running relatively complex and time-consuming interactive information retrieval studies [362]. Participants were selected based on the following criteria: age range (18+ years old); fluent level of English - to ensure that participants could understand the instructions; and had no eating disorder - to ensure minimal risk to participants for viewing different food items.

A 35% of the participants reported having studied computer science or some related field. Our participants were from 30 different countries, with most participants reportedly born in the United Kingdom (35), Germany (26), the USA (20), and India (20). Participants were informed about the nature of the task and the total completion of around 35 minutes. Those who accepted our task received brief instructions about the task and were asked to sign an informed consent before beginning their task session.

The study was approved by the Human Research Ethics Review Board of Delft University of Technology (IRB #2021-1779). Prolific participants received an honorarium of £ 5.43/hr for their participation. All participants were provided an option to participate in 5x 15 Euro Amazon gift voucher raffle prize.

### 5.4.2 TASK DESIGN

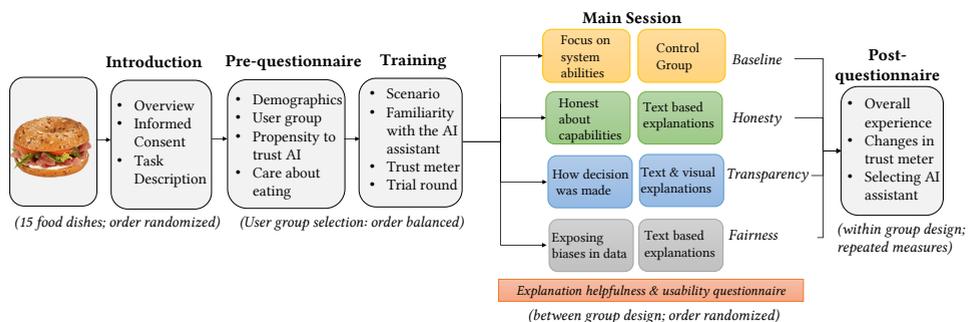


Figure 5.1: This figure illustrates the experimental design of the user study. Each participant was assigned to a experimental condition (Baseline, Honesty, Transparency and Fairness about risk) and they finished 15 rounds in approximately 35 minutes with a 2-minute break after 7 rounds to avoid fatigue effect [70].

We aimed to establish *human-in-the-loop* collaboration in our experiment; *i.e.*, a human making a decision with the assistance of an AI assistant. In our experiment, participants were asked to estimate the calories of different food dishes based on an image of the food.

We designed this task around calories as an approachable domain for our participants. The food dishes in the our experiment were specialized dishes from different countries around the globe. It is rare that participants can judge all the food dishes well but are often good at judging their own cuisine. Therefore, we told participants that there is an AI assistant to help them in identifying the correct amount of calories.

During the brainstorming session of the authors, we decided to use the Food-pics database [34] for selecting our dishes. We selected this database because it contains most popular dishes for European and North-American populations from across the globe along with detailed meta-data of the dishes. Fifteen randomly selected food dishes (referred to as ‘rounds’ hereafter) were taken from this database in the main experiment. Each round consisted of five steps.

**Steps of the task:** At the *first* step, participants were shown an image of a food dish. They were asked to select their confidence in correctly estimating the calories of the food dish. Specifically, we asked our participants, on a scale of 1-10, with 1 being ‘Not at all confident’ & 10 being ‘Fully confident’ - How accurately can you estimate the calories of this food image (Q1)? A zoom-in option was also provided to participants to have a closer look at different ingredients of the food image. Subsequently, they were asked to guess one of the four options they believed to be closest to the correct amount of calories in the dish. One option out of four was always the correct answer, and the first step only involved guessing the correct answer.

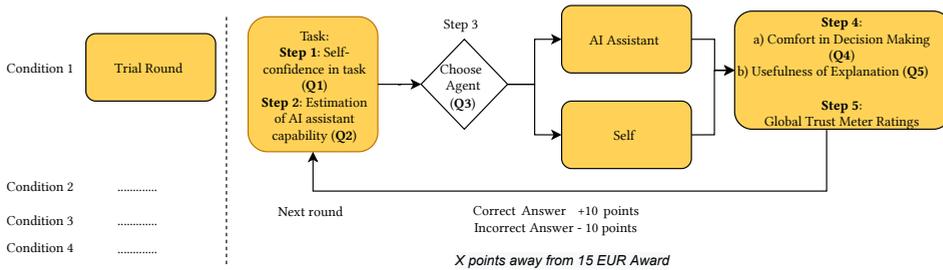


Figure 5.2: A between-subject measure of demonstrated trust. Participants interact with an agent, and then must choose whether to rely on themselves or the AI assistant to complete each task in a sequence of tasks. An incorrect answer is a risk to the trustor causing reduction of 10 points and further away from the required points to receive an award.

At the *second* step, an AI assistant guessed the correct answer from the same options as step one. The AI assistant provided a list of ingredients that it believed to be a part of the dish and the dish name with confidence scores (*for details refer Figure 5.3*) in real time. The AI assistant also explained the reasoning for an answer by providing explanations. Additionally at this step, participants were asked (Q2) to tick a checkbox if they believed that the AI assistant could better estimate the calories than themselves. At the *third* step, participant selected their final decision by choosing between themselves or the AI assistant (Q3). At the *fourth* step, participants rated their comfort level in making the decision (Q4) and usefulness of explanations (Q5). Finally, at the *fifth* step the correct answer was shown to the participants and participants were asked to adjust their trust level in the AI assistant.

**Panel A: Round 1 Step 1**  
 Question: What is the total calorie count of this food image?  
 On a scale of 1 - 10, with 1 being 'Not at all confident' & 10 being 'Fully confident: How accurately can you estimate the calories on this food image?'  
 Q1:  300  750  900  1100  
 Go to Step 2

**Panel B: Round 1 Step 2**  
 Question: What is the total calorie count on the food plate?  
 Click on the below button to ask the AI agent & carefully read its explanation!  
 Ask the Agent  

| Ingredient | Confidence Score |
|------------|------------------|
| fruit      | 0.9904504        |
| vegetable  | 0.9809111        |
| beef       | 0.96175015       |
| pork       | 0.94082177       |
| sauce      | 0.92987496       |
| onion      | 0.9166277        |
| tomato     | 0.8752977        |
| soup       | 0.84438676       |
| chicken    | 0.83761346       |

 I have limited training data matching with the specified food item in the image. The items that I correctly identified are in the table. The information I use is based on the data taken from UNESCO food nutrition website. On making the list, I used a total of 100 calories. Therefore, I could not take the option 750 based on the identified ingredients.  
 300  750  900  1100  
 Select this box if you think that the AI agent is better able to estimate the calories than yourself.  
 Go to Step 2

**Panel C: Round 1 Step 3 - Final Step**  
 Question: What is the total calorie count on this food plate?  
 Final Selection:  300  750  900  1100  
 Next

**Panel D: Round 1 Survey Questions & Correct Answer**  
 Q4 Question 1: On a scale of 1 - 10, with 1 being 'Not at all comfortable' & 10 being 'Very Comfortable': How do you feel about your decision?  
 Q5 Question 2: On a scale of 1 - 10, with 1 being 'Not at all helpful' & 10 being 'Very Helpful': Was the explanation by the agent helpful in making the decision?  
 Your Selection: 300 Agent Selection: 750 Final Selection: 300  
 The correct answer to previous question is: 150; It's a Chicken Curry from India!  
 We're so sorry, that's a wrong answer! Your team score is -10 points  
 Global Trust Meter: Adjust your trust level in the agent > Complete Distrust — Complete Trust  
 Next! Round 2

Figure 5.3: Illustration (a simplified version) of the four steps performed by a participant of the user-study. In step 1, participants rate their confidence in accurately identifying the calories (Q1). In step 2, the AI agent selects its answer with its reasoning in form of explanations and confidence scores (Q2). In step 3, the participants makes their final decision (Q3). Finally, in step 4, participants rate their comfort in decision making and usefulness of the explanations (Q4 and Q5).

An overview of the above steps is visualized in Figure 5.2.

**Scoring method:** Each correct answer yielded +10 points, and an incorrect answer cost -10 points. We specifically applied -10 points for a wrong answer to involve the risk factor associated with trust. Additionally, participants were informed that if they end up in the top three scorers on the leaderboard, they will qualify to receive a 15 Euro gift voucher. The idea to include the leaderboard was to turn a single-player experience into a social competition and provide participants with a clear goal. Participants were only informed about the top scores of the leader board and their rank once they finished the task. We did this to ensure that participants make an informed selection till the end of the task to qualify for the prize. Based on our exit interviews, participants were careful with their selection as they wanted to maximize their chance of winning the award.

### 5.4.3 MEASURES

We used two types of measures. Firstly, subjective measures where users directly report their opinion (referred to as 'subjective measurement' hereafter) (e.g., [82, 119, 409]). Secondly, behavioral measures (e.g. reliance [49, 99] and trustworthiness e.g., [114, 117, 170]). We used the wording AI assistant instead of AI agent for the ease of participants.

**Subjective measures:** Guided by the trust definition in the human communication research domain [397], we measured participant's trust inspired by Yang et al. [406] as four different

measures: (1) cognitive trust to understand human estimation of AI agent capabilities [177], (2) participant's comfort level in making a decision [406], (3) usefulness of the AI assistant explanation [406], and (4) a global trust meter that captures changes in trust [189].

*First*, human cognitive trust to follow the AI assistant recommendation was measured via Q2: Select this [check] box if you think that the AI agent can better estimate the calories than yourself. We informed our participants that by selecting the check-box they believe that the AI agent is better at the task than themselves.

*Second*, human comfort was measured by the question: Q4 - "How do you feel about your decision?" this question measured participants' comfort in taking a decision and was rated on a 10-point Likert scale from *Not at all comfortable (1) to Very comfortable (10)* with a step size of 0.2 *i.e.*, step sizes were 1.0, 1.2, 1.4...9.8, 10.0. We included this question in our user study for two reasons: (1) based on recent work by Yang et al. [406] indicating the importance of human comfort in decision making and (2) based on our pilot study where participants often used the word 'comfortable' to describe their decision which also matches with prior work by Wangberg & Muchinsky [387].

*Third*, AI assistant explanation was measured by the question Q5: "Was the explanation by the AI assistant helpful in making the decision?" This item was rated on a 10-point Likert scale from *Not at all helpful (1) to Very helpful (10)* with a step size of 0.2.

*Finally*, a linear "Trust Meter" ranged from complete distrust (0) to complete trust (+100), inspired by Khasawneh et al. [189]. Participants were asked to adjust the trust meter after every round if their trust in the AI assistant changes. The trust meter was always available to participants and took the previous round's trust meter value in every new round. For the first round, the default value of the trust meter was set at 50.

**Behavioral measures:** For trustworthiness and reliance on the system, we looked at what the participant and AI agent did. *First*, our trustworthiness (TW) measurement was about who was better at the task, so could be either the participant, the AI agent, or both. It was measured by considering how far the selected option was from the correct answer. No two options among the four options were equal distance from each other. For example, if available options are 25, 66, 97 & 143 of which the correct answer is 97, and human selection is 66 & AI agent selection is 143 then human TW is higher than the AI.

*Second*, participants were asked to 'Select your final decision by selecting among the two options – yourself or the AI assistant's guess' (Q3). With Q3, we measured reliance (distinct from trust as we discussed in the introduction) by analyzing the behavior of the participants. If they followed the AI assistant's advice or decision and selected it, they were considered to rely on it. If they switched their answer to another answer than the advised answer, they did not. In case, the two options were same, participants were asked to still decide based on the reasoning for calories of the dish, classification of ingredients, and confidence levels. Their choice determined their reliance behaviour on the AI agent.

It is important to note that although trust and reliance are related concepts, they should be measured as independent concepts. In this work, we follow this distinction as pronounced by Tolmeijer et al. [363], where trust is the belief that "an agent will help achieve an individual's goal in a situation characterized by uncertainty and vulnerability" [217, p. 51], while reliance is "a discrete process of engaging or disengaging" [217, p. 50] with the AI agent.

#### 5.4.4 EXPERIMENTAL SETUP

The study was a mixed between- and within-subject design. The within subject factor was subjective ratings and between subject factor was the integrity condition. This design choice was inspired by Hussein et al. [162]. Participants were randomly assigned to one of four different experimental conditions ("Baseline", "Honesty", "Transparency", and "Fairness"). Each condition had an equal number of participants. We did not manipulate other factors such as time [260] and workload [84], but we controlled reliability [219] and risk factors [249]. The advantage of this experimental setup as stated by Miller [262] is that we can perform detailed analysis on the relationship *Trustworthiness* → *Perceived Trust*, which in turn helps in understanding appropriate trust.

We utilized clarifai Predict API with the 'Food' model to recognize food items in images down to the ingredient level.<sup>2</sup> Our visual classifier returned a list of concepts (such as specific food items and visible ingredients) with corresponding probability scores on the likelihood that these concepts are contained within the image. Our pre-trained classifier accuracy was about 75% (11/15=73.33%), roughly matching the average actual classifier's accuracy of 72%. The list of ingredients along with their confidence score was represented in the form of a table as shown in Figure 5.3.

5

**Sequence of trials:** Each participant finished all 15 rounds, including a trial round. The number of rounds was decided to (1) compare with other experiments that studied trust (e.g., [362, 406]), (2) have enough trials to develop trust but prevent participants from memorizing the order (serial position effects [156]), and (3) have sufficient data for all the integrity conditions.

In each condition, participants finished a sequence of trials. All the sequences had the identical order of correct/ incorrect recommendations by the AI assistant. This identical order allowed us to compare different conditions. We also ensured that the AI agent response in the trial round was always correct to protect trust in an early stage and to not skew or strongly bias towards wrong [239]. Food dishes in the sequence were randomized, and the instances used for training and practice were excluded in the main trials. On completion, participants were asked to fill in a post-experiment questionnaire targeted towards (a) their overall experience, (b) possible reasons for their changes in trust meter and (c) their decision to select themselves or the AI assistant.

**Pilot Study & Pre-test of Explanations:** We used a think-aloud protocol with three participants for a pilot study. The aim of the pilot study was to test the experiment design and check the explanations manipulations. In our experiment, participants were comfortable with estimating calories of the food dishes based on their familiarity with the cuisine and often choose the AI agent when they were not confident. For example, a participant who identified himself as an American often relied on the AI agent to guess food dish from Myanmar. Similarly, another participant who identified herself as an Asian often relied on the AI agent for a Mexican food dish. Based on these observations and UI layout feedback from the participants, we fine-tuned the questions and instructions. After the experiment was finished, we checked for manipulation of explanations. We asked our participants to describe the principle of integrity they saw in the experiment from the

<sup>2</sup><https://www.clarifai.com/models/ai-food-recognition>

note cards that we used earlier with the explanation creators. All participants correctly identified the integrity principles from the note cards. This result helped us in pre-testing our explanation and start with the main experiment. We excluded these three participants from the main experiment.

#### 5.4.5 PROCEDURE

After participants provided informed consent, they saw an overview of the experiment. As shown in Figure 5.1, participants were first asked to complete a pre-task questionnaire consisting of (i) demographic questions about their age and gender, as well as (ii) the propensity to trust scale [118] (Q6) and a balanced diet eating question (Q7) on a 10-point Likert scale from *'I don't care of what I eat'* to *'I care a lot of what I eat'*.

At the beginning of the experiment, we told participants that they would work with an AI assistant and hinted that it could be wrong in its recommendation. They then took part in a trial session, read the instructions, saw an example of a food dish, and practiced using the trust meter. Participants then proceeded to the main session. For each *step*, as explained in the sub-section 5.4.4, participants first saw an introduction of what they could expect to see. In addition, they were asked to focus on the table generated by the AI assistant for specific food items and visible ingredients with corresponding probability scores. The screenshots of each step are in Figure 5.3.

## 5.5 RESULTS

One hundred eighty-two participants participated in the user study, of which 19 (18 from prolific and one from the university mailing list) did not pass our attention checks, leaving us with 163 participants. Furthermore, one participant selected the AI agent, and two always selected themselves, with a total experiment time of only eight minutes, indicating potentially invalid data. Hence, we removed the data of those three participants. Thus, the results and analysis include the remaining (160 participants (female = 85, male = 75; mean age = 23.6 years, SD = 2.8 years). A power analysis of the mixed ANOVA with G\*Power tool [109] revealed that with 40 participants per group, we have a power of 0.93 (considering a medium effect size of  $f = 0.25$ ,  $\alpha_{new} < .046$ ) [122].

### 5.5.1 EFFECT OF DIFFERENT PRINCIPLES OF INTEGRITY ON APPROPRIATE TRUST

In this sub-section, we analyzed how does the expression of different principles of integrity through explanation affect appropriateness of the trust of a human in that agent (RQ1)? For this analysis, we first conducted a descriptive statistics and then performed inferential statistics on the collected data to study the effect of explanations. The post-experiment questionnaire responses were analyzed to support the results and are reported in Section 5.6.1.

The categorization of trust categories was calculated based on Table 5.1. Following the equations in the table, Higher TW was derived based on the TW measurement (as described in Section 4.3). The value for Human trusts who? was based on the participant's response for Q2, and for Human selection, it was based on Q3. On entering these values in Table 5.1, we got our five different trust categories as described in Section 2.2.

**Frequency Distribution:** To understand the frequency distribution of different trust categories as observed for the explanations expressing different principles of integrity in Table 5.3 as % distributions. For example, consider a participant who viewed explanations expressing honesty about uncertainty and who fell into the appropriate trust category seven times, inconsistency (good and bad outcome) two times each, under trust three times, and over trust once. Then, for the honesty condition in Table 5.3 we report appropriate trust as 0.46, inconsistency (good and bad outcome) as 0.13 each, over trust as 0.20, and under trust as 0.06 on a scale of 0-1. Each condition consists of data of 40 participants collected over 15 rounds *i.e.*, 600 data points per condition.

| Condition    | App. Trust | Inconsistency (Bad) | Inconsistency (Good) | Under-trust | Over-trust |
|--------------|------------|---------------------|----------------------|-------------|------------|
| Baseline     | 0.418      | 0.078               | 0.327                | 0.123       | 0.050      |
| Honesty      | 0.433      | 0.068               | 0.285                | 0.158       | 0.053      |
| Transparency | 0.410      | 0.068               | 0.302                | 0.153       | 0.065      |
| Fairness     | 0.552      | 0.060               | 0.218                | 0.088       | 0.088      |

Table 5.3: A contingency table of frequency distribution illustrating number of times different trust categories were observed given explanations highlighting different principles of integrity. Occurrences are scaled as % distributions between 0 (no occurrence) - 1 (always occurred).



Figure 5.4: Figure 4 illustrates the frequency distribution of appropriate trust across 15 rounds and how it is affected by the wrong answers.

**Effect of Integrity Expressions:** We found a statistically significant effect of the integrity principles expressed through explanation on trust categories. A chi-square test of independence  $\chi^2(12, N = 40) = 55.11, p < .001, \phi_c = 0.30$  showed that there is a significant relationship between trust categories and experimental conditions. We further analyzed

our contingency table (Table 5.3) as a mosaic plot [150] to investigate relationships between different trust categories and conditions. While constructing the mosaic plot we extracted Pearson residuals from the output of the  $\chi^2$  results.

We visualized Pearson residuals contribution to the total chi-square score using the correlation plot (for details refer Appendix B, Figure B.1) as our exploratory analysis. Following the correlation plot, a correlation value of  $\rho = 3.45$  between the 'Fairness about risk' explanation and appropriate trust category was found. Following Hong & Oh [154], this correlation implies a strong association between the 'Fairness about risk' explanation and the appropriate trust category.

We were also interested in understanding how different trust categories build up or are relatively stable over time and how they are affected by the wrong answer. Figure 5.4 illustrates the frequency distribution of appropriate trust across 15 rounds. The figure shows that appropriate trust drops with the first wrong answer across the four conditions. However, this effect does not perpetuate in later rounds. It is interesting to note that appropriate trust builds up over time (rounds 1 to 4) and recovers slowly after each wrong answer. We also provide a similar graph as Figure 5.4 in the supplementary for other trust categories.

**Predictors for Trust Categories:** The trust categories were binary variables in our study: either the participant achieved appropriate trust or not. For this reason, we also conducted a multilevel logistic regression per category, predicting proportions of the five trust categories separately. In our model, each round was treated as one observation i.e., each row was one observation, with 15 rows per participant.

*Baseline Model:* We first created a baseline model, which comprised of a random intercept per participant and the different explanation conditions. Next, we added the 'Wrong answers by the AI agent' as additional fixed effects factor to our baseline model. Our dependent measure indicated whether this behaviour is an appropriate trust behavior or not (similarly for other trust categories). Furthermore, we added a lag factor as a fixed effect to observe the effect of the previous round answer on the trust rating of the current round. The lag factor was coded as 1 if the previous trial was correct and 0 if not.

*Baseline Model plus Covariates:* We added three covariates 'Care about eating' responses, 'Propensity to trust' responses, and human confidence in estimating the calories (Q1) to our baseline model one by one. Since the  $\chi^2$ -based ANOVA comparison showed no significant improvement in the goodness-of-fit of the model upon adding the covariates and none of the covariates were significant predictor of any trust category, we decided not to include them in the models. For comparing the models for goodness-of-fit, AIC & BIC values are provided in the Appendix B, Table B.5. We also report an marginal and conditional R-squared values, which indicates variance explained by both fixed and random effects, see Table B.1.

*Appropriate Trust:* For the appropriate trust category, the 'Fairness about risk' explanation was the only statistically significant predictor. The coefficient value of 'Fairness' ( $\beta = 0.591$ ,  $p < .001$ ) is positive. Thus, we can say that when a participant interacted with the AI agent explaining with a focus on fairness through exposing risk and bias, the participant was more likely to achieve an appropriate level of trust in the AI Agent.

*Inconsistency:* For the inconsistency with a bad outcome trust category, we did not find any statistical significant predictor variable in our analysis. However, for the inconsistency

with a good outcome trust category, the ‘Fairness about bias’ explanation was again the only statistically significant predictor variable. The coefficient value of ‘Fairness about bias’ ( $\beta = -0.526, p < .001$ ) is negative. Thus, we can say that when participants interacted with the AI agent explaining with a focus on being fair by exposing bias and risk, the participants were less likely to end up in the inconsistency with a good outcome trust category.

*Under-trust and Over-trust:* For both the under-trust and over-trust categories, we did not find any statistically significant predictor variable in our analysis.

### 5.5.2 EFFECT OF DIFFERENT PRINCIPLES OF INTEGRITY ON SUBJECTIVE TRUST

In this sub-section, we analyzed how does human trust in the AI agent change given these different expressions of integrity principles (RQ2)? For this analysis, we performed a similar approach as RQ1 first to conduct descriptive statistics followed by inferential statistics where we focused on a multilevel regression model. Here also, post-experiment questionnaire responses were analyzed to support the results and are reported in Section 5.6.2.

5

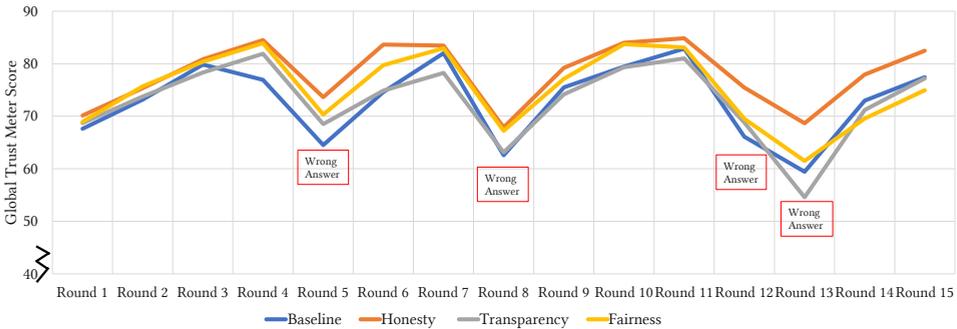


Figure 5.5: Illustration of mean responses for changes in Global Trust Meter over 15 rounds. The red coloured boxes represents when the AI agent provided a wrong answer *i.e.* round 5, 8, 12 and 13.

**Change in Trust level Overtime:** We used a global trust meter to capture changes in trust over time. First, we calculated changes in human trust towards the AI agent over time by subtracting differences in trust meter values between every two subsequent rounds. As can be seen in Figure 5.5, trust in the AI agent dropped whenever the AI agent provided a wrong answer. We recorded an average drop of 15 points in trust score when a wrong answer was preceded by a right answer by the AI agent. This drop was more than twice the number of points when there were two wrong answers in a row *i.e.*, around 35 points. These results seem to confirm that the AI agent’s accuracy influences trust.

**Predictors of Subjective Trust Scores:** Our dataset includes one row for each participant and one column for each variable or measurement on that participant. In the context of longitudinal data, this mean that each measurement in time would have a separate row of its own, therefore we analyzed this data using a multilevel regression model following the instructions by Finch et al. [115, Chapter 5].

*Baseline Model:* We analyzed the global trust meter responses as our dependent variable to test the effect of different principles of integrity expressed through the explanations with a multilevel regression model with random intercept for trials. In addition, we added the current round correctness and lag as additional factors in our baseline model to test the effect of it on subjective trust scores. Since our data is linear we used the LMER method for this analysis with the lmerTEST package v3.1 [203].

*Baseline Model with a lag factor plus Covariates:* We now added fixed interaction effect between the correct/incorrect answer and the lag variable to the model. Furthermore, we also examined the two-way interaction effect between the correct/incorrect answer with different explanations types. This model was significantly better than the other two models for the goodness-of-fit,  $Pr(> \text{chisq}) < 0.05$  (refer Appendix B, Table B.4 for further details). Hence, we finalized this model as reported in Table 5.5.

Following the same procedure as RQ1, we further explored adding same covariates to our model. Adding these covariates did not improved our model and therefore we did not include those variables in our final model. Finally, we added human comfort and usefulness of explanations ratings to the model and found that only the usefulness of explanations helps in improving our model.

Based on the regression results we can observe that the honesty explanation is a significant predictor of the trust score compared to other explanations expressing integrity ( $\beta = 7.84, p < .05$ ) i.e., participants who saw the honesty explanation rated their subjective trust in the AI agent higher than the other conditions. Furthermore, as shown in Table 5.5, both the correct/incorrect answer and the lag variable are statistically significant predictors of the subjective trust ratings ( $p < .05$ ). This result confirms our intuition observed from Figure 5.5 where the effect of the correct/incorrect answer on the trust scores can be observed. Interestingly, the significance of the lag variable show the effect of the previous round correctness on the current round trust score. In other words, as it is important to study the effect of the correct answer on the trust score for the current score, it is equally important to study how the AI agent performed a round before to observe the changes in the trust score.

Additionally, our results show that the interaction effect between the correct/incorrect answer and the lag variable is significant ( $\beta = -3.38, p < .05$ ). Given that the sign on the interaction coefficient is negative, we would conclude that there is a buffering or inhibitory effect. Analyzing the correct/incorrect answer, lag, and their interaction reveals the drop and restoration of global trust ratings. For instance, two consecutive correct trials yield a combined score of 21.44, while a correct trial followed by an incorrect one results in a high initial drop of 7.77. Similarly, an incorrect trial followed by a correct one leads to a recovery to 17.05, almost reaching 21.44 again. Two consecutive incorrect trials cause a complete drop to 0, followed by a gradual recovery to 7.77, 17.05, and 21.44. These findings align with the results in Figure 5.5.

Moreover, there is a significant interaction effect between the correct/incorrect answer and the honesty explanation ( $\beta = -4.63, p < .05$ ). This indicates that the impact of errors is smaller in the honesty condition, as depicted in Figure 5.5. Also, usefulness of explanations is a predictor of global trust ratings ( $\beta = 9.45, p < .0001$ ). This result means the participants found the explanations helpful in adjusting their trust levels after each round.

### 5.5.3 EFFECT OF DIFFERENT PRINCIPLES OF INTEGRITY ON USEFULNESS

In this sub-section, we analyzed how do different expressions of integrity principles influence the human's decision-making, and do people feel these explanations are useful in making a decision? (RQ3)? For this analysis, we performed a similar approach as in RQ2.

**Descriptive statistics:** We used human comfort ratings (Q4) and usefulness of explanations ratings (Q5) by participants to analyze our responses for RQ3. These ratings were measured after each trial. Therefore, we followed the same analysis method as for RQ2. For the human comfort ratings, we did not find any major differences among the four conditions, refer Figure B.2, Appendix C. The mean ratings for the baseline condition was 6.178 (1.981), for honesty 6.285 (1.863), for transparency 6.246 (1.811), and for fairness 6.128 (1.948). Similarly, for the helpfulness of explanations ratings, we also did not find any major differences among the four conditions, refer Figure C.1, Appendix B. The mean ratings for the baseline condition was 6.333 (2.053), for honesty 6.675 (1.764), for transparency 6.486 (1.845), and for fairness 6.423 (1.831).

**Predictors of Comfort and Explanations Helpfulness:** We analyzed the human comfort ratings and usefulness of explanations responses as our dependent variable to test the effect of different principles expressed through the explanations with a multilevel regression model with random intercept for participants. We followed the similar model as for RQ2 in analyzing the results of this RQ. Adding the covariates from RQ1 did not improved our both the models (human comfort and explanations help). Also, adding the interactions as in Table 5.5 was not helpful in improving the model statistics. Therefore, we did not include them in our final models. We report the regression model of predicting the usefulness of explanations in Table 5.6 and human comfort in Table B.2, Appendix B.

Based on the Table 5.6, we can observe that the trust score is a significant predictor of the usefulness of the explanations ( $\beta = 0.02$ ,  $p < .001$ ), *i.e.*, participants who rated their subjective trust in the AI agent could have found the explanations provided by it more helpful. Similarly, we found that human comfort in decision-making is another significant predictor of the usefulness of the explanations score ( $\beta = 0.34$ ,  $p < .001$ ). None of the other covariates were found to be significant predictors of the human comfort score except the helpfulness of explanations ( $\beta = 0.35$ ,  $p < .001$ ).

|                                     | Independent variables | Coefficient |       | z value | Pr(> z ) | Significance |
|-------------------------------------|-----------------------|-------------|-------|---------|----------|--------------|
|                                     |                       | $\beta$     | SE    |         |          |              |
| <b>Appropriate Trust</b>            |                       |             |       |         |          |              |
| (Intercept)                         | Participants          | -0.262      | 0.135 | -1.931  | 0.053    |              |
|                                     | Fairness about bias   | 0.591       | 0.162 | 3.650   | <0.001   | ***          |
|                                     | Honesty               | 0.083       | 0.161 | 0.517   | 0.604    |              |
|                                     | Transparency          | -0.009      | 0.162 | -0.594  | 0.552    |              |
|                                     | Wrong Answer          | 0.077       | 0.097 | 0.793   | 0.427    |              |
|                                     | Lag (Wrong Answer)    | -0.139      | 0.097 | -1.439  | 0.150    |              |
| <b>Inconsistency (Bad outcome)</b>  |                       |             |       |         |          |              |
| (Intercept)                         | Participants          | -2.572      | 0.240 | -10.692 | <0.001   | ***          |
|                                     | Fairness about bias   | -0.395      | 0.274 | -1.441  | 0.150    |              |
|                                     | Honesty               | -0.181      | 0.263 | -0.689  | 0.491    |              |
|                                     | Transparency          | -0.170      | 0.263 | -0.645  | 0.519    |              |
|                                     | Wrong Answer          | -0.291      | 0.196 | -1.486  | 0.137    |              |
|                                     | Lag (Wrong Answer)    | 0.149       | 0.190 | 0.786   | 0.432    |              |
| <b>Inconsistency (Good outcome)</b> |                       |             |       |         |          |              |
| (Intercept)                         | Participants          | -0.843      | 0.128 | -6.570  | <0.001   | ***          |
|                                     | Fairness about bias   | -0.526      | 0.147 | -3.571  | <0.001   | ***          |
|                                     | Honesty               | -0.225      | 0.142 | -1.583  | 0.113    |              |
|                                     | Transparency          | -0.059      | 0.140 | -0.425  | 0.670    |              |
|                                     | Wrong Answer          | -0.108      | 0.106 | -1.020  | 0.307    |              |
|                                     | Lag (Wrong Answer)    | 0.162       | 0.106 | 1.519   | 0.128    |              |
| <b>Under trust</b>                  |                       |             |       |         |          |              |
| (Intercept)                         | Participants          | -2.180      | 0.224 | -9.720  | <0.001   | ***          |
|                                     | Fairness about bias   | -0.485      | 0.289 | -1.681  | 0.092    |              |
|                                     | Honesty               | 0.333       | 0.266 | 1.254   | 0.209    |              |
|                                     | Transparency          | 0.246       | 0.268 | 0.915   | 0.360    |              |
|                                     | Wrong Answer          | 0.236       | 0.141 | 1.667   | 0.095    |              |
|                                     | Lag (Wrong Answer)    | -0.140      | 0.143 | -0.978  | 0.328    |              |
| <b>Over trust</b>                   |                       |             |       |         |          |              |
| (Intercept)                         | Participants          | -7.006      | 1.099 | -6.373  | <0.001   | **           |
|                                     | Fairness about bias   | 1.000       | 1.001 | 0.982   | 0.326    |              |
|                                     | Honesty               | -0.125      | 1.014 | -0.124  | 0.901    |              |
|                                     | Transparency          | 0.080       | 1.007 | 0.080   | 0.936    |              |
|                                     | Wrong Answer          | -0.031      | 0.229 | -0.137  | 0.891    |              |
|                                     | Lag (Wrong Answer)    | 0.184       | 0.233 | 0.790   | 0.430    |              |

Table 5.4: Results of GLMER analysis for RQ1 (\*:  $p < .05$ , \*\*:  $p < .01$ , \*\*\*:  $p < .001$ ). The marginal and conditional  $R^2$  values are provided in the Appendix B for each model of the trust category.

| Independent variables       |                                | Coefficient |      | t value | Pr(> t ) | Significance |
|-----------------------------|--------------------------------|-------------|------|---------|----------|--------------|
|                             |                                | $\beta$     | SE   |         |          |              |
| <b>Global Trust Ratings</b> |                                |             |      |         |          |              |
| (Intercept)                 |                                | 46.11       | 2.90 | 15.88   | <0.001   | ***          |
| Participants                | Round                          | 0.01        | 0.08 | 0.12    | 0.907    |              |
|                             | Fairness about bias            | 4.06        | 3.20 | 1.27    | 0.205    |              |
|                             | Honesty                        | 7.84        | 3.20 | 2.45    | 0.015    | *            |
|                             | Transparency                   | 0.68        | 3.20 | 0.21    | 0.831    |              |
|                             | Correct/Incorrect Answer       | 17.05       | 1.68 | 10.14   | <0.001   | ***          |
|                             | Lag Correct/Incorrect          | 7.77        | 1.30 | 6.00    | <0.001   | ***          |
|                             | Correct/Incorrect*Lag          | -3.38       | 1.46 | -2.31   | 0.021    | *            |
|                             | Correct/Incorrect*Fairness     | -2.71       | 1.80 | -1.51   | 0.132    |              |
|                             | Correct/Incorrect*Honesty      | -4.63       | 1.80 | -2.57   | 0.010    | *            |
|                             | Correct/Incorrect*Transparency | -1.55       | 1.80 | -0.86   | 0.391    |              |
|                             | Usefulness of Explanations     | 1.72        | 0.18 | 9.45    | <0.001   | ***          |
| Marginal R <sup>2</sup>     |                                |             |      |         | 0.136    |              |
| Conditional R <sup>2</sup>  |                                |             |      |         | 0.534    |              |

Table 5.5: Results of LMER analysis for RQ2 (\*:  $p < .05$ , \*\*:  $p < .01$ , \*\*\*:  $p < .001$ ) with LMERTest R Package

| Independent variables      |                          | Coefficient |      | t value | Pr(> t ) | Significance |
|----------------------------|--------------------------|-------------|------|---------|----------|--------------|
|                            |                          | $\beta$     | SE   |         |          |              |
| <b>Explanations Help</b>   |                          |             |      |         |          |              |
| (Intercept)                |                          | 2.72        | 0.25 | 10.88   | <0.001   | ***          |
| Participants               | Round                    | 0.01        | 0.01 | 1.24    | 0.217    |              |
|                            | Fairness about bias      | 0.09        | 0.19 | 0.45    | 0.654    |              |
|                            | Honesty                  | 0.29        | 0.19 | 1.48    | 0.141    |              |
|                            | Transparency             | 0.13        | 0.19 | 0.69    | 0.491    |              |
|                            | Correct/Incorrect Answer | -0.05       | 0.08 | -0.60   | 0.548    |              |
|                            | Lag Correct/Incorrect    | -0.14       | 0.08 | -1.87   | 0.062    |              |
|                            | Trust Score              | 0.02        | 0.00 | 10.30   | <0.001   | ***          |
|                            | Human Comfort            | 0.34        | 0.02 | 17.93   | <0.001   | ***          |
|                            | Marginal R <sup>2</sup>  |             |      |         |          | 0.230        |
| Conditional R <sup>2</sup> |                          |             |      |         | 0.391    |              |

Table 5.6: Results of LMER analysis for RQ3 - Helpfulness of Explanations (\*:  $p < .05$ , \*\*:  $p < .01$ , \*\*\*:  $p < .001$ )

## 5.6 DISCUSSION

Our results offers three major contributions for discussion in the field of the Human-AI interaction.

1. We can measure appropriate trust through a formal computation method in the context of a specific task.
2. Appropriate trust can be enhanced by providing expressions of fairness principle of integrity in the context of human-AI interaction. Furthermore, appropriate trust builds up overtime and recovers slowly if an AI agent provides an incorrect output.
3. Subjective trust builds up and recovers better by providing expressions of honesty principle of integrity in the context of human-AI interaction.

Therefore, in this section we will discuss our results about how the explanations expressing different integrity principles influenced appropriate trust. Next, we will discuss how participants perceived the AI agent's advice and made their decision based on theories from psychology and social sciences which possibly led them to select the AI agent. Finally, we will discuss the limitations of our work and possible future directions.

5

### 5.6.1 EXPRESSIONS OF INTEGRITY AND APPROPRIATE TRUST

We found that the 'Fairness about bias' explanations were the most effective for fostering appropriate trust in the AI agent. We know from previous work by Asan et al. [15] that knowing about biases can influence human trust, which perhaps also explains why trust becomes more appropriate if human can intervene in AI decision making.

A closer look at our findings shows that in our case, the explanations highlighting potential bias and risks actually improved appropriate trust through increasing trust rather than decreasing it. This makes intuitive sense as fairness explanations could have triggered more cognitive effort resulting in increase of people's cognitive thinking for engaging analytically with the explanation [43]. Furthermore, recent education research has shown that students actual learning and performance was better with the more cognitively demanding instructions [91]. Overall, Our findings seem to support the proposition that we should be building explainable and bias-aware AI systems to facilitate rapid trust calibration leading to building appropriate trust in human-AI interaction [365].

Interestingly, irrespective of which integrity principle was highlighted, explanations seem to have helped our participants in correcting under-trust and over-trust (see Figure 5.3). In particular, being explicit about potential biases and risks actually decreased inconsistent behaviour with a good outcome over the other explanation types in some cases (including those cases where trust was appropriate). A possible reason is that these explanations exposed potential bias(es) in the data or the model which could have convinced the participants to follow the AI agent. For example, P62 reported that *"If the AI Assistant says dataset is biased, then [it's] true I suppose and it's more trustworthy than my common sense because I haven't seen the data, so I will stick to my initial trust decision (P62, Fairness about bias condition)"*. Similarly, P133 reported *"I feel like the results of [the model] were strange hence I went with my decision first but I was wrong, so next time for a similar round I choose the [AI] Assistant and it was right. Hence, I decided to follow him [AI Agent]! (P133, Fairness about bias condition)"*

Another finding of our study was that irrespective of what principle of integrity was expressed in the explanation, around 30% of the time participants ended up in the inconsistency (good and bad outcome) trust category. This shows that even when participants reported that they trusted the AI agent to be better than themselves, they still quite often chose not to rely on it. Based on our exit interviews, we found that participants acted inconsistently several times during the experiment to increase their score leading to winning the gift voucher. For example, P20 told us *"I think [AI Agent] it is better in identifying this dish, but it was also wrong with a similar dish in one of the previous rounds, so I will choose myself because I do not want to lose any points."* Similarly, P77 said *"Ahh, I was just checking if I say I trust [AI Agent] him but do not go with him then what will happen. If it turns out to be good, I will do this again to keep my score up"*.

We found none of the covariates, 'Care about eating' and 'propensity to trust' a predictor of subjective trust score and any trust category. For 'Care about eating', a potential reason could be that people who rated higher on caring about their eating behaviour were more aware of the different ingredients with their calories level that were known to them and *vice-versa*. Given the images of the food items in our experiment were diverse, this could have impacted their skills to judge the calories well. For example, P97 with a score of 10 for the 'Care about eating' question reported that *"I am very picky about what I eat as I need my balanced diet. However, this task is not easy as it has many international cuisines!"*. For, 'propensity to trust' one possible explanation can be that this dispositional covariate became less important as system experience increased. Alternatively, this covariate could influence trusting behaviors more than trusting beliefs. More research is needed on the effect of propensity to trust factors over time.

5

### 5.6.2 SUBJECTIVE TRUST, HELPFULNESS AND COMFORT

Subjective trust is not the same as appropriate trust [406]. Chen et al. [61] identified in their study that participant's objective trust calibration (proper uses and correct rejections) improved as intelligent agent became more transparent. However, their subjective trust did not significantly increase. The 'Fairness about bias' explanation in our work helped in fostering appropriate trust in the AI agent. However, it did not necessarily improve participant's (subjective) trust. This result is in line with Buçinca et al [43] who showed that there exists a trade-off between subjective trust and preference in a system of human+AI decision-making.

From Figure 5.5 and Table 5.5, it is evident that the subjective trust ratings for the 'Honesty' explanations are significantly higher compared to the other explanation types. This observation can be explained by the explicit references to honesty by the AI agent as reported by P102 *"It [AI Agent] mostly talks about being honest and based on all rounds - I think it is, so I trust it (P102, Honesty condition)"*. We can recall that the AI agent in the 'Honesty' condition expressed its honesty by stating it cared about honesty and adding further information about uncertainty in the decision-making. This expression of honesty resonates with Wilson [399] who argue that as long as communication is performed in a honest way, it produces ecological integrity affecting trust.

We also found the effect of the current and the previous round correctness on the subjective trust ratings, refer Table 5.5. This result is echoed from prior study by Tolmeijer et al. [362] who showed that system accuracy influences the trust development in an AI

system. Furthermore, the effect of the previous round correctness *i.e.*, the lag in Table 5.5 had a influence on the trust score as well. This result indicates that trust is not only influenced by how the system is performing now but also on how it performed before. Human trust develops over time and depends on many factors. Also, each interaction with a system can alter the trust in that system. For example, Holliday et al. [152] looked at trust formation within one user session, they found that the impression of system reliability at each time point shapes trust. Our results are in line with van't Wout et al. [376] who show that the outcome on a previous round (whether trust was repaid or abused) had an effect of how much a participant trust other participant to send money over.

Turning to the transparency explanations, based on post questionnaire responses, the participants found the visual part of the explanation difficult to follow. For example *"I can see there is best, good and unsure match but I have no idea it really helps as everything looks almost same! (P140, Transparency condition)"*. Additionally, we believe that the combination of visual with textual explanations may have hampered understandability as reported by P17 *"That's simply too much of information for me!" (P17, Transparency condition)*.

Overall, trust scores exhibit a consistent level of stability, particularly an initial overall level of trust that remains steady over time, except in cases where an error occurs. (Figure 5.5). This is in line with our intuition of how trust works. Interestingly, while an increase of trust between rounds three and four was expected, trust recovers to same levels between rounds six & seven and nine & ten. A potential explanation can be that the AI agent's early impressions positively influenced the AI agent's perceived reliability, leading to increased trust even after inaccurate advice.

The result in the Table 5.6 demonstrates no effect of type of explanations on participant's usefulness of explanations ratings. However, we found that participant's trust and human comfort scores significantly predicted the usefulness of explanations ratings. We can understand this result as if an explanation was helpful; participants often rated their trust and comfort in the decision-making process higher than the non-helpful explanations.

We also found that participant's decision-making comfort levels were similar across conditions. However, the explanations score significantly predicted the participant's comfort level. A potential reason might be that other individual factors more strongly influence the subjective notion of the comfort of decision-making than the differences between our explanations. Another possible explanation is that different types of explanations by the AI agent did not necessarily improve the comfort level but only assisted in decision-making. A previous study focusing on integrity among colleagues reported that showing integrity did not increase the comfort level of employees to rely on each other [396]. This result aligns with our findings, where it is hard to establish human comfort by expressing principles related to integrity.

### 5.6.3 UNDERSTANDING HUMAN PSYCHOLOGY ADVICE UTILIZATION

Advice utilization has been studied in the literature of psychology to understand how humans utilize the advice given by others [231]. Jodlbauer and Jonas [175] found that while three different dimensions of trust (competence, benevolence, and integrity) mediate between advisor and listener, for the acceptance of advice, trust in advisor integrity played the strongest mediating role in human-human interaction.

Given that all the AI agents in our user study had the same competence level, the only

difference was what principle of integrity was highlighted in the explanation of the AI agent. This might partly explain why integrity expressions of fairness through exposing potential bias and risk; and honesty about uncertainty in decision making could have worked for RQ1 and RQ2 compared to the baseline condition.

The theory by Bazerman and Moore [29] can help us partly understand why explanations exposing potential bias and risk were significantly different from the other explanations used in this study. They showed that humans are limited in their rationality and are often subject to cognitive bias. Furthermore, when decisions involve risks based on unbiased advice and people cannot weigh all relevant information, decision-makers often use the advice [29] which helps in reducing their own bias. Therefore, participants' trust in the 'fairness about risk' condition was more appropriate compared to other conditions. For example, P73 reported that *"I was not sure about different type of vegetables in the salad but the AI told me correctly that it was also not sure, hence I decided not to trust it and went with my best possible option - which was eventually correct!"*.

## 5

#### 5.6.4 REFLECTIONS ON DESIGN CONSIDERATIONS FOR BUILDING APPROPRIATE TRUST

In the prior research, appropriate trust is often linked with [not] relying on the AI system when it makes a [in] correct decision. This notion of appropriate trust heavily relies on the capability of the AI system leaving out other factors that can influence trust, such as integrity or benevolence. Here, our work serves as an example of how expressing different principles related to integrity through explanations can establish appropriate trust in human-AI interaction. Therefore, an essential focus of designing AI for fostering appropriate trust should be both on the capability as well as the integrity of the AI system. However, this comes with the challenge of obtaining accurate measurement information regarding the machine learning models' performance, bias, fairness, and inclusion considerations.

Lord Kelvin has promoted measurement with his memorable statement: "If you cannot measure it, you cannot improve it [187]" There is much discussion on the AI systems to be appropriately trusted. However, there are very few suggestions for measuring the appropriate trust. Part of this lack of literature on measurement is because trust is subjective in nature. What seems an appropriate trust for person A won't be appropriate for person B. Nevertheless, it is also crucial for humans to calibrate their trust, recognizing that AI systems can never be 100% trustworthy. Likewise, we made an attempt to capture trust into various categories (appropriate, over/under trust, inconsistency) through formal definitions.

We believe that our proposed formal definitions can help facilitate communication between researchers, practitioners, and stakeholders by providing a common language and understanding of what is meant by measuring appropriate trust. Furthermore, it can set clear expectations for how trust should be measured, can promote a better understanding of what trust means and what aspects of trust should be considered [32]. We hope this work highlights the need for guidelines to incorporate a method to capture appropriate trust and develop an understanding of human decision-making with psychological theories such as advice utilization.

### 5.6.5 LIMITATIONS AND FUTURE WORK

Our work limits itself to exclusive decision making, which does not represent the full spectrum of possible human-AI interaction. Our task was inspired by scenarios in which a human needs to make a conscious choice to either follow the system advice or their(s) own; such as the autopilot mode or cruise control in a car. Therefore, our findings may not generalize to every scenario such as human-AI teaming where the focus is more on the collaboration. Additionally, in our definition of appropriate trust, we did not further explore the reasons for the selections made by the human. Interesting notions for further study are how our notions of appropriate trust can be influenced by the delegation of responsibility focusing on different choices people make in the delegation. For example, people are more likely to delegate a choice when it affects someone so as not to take the blame if something goes wrong [351].

In our user study, we used images of various food items for estimation of food calories based on a machine learning model. In our day to day situations, people hardly use such technological advances. Therefore, the level of realism can be further improved in the future studies. Furthermore, our users got 15 trials in the same condition which could have led to possible learning or fatigue effects even though we provided a break after seven rounds. Also, the order of the wrong AI advise was same across the conditions which made us hard to control the possible fatigue effects.

We have utilized situation vignettes to craft our explanations. In our work, custom build explanations to highlight different principles related to integrity were better suited to our user study, i.e., by explicitly revealing the importance of individual notions of integrity (honesty, transparency and fairness) in a calories estimation task. In this, we attempted to keep other variables (e.g. length) mostly the same, but for instance it was inevitable that the baseline explanation would be shorter. The style was controlled for in some way by having the same authors for all explanations, but here too, differences might exist between conditions. For instance, the 'fairness about risk' explanation might have been a little more technical, as it explained where in the process risks could come from (e.g. bias in training data). Although we cannot exclude such influences, we would argue that such slight differences will always be inevitable when expressing different principles in explanations. More research on e.g. style of writing, length, etc. would be relevant to further control for such factors [380].

Finally, our explanations express the related principles of integrity in one specific way, and different methods of expressing these might have different effects on trust than what we found. However, with this work we show a method for the AI agent to express its integrity in form of explanations and our aim for this research was not to design effective explanations but to study the how different expressions of integrity can help in building appropriate trust.

A future research direction to scale this work could look at how we can create vignettes by systematically combining actions of the agent based on the affect control theory [143] in the real time. For example, one could adopt ensemble machine learning methods as they are shown to perform well and generalize better for generating action based situations [94]. One could also look at PsychSim [304] framework which combines two established agent technologies: decision-theoretic planning and recursive modeling for crafting explanations using machine learning models.

Furthermore, the understandability of explanations might be further enhanced by design specialists, and tested by crowdsourcing with a diverse demographic sampling. Broader findings would further enable designers to craft explanations to make AI systems more understandable and trustworthy. Finally, further work can explore trusting behaviour targeting both integrity and benevolence as antecedents of trust.

## 5.7 CONCLUSION

Our user study was a means to employ the formal definition of appropriate trust and understand how expressions of principles related to integrity through explanations can help in fostering appropriate trust. In this chapter, we (a) provided a formal definition of appropriate trust following the interpersonal perspective of trust, (b) investigated different ways of expressing principles related to integrity through explanations – honesty about uncertainty; transparency about the decision making process; and fairness in terms of being open about potential bias & risk by an AI agent, and (c) showed the effect of these different types of integrity based explanations on the end user's appropriate trust. Our task involved an exclusive decision making process where participants were required to select either themselves or rely on the AI agent for the task. Our results show a strong correlation between expressing integrity focused on fairness in openness about biases & appropriate trust. In summary, the two key takeaway messages of this work are (1) a measurement method for appropriate trust in exclusive decision making task and (2) expressing integrity principles in explanations given by an AI agent has the potential to improve end users' appropriate trust and enhance the appropriate use of AI systems.

# III

## APPLICATION LENS



## 6

# FOSTERING APPROPRIATE TRUST IN AI-BASED PREDICTIVE POLICING SYSTEMS: A CASE-STUDY

6

Law enforcement agencies worldwide are increasingly using machine learning systems for crime prevention and resource allocation. Predictive policing, a notable example, employs data analysis and algorithms to predict criminal activity and optimize resource deployment. Concerns regarding user trust levels in such systems have garnered significant attention. Under-trust may lead to inadequate reliance, while over-trust can result in over-compliance, negatively impacting tasks. Users must maintain appropriate levels of trust. Past research indicates that explanations provided by AI systems about their inner workings can enhance user understanding of when to trust or not trust the system. The role of explanations in building trust varies based on the task and user expertise. This study explores the impact of different explanation types (text, visual, and hybrid) and user expertise (retired police officers and lay users) on establishing appropriate trust in AI-based predictive policing systems. While we observed that the hybrid form of explanations significantly increased the subjective trust in AI for expert users, no form of explanation significantly helped in building appropriate trust. The findings of our study underscore the nuanced relationship between explanation types and user expertise in establishing appropriate trust, emphasizing the importance of reevaluating the use of explanations. Finally, based on our results we synthesize potential challenges along with policy recommendations to design for appropriate trust in AI-based predictive policing systems.

## 6.1 INTRODUCTION

### 6

Artificial Intelligence (AI) is rapidly reshaping public organizations globally, mainly through machine learning approaches that automate routine administrative tasks and support decision-making [45]. One of the key components to achieving effective decision-making is a user's appropriate trust in AI systems. Despite recent efforts towards enhancing trust in algorithmic decision-making systems (e.g., adding *explanations* [95, 252, 406], *human oversight* [282, 338, 363], and *confidence scores* [26, 417], comparatively little attention has been paid to building appropriate trust in them. Both under-trust and over-trust are deemed inappropriate [217, 294, 314], instead we require trust to be appropriate. Under-trust can lead to under-reliance, and over-trust can lead to over-compliance, which can negatively impact the task [252]. In this work, we study whether we can improve the appropriateness of trust in an AI decision support system (*goal*). We propose to do this through explanations (*means*), and position our work in the context of AI-based predictive policing (*context*).

Explainable AI (XAI) is meant to give insight into the AI's internal model and decision-making [395] and has been shown to help users understand how the system works [52, 287]. Efforts to ensure that AI is trusted appropriately are often in the form of explanations [24, 226, 252]. While there are few prior works showing explanations not helping in trust calibration [24, 43, 417], there is no consensus on this in the community. Several recent works show the usefulness of explanations in trust calibration [211, 218, 252, 377]. Given this clear lack of consensus, the effect of XAI on appropriate trust requires further exploration.

Explanation can be communicated in different forms by an AI system, making it difficult to choose an appropriate design. For example, explanations could be visual in the form of a graph such as a saliency map [4], or textual in the form of words and phrases or analytical, allowing users to explore the data and the model [142, 194]. Despite the considerable interest each of these methods received individually, only a few studies have

compared these different presentation methods to learn when, why, and for whom they work [277, 295, 313, 353]. To address this research and empirical gap, in this study, we investigate how users interact and perceive different explanations, such as textual, visual, or a combination of text and visual (*i.e.*, hybrid) to foster appropriate trust.

We choose AI-based predictive policing as our use case because it represents a critical domain where appropriate trust is pivotal due to the high-stakes nature of decision-making. Furthermore, examining how different explanations impact appropriate trust in this context is highly practical as it informs the design and deployment of AI systems particularly in sensitive areas like law enforcement, promoting responsible and effective use. [257].

According to Ribera et al. [309], the goal of XAI depends not only on the presentation of explanations but also on the type of end-user that is on the receiving end of the explanations. For example, developers and AI experts might use the explanations to verify that the system is working as expected. Previous research has shown that the hybrid form of explanations significantly improves correct understanding for lay users compared to visual explanations [353]. However, little work has been done so far to compare the utility of different explanation methods in building appropriate trust between expert users and lay users. In the predictive policing use-case, this comparison is relevant as some police officers might have less professional experience to draw on than others, for example, police officers who have recently joined the department. Therefore, in this study, we will compare the utility of different types of explanations with both expert and lay users (*moderation factor*).

As we wanted to understand the moderating factor of user expertise in our study design, we perform a application-grounded evaluation followed by a human-grounded evaluation. This is in line with Doshi-Velez & Kim [97] (Fig. 1, Page 2) who identify the value of different layers of evaluation within XAI focusing on functionality, grounding and presentation. Our sample of police officers corresponds to application-grounded evaluation which includes real humans and real tasks. On the other hand, our sample of lay users corresponds to human-grounded evaluation which includes real humans and proxy tasks. Human-grounded evaluation is appealing when experiments with the target community is challenging to recruit to obtain the enough sample size [193, 209] which was the challenge for us in this study. Furthermore, we echo Doshi-Velez & Kim [97] recommendation that "*Just as one would be critical of a reliability-oriented paper that only cites accuracy statistics, the choice of evaluation should match the specificity of the claim being made.*". Our contribution is centered around a specific application warrants assessment within the framework application-grounded and human-grounded evaluation to answer our following research questions (RQ3 and RQ4).

We aim to address the following research questions:

- RQ1:** What effect do different types of explanations (no explanation, textual, visual or hybrid) have on building appropriate trust in AI-based predictive policing systems?
- RQ2:** How does human trust in the AI assistant change given these different types of explanations?
- RQ3:** Do lay users or experts find these explanations useful in making a decision?
- RQ4:** Is there a moderating effect of user expertise influence the role of explanations in establishing appropriate or subjective trust?

We investigated the first question by prompting users with a selection of hotspots for predictive policing that gauge their understanding of the explanation at hand and see whether they can appropriately trust the system. For measuring appropriate trust, we adopt several existing measures from the literature. The second question is addressed by asking participants to rate their perceived trust in the AI assistant, and the third by asking participants to rate the usefulness of the explanations over multiple rounds in our study. Finally, we address the last question by recruiting and comparing participants with different expertise (police officers and lay users) and studying the role of explanations in building appropriate and subjective trust.

**Original Contributions.** Through our work in this chapter, we make the following contributions:

1. We present the first study exploring the effect of different types of explanations on building appropriate trust in AI-based predictive policing systems.
2. We illustrate the effect of user expertise on different types of explanations for building appropriate trust.
3. By conducting two user studies ( $N=192$ ), we show how explanations can shape perceptions of the AI system and how participants often end up in the trap of confirmation biases.
4. Based on our results, we highlight research challenges and recommendations for the design of XAI-based predictive policing systems.

## 6.2 BACKGROUND AND RELATED WORK

### 6.2.1 APPROPRIATE TRUST

Appropriate Trust in AI systems has rapidly become an important area of focus for researchers and practitioners. As technology evolved from automated machines to decision aids, virtual avatars, robots, and AI teammates, appropriate trust has been studied in depth and breadth across various domains. Appropriate trust is often linked to the alignment between the perceived and actual performance of the system [406]. Mehrotra et al. argue that human trust in the AI system must be appropriate because, with appropriate trust in AI, people may be simultaneously aware of AI's potential and limitations [252]. This should reduce the harms and negative consequences of misuse and disuse of AI [217].

#### DEFINITIONS

It is important to understand how we define appropriate human trust in AI (Human-AI trust) when trying to achieve it. On the one hand, the definitions of appropriate trust are linked to system performance or reliability, such as:

1. Appropriate trust is the alignment between the perceived and actual performance of the system [243, 247].
2. Appropriate trust is to [not] follow a [in]correct recommendation. Other cases (where trust is not appropriate) lead to over-trust or under-trust [406].

3. In human-robot teaming, appropriate trust is maintained when the human uses the robot for tasks or sub-tasks the robot performs better or safer while reserving those aspects of the task the robot performs poorly to the human operator [282].

On the other hand, the definitions of appropriate trust are related to trustworthiness and beliefs such as:

1. Appropriate trust in teams happens when one teammate's trust towards another teammate corresponds to the latter's actual trustworthiness [114].
2. We can understand 'appropriate trust' as obtaining when the trustor has justified beliefs that the trustee has suitable dispositions [81].
3. Appropriate trust occurs when (i) the human estimates that the AI agent is better at the task than the human, (ii) also the actual TW of the AI agent is equal to or higher than the human's TW and (iii) the human selects the AI agent for the task and *vice-versa* (Chapter 5).

Inevitably, despite the crucial role of appropriate trust in ensuring the successful use of AI systems, there is currently a fragmented overview of its understanding. This conclusion resonates with Jacovi et al.'s overview [166], who calls for these definitions to be more precisely defined and differentiated.

### THE USE OF EXPLANATIONS FOR FOSTERING APPROPRIATE TRUST

A common method to achieve appropriate trust is by adding transparency to the system through explanations. Intuitively, this makes sense as understanding an AI system's inner workings and decision-making should, in theory, also allow a user to understand better when to trust or not trust a system to perform a task [252]. Explanations focus on the inner workings of AI systems, appearing either for every AI recommendation or under specific circumstances. It has been shown that an AI agent who displays its integrity in the form of explanations by being explicit about potential biases in data or algorithms achieved appropriate trust more often than being honest about capability or transparent about the decision-making process [252]. Also, Buçinca et al. have shown that explanations with 'cognitive forcing functions' were effective in trust calibration, as here the AI system adjusts to the user's attitude and behaviour following the signs of over- and under-trust [43]. Therefore, building on these prior works, we use explanations in our study to elicit appropriate trust.

### 6.2.2 AI-BASED PREDICTIVE POLICING & TRUST

AI-based predictive policing employs advanced algorithms and machine learning to analyze historical crime data, facilitating proactive crime prevention strategies [9, 139]. In light of the EU AI Act's regulatory framework, the use of predictive policing systems, which heavily relies on advanced data analysis methods, may come under scrutiny and be subject to compliance with the Act's provisions on high-risk AI applications in law enforcement [75]. This introduces a dimension of accountability and transparency in deploying predictive policing technologies within the legal and ethical parameters defined by the EU.

Currently, there are four main applications of predictive policing being used in European and American police departments: CAS (Crime Anticipation System) in the Netherlands<sup>1</sup>, ProMap and PredPol in UK<sup>2</sup>, and Soundthinking in the US<sup>3</sup>. Multiple studies have been conducted to explore the effectiveness of these applications and how police officers trust them [125, 240, 281]. For example, Meijer et al.'s study highlights two patterns of algorithmization of government bureaucracy - the 'algorithmic cage' (Berlin, more hierarchical control) and the 'algorithmic colleague' (Amsterdam, room for professional judgment) [258]. Specifically looking at trust, a study by Selten et al. shows that police officers trust and follow AI recommendations congruent with their intuitive professional judgment and that textual explanations do not affect trust in AI recommendations [332]. Building upon these prior works, we explore the role of different explanation types and user expertise on appropriate trust.

### 6.3 STUDY DESIGN

In our main user study, we sought to characterize the main and interaction effects of different explanations on appropriate and subjective trust and the role of user expertise. To understand the role of user expertise, we conducted two user studies. In the first user study, we recruited 12 ex-police officers from the Dutch police who retired in the last five years and had experience with predictive policing systems. We call this group of participants as "*Expert users*". In the second user study, we recruited 180 crowd-sourced workers without experience with predictive policing systems. We call this second group of participants "*Lay users*". As indicated in section 1, our choice of recruiting two different set of users is necessary to answer RQ3 & RQ4 and is situated in line with prior work of Doshi-Velez & Kim [97] focusing on application- and human-grounded evaluation given the challenges of recruiting experts sample.

Before data collection, we had preregistered our hypotheses, research design, and data analysis plan for the main study<sup>4</sup>.

#### 6.3.1 DESIGNING AI SYSTEM'S EXPLANATIONS

We conducted a preliminary study with three expert users to inform us about the design of explanations (Appendix C). Building on insights from this preliminary study, we sought to design our explanations as a combination of input-influence and case-based explanation types. Also, we added weather and escape route information to the explanation based on an insight which emerged during our preliminary study. Once we decided on the content of the explanations, we looked at established guidelines in the literature on crafting them. We selected the guidelines by Szymanski et al. [353] because they conducted a state-of-the-art literature survey and a formative study on XAI. The guidelines include (a) quantifying each parameter's contribution to an instance prediction, (b) what parameters lead to predictions, (c) finding instances that have similar predictions, (d) locating regions where model predictions are uncertain, and (e) displaying an overall predictions. To explain

<sup>1</sup><https://kombijde.politie.nl/vakgebieden/ict/predictiv>

<sup>2</sup><https://www.lawsociety.org.uk/topics/research/>

<sup>3</sup><https://www.soundthinking.com/law-enforcement/crime-analysis-crimetracer/>

<sup>4</sup>[https://osf.io/hka58/?view\\_only=a1cdf92f29d64e7289047c8a8da41855](https://osf.io/hka58/?view_only=a1cdf92f29d64e7289047c8a8da41855)

the predictive policing system’s decision in a textual form, we generated sentences per input parameter using the template described by Hohman et al. [151]:

*The system predicts a higher likelihood of incidents in hotspot A/ B/ C/ D based on (historical crime data OR proximity to dense forest/ highway/ sea OR last arrest of offenders) [A]. The X% confidence score reflects (strong/ weak) support [B], and the remaining X% acknowledges potential unknowns like X [C]. Major contributing factors to this decision include C1 (X%), C2 (X%), and C3 (X%) [D]. Furthermore, a similar case was found in X’s police records, where offenders were caught near X [E]. A strong/ no correlation with severe weather (snow or thunderstorms) was found while making this decision [D]. (Tip: Weather prediction for the next three days is X; allocate resources accordingly.*

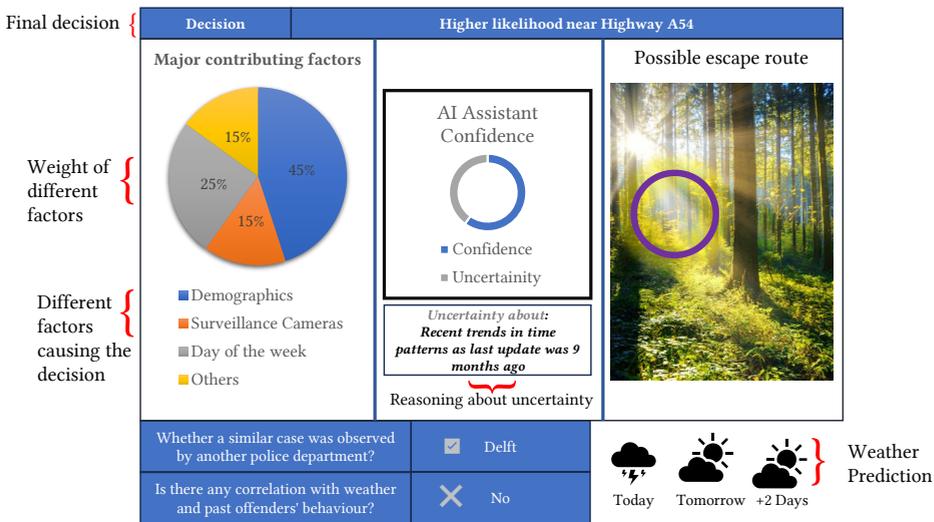


Figure 6.1: Visual explanations for a selected instance. The final decision, weight of different factors, reasoning about uncertainty and weather prediction corresponds to textual explanations.

Here, [A] denotes overall prediction, [B] denotes the confidence in the prediction, and [C] shows regions where the model prediction was uncertain. [D] quantifies each parameter’s contributions, and the name of the parameters, and [E] denotes instances that have similar predictions. The contributing factors were among the following: (C1) Historical crime data and incident reports, (C2) Geographical information, (C3) Time and day of the week, (C4) Weather information on the crime date, (C4) Demographic statistics, (C5) Resource availability and (C6) Socioeconomic data.

To enable a fair comparison, the visual explanations contained the same information as the textual explanations. Figure 6.1 shows an example of a visual explanation used in the study. Text annotations on this figure correspond to the bold alphabets [A...E] of the textual explanations. Finally, as prior research on designing hybrid explanations is limited [265], we based our design only on previous work done by Szymanski et al. [353], who combined visual explanations with text.

### 6.3.2 TRADITIONAL INVESTIGATION NOTES

Based on our preliminary study, police officers often follow traditional methods (diary notes, intel from other units and instructions from the department) for investigation in conjunction with predictive policing systems. On the one hand these notes serves as the ecological validity (in real life police officers often use their diary notes for investigation) for the task and on the other hand they make sure that there is a 'joint' knowledge for our both groups of participants. Furthermore, these notes provide evidence supporting the information presented and were added in the study to provide a reason to disagree with the AI and follow own intuitive judgement. An example of a note used in this study is as follows:

*You have (less than a year OR more than three years of experience) in this area shown on the map [A]. According to your diary notes, under the cover of darkness, the past offenders often slip through the labyrinth of narrow alleyways matching with the hotspot A/ B/ C/ D [B]. According to the intelligence department of the Police, the last fugitive vanished into the dense forest after following the alleyways [C].*

Here, [A] denotes the overall experience of the police officer in the selected area, [B] denotes diary notes and probable hotspot selection, and [C] shows the intel from the intelligence department of the police.

## 6.4 FIRST USER STUDY - "EXPERT USERS"

### 6

### 6.4.1 PARTICIPANTS

For our expert study, we recruited 12 retired police officers (aged between 65 and 70, 10M:2F) who retired in the last five years from the Dutch police. Our goal was to recruit experts who had prior experience with predictive policing systems or were in-charge of making decisions related to crime prevention. The retired police officers fulfilled these criteria as they were mostly in higher positions in their hierarchy before they retired. Furthermore, given the discussions around the new EU AI Act, police officers in the current force were hesitant to join our user study. Therefore, we decided to recruit retired police officers as they fit our goal of expert users. This study was approved by the Human Research Ethics Review Board of our university (IRB 2023-1779) and was conducted in the Dutch language.

### 6.4.2 METHODOLOGY

#### INDEPENDENT VARIABLE

- **Explanations** (*categorical, between-subjects*). We assigned each participant to one of four configurations:
  1. No explanation: participants saw hotspot selection by the AI assistant but not how this decision was made.
  2. Text-based Explanations: participants saw the hotspot selection by the AI assistant and how this decision was made in textual form as described in section 6.3.1.
  3. Visual Explanations: participants saw the hotspot selection by the AI assistant and how this decision was made in visual form as described in section 6.3.1.
  4. Hybrid (Text+Visual) Explanations: participants saw a combination of text-based and visual explanation.

**DEPENDENT VARIABLES.**

- **Appropriate Trust** (*continuous*). Adapted from [206, 252, 406, 417]. These measures are described in section 6.4.2.
- **Subjective Trust** (*continuous*). We used a global trust meter that captures changes in trust for each round ranged from completely distrust (-100) to completely trust (+100), adapted from [189, 252, 406].
- **Usefulness of Explanations** (*continuous*). The usefulness of explanations was measured on a 7-point Likert scale from Not at all helpful (1) to Very helpful (7), adapted from [406].

**DESCRIPTIVE AND EXPLORATORY MEASUREMENTS**

We use these variables to describe our sample and for exploratory analyses, but we do not conduct any conclusive hypothesis tests on them.

- **Age group** (*categorical*). Participants will select their age group from multiple choices.
- **Level of education** (*categorical*). Participants will select the highest level of education they have completed.
- **AI literacy** (*continuous*). Average score of the four items defined by [329].
- **Propensity to Trust** (*continuous*). Propensity to Trust scale by Merritt et al. [260] (adapted).
- **Personal experiences as a police officer and use of AI** a) Have you ever worked with predictive policing systems in the past? and b) Do you have prior experience with the use of AI in predictive policing systems?
- **Task stakes perception** (*continuous*). In this study we have considered scenario such as pick-pocketing as non-violent crime and sexual-offense as a violent crime based on the Dutch WODC Magazine Recidivism<sup>5</sup>. Since the stakes involved in a decision are subjective[186], we will capture task stakes perceptions using [228].
- **AI Confidence Score** (*categorical, within-subjects*) AI accuracy was communicated to participants as a part of the explanations. (1) High (*Confidence Score* > 75%) and (2) Low (*Confidence Score* < 75%) based on [416].
- **Geographical Experience** (*categorical, within-subjects*): Prior experience with policing about the shown geographic area on the map was communicated in the dairy notes. (1) Amount of professional experience : Limited (> 3 years experience) and Amount of professional experience : High (< 3 years experience).

<sup>5</sup><https://magazines.wodc.nl/wodcmagazine/2019/03/high-impact-crime-hic>

### MEASUREMENT OF APPROPRIATE TRUST

In this study, we used two measurements of appropriate trust and calibrated trust used in prior research. We used distinctive measures for appropriate and calibrated trust based on the definitions provided in the literature [89, 279, 402]. For example, Mehrotra et al. show that different definitions and measures of appropriate and calibrated trust exist in the literature [255]. We argue that it is necessary to study multiple measures to understand when trust can be classified as appropriate or calibrated, as different measures may result in slightly different outcomes. Our measures are:

1. **Measure 1 (App1):** Appropriate trust is to [not] follow an [in]correct recommendation [406].
2. **Measure 2 (App2):** Appropriate trust occurs when (a) the human estimates that the AI agent is better at the task than the human, (b) also the actual TW of the AI agent is equal to or higher than the human's TW and (c) the human selects the AI agent for the task and *vice-versa* [252].
3. **Measure 3 (Calib1):** Switch percentage, the percentage of trials in which the participant switched from their initial prediction to use the AI's prediction as their final prediction [417].
4. **Measure 4 (Calib2):** Agreement percentage, the percentage of trials in which the participant's final prediction agreed with the AI's prediction [206].

6

Table 6.1: Categorization of measures of appropriate trust

| Round | Human TW | AI TW | AI Correctness | Human Correctness |
|-------|----------|-------|----------------|-------------------|
| 1     | High     | High  | Correct        | Correct           |
| 2     | Low      | High  | Correct        | Incorrect         |
| 3     | Low      | Low   | Correct        | Incorrect         |
| 4     | High     | High  | Incorrect      | Correct           |
| 5     | High     | Low   | Correct        | Correct           |
| 6     | Low      | High  | Incorrect      | Incorrect         |
| 7     | Low      | Low   | Incorrect      | Correct           |
| 8     | High     | Low   | Incorrect      | Correct           |

For **App2**, we have classified the human expertise based on the provided notes (High or Low) and AI expertise based on the available data (High or Low). Furthermore, we have categorized the cases where AI's suggestion and human's diary notes correspond to the correct selection of the hotspot for predictive policing. The hotspot's correctness can never be proven in real life as it's more about the relative risk compared to other areas. Hence, to simplify, in our work the correctness of a hotspot was simply based on the permutations of trustworthiness and combinations of AI & Human correctness, refer to Table 6.1. As to **Calib1** and **Calib2**, the main difference between these two measures of calibrated trust was that the agreement percentage **Calib2** would count the trials in which the participants and the AI's predictions agreed and counted as the final decision. In contrast, the switch percentage **Calib1** would only consider cases where they disagreed and had to switch intentionally.

**PROCEDURE**

Our experiment aimed for human-in-the-loop collaboration, where participants made decisions assisted by an AI assistant. Participants, after providing informed consent and answering descriptive and exploratory questions, were introduced to the AI assistant. The AI assistant was hypothetically trained on the last ten years' crime data in the Netherlands. Finally, they were assigned to one of the four explanation types as per section 6.4.2.

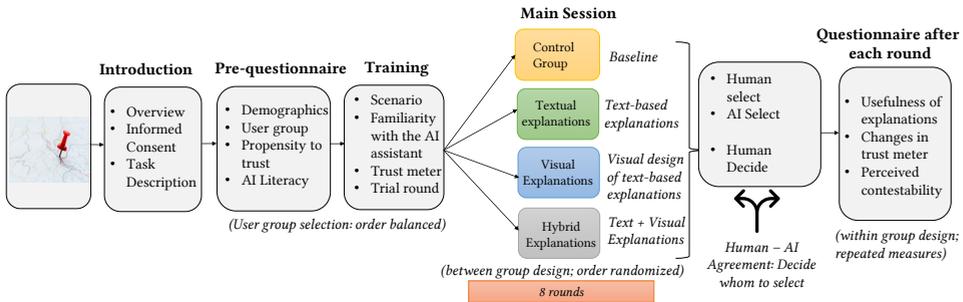


Figure 6.2: The experimental design of the user study. Each participant was assigned to a specific explanation type (Baseline - no explanation, Text, Visual and Hybrid) and they finished 8 rounds.

**Step 1: Trial Study** - Participants were tasked with choosing a hotspot for resource allocation, specifically for police patrol. They read diary notes (refer to section 6.3.2) for guidance and made their selection. The AI assistant then chose a hotspot based on hypothetical training data sourced from Het dataportaal van de politie <https://data.politie.nl/>, offering reasoning using one of four explanation configurations. Participants were then prompted to make a final selection, either affirming their or the AI assistant's choice or selecting a different hotspot and providing a reason.

**Step 2: Main Study** - After completing the trial round, participants received details about the main study, comprising eight rounds, each focusing on either violent (e.g., murder) or non-violent (e.g., pick-pocket) crimes. Participants had a limited 3-minute window for hotspot selection. They were instructed that the (*hypothetical*) intelligence unit chief would review their hotspot selections at the end of the experiment. Correct selections earned +10 points, while incorrect ones incurred a -10 point deduction. Due to time constraints, immediate result verification was not possible, prompting participants to proceed to the next round swiftly. Additionally, a bonus was promised for those achieving a top 3 score

**Step 3: End of the Study** - Participants completed the post-experiment questionnaire after the main study. They were asked to rate task stakes perception and how familiar they were with the geographical areas shown in the study. Furthermore, we also asked them two open-ended questions: (a) Do you think the AI assistant offered an appropriate explanation of the decision-making process? Why? If not, what explanation do you think it should have offered? and (b) How was your overall experience with this user study? Also, please elaborate on how the AI assistant's reasoning helped (or did not help) you to trust its decision. Once they answered these questions, they were shown their final score and whether they qualified among the top 3 scorers. A pilot study with five HCI researchers revealed no significant usability flaws, see Appendix C for details.

### 6.4.3 RESULTS

#### DESCRIPTIVE STATISTICS

Of the 12 participants in our user study, 10 had at least a Bachelor's degree. All of them claimed to have heard about predictive policing systems, whereas only two of them had heard of or had experience using AI in predictive policing. Our participants' average score of AI literacy was 6.8 (SD =  $\pm 1.1$ ) on a scale of 0 to 10, and their propensity to trust AI systems was 7.2 (SD =  $\pm 1.25$ ) on a scale of 0 to 10. The average duration of the study was 32 minutes (SD =  $\pm 3.25$ ), and each participant spent an average of 2 minutes 35 seconds (SD =  $\pm 0.60$ ) per round. Our analysis revealed no major differences in the frequency count of our measures of appropriate trust between the explanation types. However, subjective trust scores were comparatively higher for hybrid explanations (Mean = 60.98, SD = 5.62) when compared to all other explanation types (Mean = 41.23, SD = 10.83).

#### INFERENTIAL STATISTICS

We used Kendall's (tau) correlation to explore the association between our dependent variables. The correlation plot is shown in Appendix C, Figure C.1. Interestingly, there was no significant correlation between **App1** and **App2** measures, but **Calib1** and **Calib2** measures were significantly positively related. A possible explanation for this result is that both **Calib1** and **Calib2** measures look at the participant's final decision, whereas for **App1** and **App2**, the focus is on an accurate recommendation for the former and trustworthiness for the later. Additionally, we found subjective trust responses significantly positively correlated to both **Calib1** and **Calib2** and the usefulness of explanations measure. This means that as the perceived usefulness of an explanation increases, so does the level of trust that participants place in the system.

In addition to the above tests, we performed an exploratory analysis to understand our results better and identify potential interesting trends. Note that these are not confirmatory results as we did not preregister any analyses presented in this subsection. We performed correlation analysis on our exploratory variables. Our results indicate that the propensity to trust AI systems was positively significantly correlated with AI literacy ( $r=0.556$ ) and subjective trust ratings ( $r=0.255$ ). Interestingly, the propensity to trust AI systems was also positively significantly correlated with the task stakes perception. On adding our dependent variables to the analysis, we found no significant correlation between them and our exploratory variables, refer to the correlation matrix in Figure C.2, Appendix C for further details.

#### QUALITATIVE ANALYSIS

We first translated the transcripts in English with the help of two native bilingual speakers and performed qualitative analysis using a reflexive thematic analysis [39]. We inductively generated individual codes from our participants' responses to the open-ended questions and then clustered them into code groups. We identified two main areas: one related to explanation presentation & clarity, and the other related to perceptions of the AI system.

**Explanations Presentation and Clarity:** Overall, participants from the hybrid and textual explanations group found the explanations to be clear and structured. P7 (textual explanation) wrote, "*the use of public language rather than technical jargon helped decide to go with the AI assistant*". On the other hand, there were mixed reviews from participants for the visual explanations and no explanations categories. P11 (no explanations) expressed the

desire for more underpinning or context with the explanation. P5 (visual explanation) found visual explanations to be overwhelming. Also, 50% of participants wrote they followed their reasoning first and then looked at AI assistant's recommendation.

**Perceptions of the underlying AI system:** There were mixed reviews regarding the help provided by the AI assistant. From the far opposing end, there were concerns about the use of AI in predictive policing where P8 (baseline) wrote, "*the use of AI in predictive policing is fundamentally wrong because you cannot train a system to do policing*". Interestingly, we also found some quotes related to AI capabilities that supported P8's thinking. P12 wrote, "*AI does not possess human intuition and experience. Hence it cannot help in the way that my notes from my teams can*" and P4 wrote "*AI rarely captures the considerations of the perpetrator, which is important in understanding any crime as found discussed among officers*". Some police officers (P4, P6, P7, P11) found the AI assistant less appropriate than real police when its uncertainty was high. On the other end, participants appreciated the AI assistant's decision, when it aligned with the participant's understanding or AI provided additional information that was new to the participant. For example, P1 wrote, "*The information about the existing military unit was useful because it requires a cooperative operation then.*"

## 6.5 SECOND USER STUDY - "LAY USERS"

We conducted another user study to understand user expertise's role and assess explanations' role in fostering appropriate trust at a scale. We computed the required sample size using the software G\*Power [109] for an ANOVA with main effects and interactions, specifying the default effect size of 0.25, a significance threshold of  $\alpha = 0.05$ , a desired power of 0.8, 4 groups, and the respective degrees of freedom for the different hypotheses we aimed to test. The result indicated that we require approximately 179 participants.

We recruited 209 participants from *Prolific* (<https://prolific.co>). Each participant was at least 18 years old, highly proficient in English, and could participate in our study only once. Participants were rewarded based on a \$10 hourly rate, and the median completion time was 28 minutes and 11 seconds. Participants were excluded from data analysis if they did not pass at least one of the attention checks in the experiment. This led to 180 participants (age between 18 and 65+, 94M:86F), *i.e.*, 45 participants per explanation type. The study was conducted on *Qualtrics* (<https://www.qualtrics.com>), where participants authenticated with a registration token received on *Prolific*. This study was approved by our university's Human Research Ethics Review Board (IRB 2023-1780) and was conducted in English.

### 6.5.1 METHODOLOGY

Each participant had to follow the same methodology as the first user study (section 6.4.2) except that they also answered the following question - Have you worked for the police in the past, or are you currently working? In addition to the questions about personal experiences as a police officer and the use of AI (section 6.4.2). Concretely, this means we wanted to filter the participants who have worked for the police in the past or the present as they might have expert knowledge about predictive policing and would not classify as lay users.

### 6.5.2 RESULTS

### DESCRIPTIVE STATISTICS

Of the 180 participants in our user study, 29.44% were between 18 and 24 years old, 44.44% between 25 and 34 years old, 17.77% between 35 and 44 years old, and 8.33% were between 45-65+. 77% of the participants had at least a Bachelor's degree. None of them claimed neither to have ever worked for the police nor were they aware of the predictive policing system. The average score of AI literacy among participants was 4.67 (SD =  $\pm 1.25$ ), and their propensity to trust AI systems was 4.31 (SD =  $\pm 0.91$ ). The average duration of the study was 23 minutes (SD =  $\pm 4.25$ ), and each participant spent an average of 2 minutes 5 seconds (SD =  $\pm 1.03$ ) per round to make the final selection.

### INFERENCE STATISTICS

Before conducting any statistical analyses, we mapped all (seven-point) Likert scale answers onto an ordinal scale ranging from -3 (i.e., strongly disagree) to 3 (i.e., strongly agree). The result of Shapiro-Wilk shows that our data followed the normal distribution. Therefore, we conducted an ANOVA with explanations as between-subjects factors and different measures of appropriate trust as the dependent variables. We found no main effect of different explanation types on any measure of appropriate trust ( $p > 0.05$ ,  $\eta_p^2 < 0.01$ ).

We conducted another ANOVA with the same between-subjects factors but with subjective trust ratings and usefulness of explanations as the dependent variable. Next to the  $F$  statistic and  $p$ -value, we also report the partial eta squared  $\eta_p^2$  effect size. We found a significant difference between different explanation types on the perceived usefulness of the explanations ( $F(3,1436) = 4.35$ ,  $p < 0.005$ ,  $\eta_p^2 = 0.2$ ). The post hoc analysis revealed that hybrid ( $p < 0.013$ ) and visual explanations ( $p < 0.001$ ) were significantly better than no explanations for the usefulness of explanations ratings. However, we did not find any evidence indicating the effect of different explanation types on subjective trust responses ( $p = 0.479$ ,  $\eta_p^2 < 0.01$ ).

In addition to the analyses described above, we conducted multiple linear regression to analyze the association of independent and dependent variables and exploratory analyses to explore any trends in the data to better understand our results. Our results show that **App2** ( $\beta = 4.31$ ,  $p < 0.001$ ), **Calib2** ( $\beta = 8.59$ ,  $p < 0.001$ ) and AI assistant's correctness ( $\beta = 11.01$ ,  $p < 0.001$ ) predicted the measure **App1** ( $R^2 = 0.40$ , AIC = 1502, BIC = 1560), with AI assistant's correctness being the strongest predictor. Similarly, we found perceived usefulness of explanations ( $\beta = 2.16$ ,  $p < 0.001$ ) and AI assistant trustworthiness ( $\beta = 12.58$ ,  $p < 0.001$ ) predictors of **App2** ( $R^2 = 0.393$ , AIC = 1311, BIC = 1314) other than **App1**. Finally, we found **Calib2** ( $\beta = 9.34$ ,  $p < 0.001$ ) and perceived usefulness of explanations ( $\beta = 19.45$ ,  $p < 0.001$ ) predicted the subjective trust scores ( $R^2 = 0.267$ ). Finally, we did not find evidence of any exploratory variable affecting measures of appropriate trust.

### QUALITATIVE ANALYSIS

We followed a similar approach as in section 6.4.3 to perform our qualitative analysis. We identified two main topics of interest: evaluation of AI's reasoning and doubts about AI's effectiveness.

**Evaluation of AI's reasoning:** Participants, in general, had a positive attitude towards the AI assistant across all explanation types due to (a) lack of expertise for the task, P24 (no explanation): *"I think this system know what it is going, I just need to use it accordingly*

as this task is very new to me", (b) in-depth reasoning of the decision, P96 (textual explanation): "I believe various factors considered by the AI, such as historical crime data, weather, demographics, and spatial relationships are useful to decide.", and (c) breaking the tunnel vision, P77 (visual explanation): "I find visual information appealing and photos, maps, past crime patterns are right to the point, especially the link with weather is something I could never think off." Some participants expressed reassurance from AI's logical reasoning (P9, P23, P55, P149, P180) and expressed higher trust when their hotspot selection was similar to AI (P112, P106, P155, P47, P33).

**Doubts about AI's effectiveness:** Several participants (23%) expressed scepticism about the AI assistant's effectiveness irrespective of explanation types. They put forward the desire for consideration of (a) real-time factors (P25, P40, P164), (b) more transparency (P54, P109, P1172), (c) resolution of discrepancies between AI and personal judgement (P37, P111, P140), and (d) providing validation approaches for AI decision-making (P50, P74, P101). Furthermore, 12 participants reported that if the explanation was hard to understand and follow, they just followed the AI assistant's answer because it is too much work to determine whether the AI is right or wrong. For example, P78: "region around Assen is under control of military so how as a police officer can make any judgement, go with AI!"

## 6.6 DISCUSSION

### 6.6.1 EFFECT OF EXPLANATIONS ON APPROPRIATE AND SUBJECTIVE TRUST

Our findings show that explanation types, including 'no explanations' had no impact on any measures of appropriate trust in either user study (RQ1). To interpret these results, let's revisit the definitions of our measures, **App1** and **App2**. For **App1** to occur, participants must [not] follow [in]correct recommendations, and for **App2**, understanding both the trustor and trustee's trustworthiness is crucial. Our analyses indicate that, on average, participants correctly selected the hotspot four times in study 1 and three times out of eight rounds in study 2, suggesting a 50% error rate. Moreover, in user study 1, participants utilized the AI assistant to confirm their intuitive professional judgment rather than comparing trustworthiness, leading to a lack of substantial variations in explanatory formats. Therefore, regardless of expertise, participants failed to perceive meaningful distinctions in trustworthiness, leading us to conclude that there was no effect on appropriate trust, regardless of the type of explanation provided, including no explanation.

For RQ2, our findings indicated that hybrid explanations were significantly rated better on subjective trust than all other explanation types in study 1. However, this trend was not apparent in study 2. This result suggests potential variations in how different user groups perceive and respond to explanation types echoing the work by Szymanski et al. [353]. Several factors could contribute to this divergence. It is conceivable that the prior professional experience of retired police officers influenced their preference for hybrid explanations, given their familiarity with complex decision-making processes. In contrast, lay users in study 2 might have different expectations or preferences. Additionally, the significant positive correlation between (a) propensity to trust AI systems, (b) AI literacy and (c) subjective trust ratings which was only visible for retired police officers could have contributed to the observed differences.

### **INCREASE IN TRUST DOESN'T MEAN TRUST IS APPROPRIATE**

The finding that hybrid explanations were significantly rated higher on subjective trust in study 1 raises critical questions about the conventional assumption that higher trust is inherently positive. While the observed increase in trust suggests a positive perception of the hybrid explanations, the lack of a parallel effect on appropriateness challenges the simplistic notion that elevated trust equates to improved system performance. The danger lies in uncritically assuming that higher trust is universally advantageous [24, 252], as this study indicates that trust can rise without a corresponding improvement in appropriateness. Assuming higher trust is always better could inadvertently divert attention from the primary goal of ensuring the effectiveness and appropriateness of AI systems, particularly in sensitive domains like predictive policing. This calls for reevaluating trust as the sole metric of success.

We recognize that appropriate trust is a complex topic as it requires consideration of the influence of context, the goal-related characteristics of the agent, and the cognitive processes that govern the development and erosion of trust [61]. However, use of explanations to promote users trust in AI may not help them make better decisions. This calls for exploration whether alternative approaches, beyond explanations, should be explored or if the shortcomings lie in the quality of the provided explanations [24].

Advocates for diversifying appropriate trust-building methods argue for incorporating transparency, user engagement, or iterative feedback processes [114, 252]. Conversely, proponents of refining explanations contend that if appropriate trust is lacking, it might indicate a need for enhancing the clarity and quality of explanations rather than abandoning this approach [250]. We believe striking a more balanced approach should prioritize a holistic evaluation that includes appropriateness (goal), ensuring that any increase in trust aligns with the system's intended purpose (usability) and the user's needs (usefulness). We propose investigating Miller's approach using Evaluative AI which provides an alternative to XAI with hypothesis-driven decision support [263]. The author reports that evaluative AI approach helps in trust calibration because (a) it aligns with the cognitive decision-making process that people use when making judgements and decisions, and (b) there is no recommendation to follow which pushes users not to trust blindly.

6

### **6.6.2 USEFULNESS OF EXPLANATIONS, USER EXPERTISE & EXPLORATORY MEASURES**

In both of our user studies, we found a significant effect of the perceived usefulness of explanations on participants' subjective trust scores (RQ3). This result confirms that although explanations did not help build appropriate trust, participants found them helpful in making decisions and it increased subjective trust for expert users. Moreover, the significant prediction of subjective trust scores by the usefulness of explanations emphasizes the perceived quality of explanations in shaping users' overall subjective trust. This implies that users' trust assessments are not solely based on the accuracy of predictions but are also influenced by the perceived value of provided explanations.

Our results also did not indicate that user expertise moderated the role of explanations in building appropriate trust (RQ4). However, subjective trust of only expert users was significantly higher with hybrid explanations. This finding underscore the importance of considering user expertise in understanding human trust in AI systems as outlined by prior

studies [265, 309, 353]. Especially for this characterization, researchers already argued that the users are a key aspect to be considered when designing explanations [14, 48, 349]. Moreover, Gunning et al. [135] identified that a key challenge in XAI is tailoring the explanations to the expertise of end user.

A key notable difference between expert and lay users was the time spent on each round to decide. We found that most lay users followed the AI assistant advice as it is (42%), and expert users followed a more thorough, analytical approach as indicated in the open-text responses (75%). This observation can be explained by work from Wang et al. [388], who points out that inexperienced users are more susceptible to reinforcing effect. For example, when lay users interacted with unfamiliar information (rugged terrain), they often relied on the AI, which resulted in over-reliance.

Finally, on analyzing the role of our exploratory variables, we discovered a positive and significant correlation between propensity to trust AI systems and AI literacy, subjective trust scores, and crime classification by the AI system in study 1. However, none of these exploratory variables affected the dependent variables in study 2, suggesting contextual or participant-specific factors may contribute to the observed differences. For instance, participants in study 1, with a higher trust in AI, also exhibited greater AI literacy (average score of 6.80) compared to study 2 (average score of 4.67). Additionally, variations in the characteristics of the AI system related to predictive policing may have influenced variable relationships differently in study 2. Further exploration of these contextual differences could provide insights into the nuanced factors influencing trust and its correlations with various variables. Identified predictors of appropriate trust, including AI assistant correctness and trustworthiness, reveal commonalities in the importance of effective communication between AI systems and users across both expert and lay user contexts. These findings underscore the need for tailored approaches to communication irrespective of user expertise levels. Overall, recognizing the differences in expertise levels can guide the development of user-specific strategies, ensuring that AI technologies effectively meet the diverse needs of both experts and lay users.

### 6.6.3 POLICY IMPLICATIONS - CHALLENGES AND RECOMMENDATIONS

Fundamentally, the advantages of AI-based predictive policing lies in considering temporal and spatial dimensions, distinguishing it from traditional methods. However, certain challenges must be addressed first to utilize this technology fully. Here, we will first illustrate the challenges based on this study and propose recommendations to overcome them.

- (1) Both expert (retired police officers) and non-expert users (lay users) appear prone to **confirmation bias when interpreting AI recommendations**.
- (2) Both expert and non-expert users sometimes align with their **biased judgment**. This is especially worrisome as AI systems are prone to reproducing human decision-making biases and errors [332].
- (3) The use of explanations in predictive policing scenario can lead to **higher subjective trust but not appropriate** which do not necessarily help the users make better decisions.

Given these challenges, the outcomes of our studies carry significant implications for shaping policies surrounding responsible AI practices. The apparent finding that increase in subjective trust with the use of explanations does not lead to better decisions underscores

the importance of looking at alternatives beyond explanations. Policy initiatives should encourage a shift in the focus of AI system development and deployment towards fostering appropriate trust rather than solely prioritizing an increase in subjective trust through explanations. To achieve this, we advocate for following policy measures based on our findings:

- (1) **Performance Metrics that Reflect Decision Quality:** We encourage the adoption of evaluation metrics that go beyond subjective trust scores and incorporate measures of decision quality such as appropriate trust being the end-goal or examining usability. This ensures that the effectiveness of AI systems is assessed based on their impact on decision-making rather than just the perceived trust.
- (2) **Understanding user expertise:** Our study showed multiple differences between the two user groups. Therefore, we propose integrating user-centric design principles into policy frameworks, emphasizing the importance of tailoring explanations to meet the user expertise.
- (3) **Maintaining human discretion to overturn AI recommendations** - We found that expert users do not unquestioningly trust and follow all AI recommendations but weigh them against their professional knowledge. Therefore, evidence based expert judgment, such as the human-in-the-loop, should be included to override decisions. For example, the system was used less restrictively in the Netherlands and seen as a "helping hand". In contrast, in Germany, police officers could less easily divert from recommendations by the predictive policing system [258]

## 6

#### 6.6.4 LIMITATIONS AND FUTURE WORK

A major limitation of our study is the small sample size in Study 1, limiting the generalizability of findings. While our sampling method aligns with standards in predictive policing studies [124], the results may not represent the broader population. Future work should consider a larger sample size and a different XAI approach to better understand how to establish appropriate trust in such systems. Additionally, the hypothetical nature of the scenarios may not fully capture the complexities of real police decision-making. Our intentional use of scenarios requiring individual hotspot selection with an AI assistant contrasts with the likely team-based approach in reality. These scenarios are akin to those discussed by Ferguson [112] addressing issues related to the rise of "big data policing" who also points out to concern with the underlying data than trust in such systems.

Given the inherent limitations of the study, experimental conditions and fictitious vignettes cannot fully replicate authentic police decision-making, as key factors influencing complexity were not thoroughly delineated. Future research could enhance robustness by employing more rigorous methodologies, incorporating stronger manipulations, or utilizing real-world interventions. Exposure to police decision-making or activities using advanced technologies like virtual reality [244], or employing deliberative polling, represents potential avenues for more comprehensive investigations.

### 6.7 CONCLUSION

In this chapter, we looked at the effect of different type of explanations (text, visual and hybrid) and user expertise (retired police officers and lay users) on fostering appropriate and subjective trust in an AI-based predictive policing system. Our results show that hybrid

form of explanations raised subjective trust of expert users compared to lay users in the AI system. However, none of explanation types helped participants in forming appropriate trust in the system. We argue that this result of an increase in trust is worrisome, as it does not lead to better decisions. Based on these results, we highlight challenges in building appropriate trust in human-AI interaction, and propose important policy recommendations centered around fostering appropriate trust in AI-based predictive policing systems.

We hope this chapter will serve as a “call to action” for the AIES community to shift focus from the use of explanations for just promoting trust in AI systems to fostering appropriate trust instead.

## **6.8 IMPACT STATEMENTS**

### **6.8.1 ETHICAL CONSIDERATIONS STATEMENT**

In formulating our survey, we adhered to the ethical guidelines outlined in HCI methodology [42] to ensure the anonymity of participants and the protection of their data privacy. Furthermore, our study and data management plan received approval from the ethics committee at our university. Retired police officers were recruited through open advertisements on retired police officers’ online community channels. Lay participants were recruited via the Prolific platform and were compensated for their involvement. To ensure a representative sample, we established pre-screening criteria including country of residence, self-reported socio-economic status, ethnicity, and geographic location. No identifiable information was collected, and all responses were temporarily stored on a secure server. To prevent confusion regarding the predictive policing system, a statement was included at the conclusion of the study clarifying that the system described was only a prototype, albeit loosely based on existing systems. Additionally, it was emphasized that the research aimed to explore whether system explanations contribute to trust formation, and contact information for the lead author was provided.

We recognize that our study, which delves into the question of which types of explanations foster appropriate trust in a system, raises concerns regarding potential dual use. For instance, malicious entities could exploit our findings to manipulate trust, particularly by framing technical system functionalities in a way that appeals to cognitive biases and by highlighting their perceived “intelligence.” Moreover, these entities could utilize insights from our study to target specific demographics that appear more prone to trusting systems.

### **6.8.2 RESEARCHER POSITIONALITY STATEMENT**

We present our research team positionality according to the work by Doshi-Velez & Kim [97]. Our focus in this study is on exploring effect of different presentation type of explanations on appropriateness of trust. We used predictive policing system as a use-case in our study as such systems are already in use and it demonstrates a critical real life example of human decision-making involving high risk.

The first author of this work has a background in HCI and AI governance. He positions himself as an enthusiast of (mixed methods) research methodology. His research career has focused on understanding how people interact with technology, and how technology impacts human cognition. The second author of this work is an active police officer with 10+ years of experience working for the Dutch police in their software division. He positions

himself at the intersection of designing effective technological solutions for the police and understanding their impact on the society. The third author positions herself with the design of collaboration processes and practices for knowledge-intensive work using the capabilities of modern ICT. The fourth author positions himself in the domain of Crowd Computing and Human-Centered AI. He has worked extensively in exploring challenges towards building better AI systems and facilitating better reliance of humans on AI systems. The fifth author positions herself in conducting decades long research on computational trust and automated negotiation in multi-agent systems and human-agent teams. Finally, the sixth author has a background in cognitive artificial intelligence and HCI. She positions herself towards understanding how we can create technology which interacts with humans in a meaningful way.

### **6.8.3 ADVERSE IMPACT STATEMENT**

However, the ethical dilemmas associated with predictive policing and its impact on affected individuals are not the primary focus of this study. For readers interested in exploring these issues further, we direct them to refer to section 2.2. Our research ultimately advocates for a balanced, thoughtful, and well-informed approach to the utilization of predictive algorithms. This approach acknowledges the inherent biases present in technology and emphasizes the pivotal role of human agency and law enforcement in shaping the objectives of algorithms, guided by scientific evidence tailored to each unique technological application.

# IV

## CONCLUSIONS



# 7

## CONCLUSION

## 7.1 CONCLUSIONS

The research presented in this thesis focuses on designing AI systems to foster appropriate human trust. It uses three lenses, formal, social and application, to study and design appropriate trust in human-AI interactions. Starting with a formal lens, it provides a logical understanding of what we mean by appropriate trust as well as an understanding of trust in human-AI dyads, and formalizes the integrity of the AI systems (Chapter 2). Next, the formal lens was adopted to draw up a novel Belief, Intentions, and Actions (BIA) mapping to study commonalities and differences in the concepts related to appropriate trust based on a systematic review (Chapter 3).

The social lens was adopted to understand how a human and an AI agent value similarity influences a human's trust in that AI agent (Chapter 4). Also, within the social lens, we studied how expressing integrity through explanations can help build appropriate trust in AI systems, along with a measurement method for appropriate trust based on a specific task in human-AI interaction (Chapter 5). Finally, we adopted the application lens to study how user expertise and different styles of explanations affect user's appropriate trust in a specific application area – Predictive Policing (Chapter 6). The main research question of this thesis is:

**How can we design for appropriate human trust in human-AI interaction?**

From this main research question, five sub questions were derived. These questions focus on the different aspects of fostering appropriate human trust in the AI agent.

7

1. **RQ1:** How can we formally define appropriate human trust in human – AI interactions?
2. **RQ2:** What's state-of-the-art in fostering appropriate trust in AI systems?
3. **RQ3:** How does human and AI agent value similarity influences a human's trust in that AI agent?
4. **RQ4:** How does the expression of different principles of integrity through explanations affect the appropriateness of human's trust in the AI agent?
5. **RQ5:** How do different types of explanations can help build appropriate trust in a predictive policing AI system?

The chapters presented in this thesis was aimed to answer these research questions. The conclusions from these chapters are as follows:

### **RQ1: HOW CAN WE FORMALLY DEFINE APPROPRIATE HUMAN TRUST IN HUMAN – AI INTERACTIONS?**

To address the first research question, we posited that how trustworthy an AI agent is for a human and how a human trusts the AI agent (human's belief in the AI agent's trustworthiness) should be similar to get appropriate trust. When human's beliefs diverge, it can lead to either under-trust or over-trust in the AI agent. Based on the theoretical

constructs, we conceptualized appropriate trust through formalism of nested beliefs and proposed to view them in a human-AI agent dyadic relationship. Furthermore, we studied the salient role of not only mutual human – AI agent trust but also a dyadic relationship between the human and the AI agent, including belief formation, in shaping human trust. Finally, we also studied an AI agent’s integrity and benevolence towards a human for a specific task. Overall, we provided a formal understanding of what we mean by appropriate human trust in human–AI interactions and the factors affecting it, particularly the integrity of the AI agent.

### **RQ2: WHAT’S STATE-OF-THE-ART IN FOSTERING APPROPRIATE TRUST IN AI SYSTEMS?**

To answer the RQ2, we studied the state-of-the-art literature on building appropriate trust by examining its evolution, definitions, related concepts, measures, and methods. First, we provided an overview of the history of appropriate trust in automated systems. This overview highlighted the focus on studying the over- and under-trust in automation in the 1980s to the emergence of calibrated trust as a key research topic across various disciplines in the 2010s.

We proposed a Belief, Intentions, and Actions (BIA) mapping to highlight commonalities and differences between appropriate trust and their related concepts. Following this mapping, we presented the results of the systematic review, discussing different ways to measure appropriate trust – (a) perceived, (b) demonstrated, and (c) mixed measures; types of tasks used, approaches to building it – (a) transparency, (b) perception, (c) guidelines and (d) studying the continuum of trust, and results of the appropriate trust interventions – (a) positive and (b) negative. We identified three key challenges in studying appropriate trust (a) discord and diversity in concepts related to appropriate trust such as calibrated trust, justified trust, responsible trust etc., (b) a strong focus on appropriate trust in capability, leaving out other aspects of trust such as benevolence and integrity, and (c) issues involved in adequately measuring appropriate trust such as when to categorize trust is appropriate. Finally, we summarised our observations as current trends, potential gaps, and research opportunities for future work.

### **RQ3: HOW DOES HUMAN AND AI AGENT VALUE SIMILARITY INFLUENCE A HUMAN’S TRUST IN THAT AI AGENT?**

To answer the third research question this thesis presented a user-study with 89 participants who teamed up with five different AI agents, which were designed with varying levels of value similarity to that of the participants. In a within-subjects, scenario-based experiment, AI agents gave suggestions on what to do when entering the building to save a hostage. We analysed the agent’s scores on subjective value similarity, trust and qualitative data from open-ended questions.

Our results showed that agents rated as having more similar values also scored higher on trust, indicating a positive effect between the two. Furthermore, we found that 42% of the participants responses from the open-ended questions were related to common values between the participant and the AI agent they chose. With these results, we concluded that people base their trust judgments on whether they feel that the AI agent shares similar goals, thoughts, values, and opinions.

**RQ4: HOW DOES THE EXPRESSION OF DIFFERENT PRINCIPLES OF INTEGRITY THROUGH EXPLANATIONS AFFECT THE LEVEL OF APPROPRIATE TRUST IN THE AI AGENT?**

To answer this research question, we conducted study of how different integrity-based explanations made by an AI agent affect the appropriateness of trust of a human in that agent. To explore this, (1) we provided a formal definition to measure appropriate trust, and (2) presented a between-subject user study with 160 participants who collaborated with an AI agent in food calories estimation task.

Based on our results, we concluded that (a) an AI agent who displays its integrity by being explicit about potential biases in data or algorithms achieved appropriate trust more often compared to being honest about capability or transparent about the decision-making process, and (b) subjective trust builds up and recovers better with honesty-like integrity explanations. With this study, we contributed to the design of agent-based AI systems that guide humans to appropriately trust them.

**RQ5: HOW DO INTEGRITY-BASED EXPLANATIONS HELP BUILD APPROPRIATE TRUST IN A PREDICTIVE POLICING AI SYSTEM?**

To answer our final research question, RQ5, we designed three types of explanations provided by an AI agent to support police officers in predictive policing: (a) Text-based, (b) Visual, and (c) Hybrid: Text + Visual. We wanted to understand the impact different types of explanations have on establishing appropriate trust in an AI-based predictive policing system. Based on our results, we concluded that (a) with crowdsourced workers, there was no difference in the fostering appropriate trust with different types of explanations; however, AI correctness and trustworthiness was significant predictor of participants appropriately trusting the AI system output (b) with retired police officers, hybrid explanations helped build appropriate trust in the AI system, and AI correctness along with its trustworthiness were significant predictors of appropriate trust.

## 7.2 LIMITATIONS

To fully appreciate the findings presented in this thesis, it is important to consider the limitations of our overall method of approaching the RQs and the studies described. Two of the main limitations of our method of using three lenses are the complexity of synthesis and our reductionist perspective.

### 7.2.1 COMPLEXITY OF SYNTHESIS

Adopting the three lenses involved translating findings from one lens to another. This could have eventually resulted in majorly translating only the transferable findings and leaving behind the others. For example, firstly, we adopted the formal lens to define appropriate trust, and a vital attribute of that definition was the concept of nested beliefs. Secondly, when using the social lens in Chapter 4 we reworked the definition of appropriate trust from a specific angle in this paper. Our new definition was based on the context of a specific task type and involved comparing trustworthiness and subjective human trust. Here, we did not specifically use the attribute of nested beliefs due to the complexity of measurement and specificity of the task. Hence, the risk of overlooking nuanced aspects arose when attempting to synthesize findings across these diverse lenses, possibly limiting the depth of understanding and analysis.

### 7.2.2 REDUCTIONIST PERSPECTIVE

A reductionist perspective is another inherent limitation in our approach. By categorizing trust into three lenses, there is a risk of neglecting the intricate and context-dependent aspects of trust. Trust is a multifaceted concept influenced by a myriad of factors, and a reductionist approach may inadvertently oversimplify the richness of its dynamics, potentially hindering a comprehensive understanding of trust in the studied context. Consequently, this approach might not fully capture the richness and complexity of real-world scenarios, limiting the generalizability and applicability of the research findings. Furthermore, the lenses themselves might not exhaustively cover all relevant dimensions of human trust in AI interactions such as morality, trust dynamics etc., potentially omitting crucial factors that could impact the design for appropriate trust.

After discussing the limitations of our three lenses approach, we now draw upon the limitations of our methodology for each sub-research question one by one. For RQ1, the primary constraint lies in the assumption that the alignment between how trustworthy an AI agent is as perceived by a human and the actual trustworthiness of the agent is a sufficient criterion for establishing appropriate trust. This assumption oversimplifies the intricate nature of trust, potentially neglecting contextual and subjective factors that influence the human perception of trustworthiness. Furthermore, the computational derivation of an AI agent's integrity for a specific task and the exploration of psychological factors contributing to appropriate trust are based on specific methodologies and models.

For RQ2, while we included studies from computer science and engineering sciences, it is possible that some relevant studies from other disciplines were missed. Additionally, we only focused on studies published in English, which may have led to language bias. Also, our mapping to concepts related to appropriate trust based on beliefs, desires, and intentions is only one of many possible ways to organize such concepts under an umbrella. Finally, our search period only included papers from 2012 till June 2022 and the research on appropriate trust is growing at a faster pace. Therefore, papers which were published from June 2022 are missing from this review.

For RQ3, we investigated the effect of value similarity on trust with a risk-taking scenario of saving a hostage. We believe that further evaluation with more real-life examples would provide additional insights on the participant's trust. Additionally, although we cross-examined the participant's value profile with their responses to the value similarity questionnaire, we did not focus on their understanding of value-laden explanations. Also, we did not study the potential effect of cultural norms on our findings which could have resulted in different findings.

For RQ4, our work limits itself to exclusive decision making, which does not represent the full spectrum of possible human-AI interaction. Therefore, our findings may not generalize to scenarios such as human-AI teaming where the focus is more on collaboration. Additionally, in our definition of appropriate trust, we did not further explore the reasons for the selections made by the human. Furthermore, the style of explanation was controlled for in some way by having the same authors for all explanations, but here too, differences might exist between conditions. For instance, the 'fairness about risk' explanation might have been a little more technical, as it explained where in the process risks could come from (e.g. bias in training data). Finally, our explanations expressed the related principles of integrity in one specific way, and different methods of expressing these [353] might

have different effects on trust than what we found.

Finally for RQ5, a major limitation of our study was the small sample size in Study 1, limiting the generalizability of findings. While our sampling method aligns with standards in predictive policing studies [23], the results may not represent the broader population. Future work should consider a larger sample size and a different XAI approach to better understand how to establish appropriate trust in such systems. Additionally, the hypothetical nature of the scenarios may not fully capture the complexities of real police decision-making. Our intentional use of scenarios required hotspot selection by individual participants with an AI assistant, which contrasts with the likely team-based approach in reality. These scenarios are akin to those discussed by Ferguson [112] addressing issues related to the rise of "big data policing."

In synopsis, successfully answering our main RQ necessitates a clear comprehension of all lenses and how they can be effectively combined to provide a holistic view of the complex landscape of designing for appropriate human trust in human-AI interaction.

## 7.3 CONTRIBUTIONS

### 7.3.1 SCIENTIFIC

The main scientific contributions of this thesis lie in the insights gained into understanding what it means to be appropriately trusted. We adopted a formal, social and application lens model which provided us a way to study our main RQ covering the logical, empirical and application area oriented aspects of designing for appropriate trust in human-AI interaction. Moreover, the scientific contributions of this thesis extend to the delineation of how AI agents can guide users toward appropriate trust, shedding light on the practical implementation of appropriate trust-building mechanisms.

7

**Chapter 2: Formal Definition of Appropriate Trust** By establishing a formal definition of trust as a belief in directed mutual trustworthiness, we lay the groundwork for further research on trust in AI systems. This definition not only clarifies the conceptual framework but also provides a basis for empirical investigation. Understanding trust as a belief system that corresponds to actual trustworthiness sets the stage for developing AI systems capable of reasoning about trust and trustworthiness, thus fostering more effective collaboration and decision-making. Moving forward, researchers can leverage this definition to design AI systems that promote appropriate mutual trust, leading to safer and more efficient human-AI interactions.

**Chapter 3: Mapping of Appropriate Trust Concepts** Through our Belief, Intentions, and Actions (BIA) based mapping, we offer a comprehensive overview of how appropriate trust and related concepts are defined, quantified, and conceptualized in the existing literature. This mapping not only synthesizes current knowledge but also identifies gaps and areas for future exploration. By delineating similarities and differences in approaches to building appropriate trust, we facilitate a nuanced understanding of the field's landscape. Researchers can utilize this mapping to inform their theoretical frameworks, methodological choices, and empirical investigations, thus advancing the state of the art in trust research.

**Chapter 4: Impact of Value Similarity on Trust** Our user study demonstrates the positive effect of value similarity on trust in AI agents, highlighting the importance of aligning human and machine values. This finding underscores the significance of incorporating value-based reasoning into AI design, particularly in trust-critical contexts. By emphasizing the benefits of integrating value similarity despite potential development challenges, we encourage designers to prioritize ethical considerations in AI system development. This insight can inform the design of explanation and feedback mechanisms aimed at enhancing trust in AI technologies, ultimately helping in fostering ethical and trustworthy human-AI interactions.

**Chapter 5: Integrity based explanations for building Appropriate Trust** We proposed an approach for building appropriate trust in human – AI interactions through integrity laden explanations focusing on honesty, transparency, and fairness. We believe our research holds significance for two main reasons. Firstly, before we can investigate methods to establish suitable trust, we need to have a clear understanding of its meaning. Secondly, the potential for conveying integrity-related principles through explanations remains largely unexplored. Through our contributions, we aim to broaden our comprehension of fostering appropriate trust between humans and AI, which is vital for effective human-AI interaction.

**Chapter 6: Effect of Explanations on Trust Building** Our study explores the impact of different types of explanations on building appropriate trust in AI-based predictive policing systems. This research sheds light on the challenges and opportunities associated with explainable AI (XAI) in high-stakes domains, such as law enforcement. Firstly, the apparent finding that an increase in subjective trust with the use of explanations does not lead to better decisions underscores the importance of looking at alternatives beyond explanations. This could be linked to one who did not have the appropriate explanations in the appropriate form. Secondly, policy measures for trustworthy AI should focus on performance metrics that reflect decision quality, understanding user expertise, and maintaining human discretion to overturn AI recommendations.

Overall, this thesis makes noteworthy contributions in several dimensions. Firstly, by synthesizing insights from the three lenses, it provides a multidisciplinary perspective on appropriate trust in human-AI interaction. Secondly, this research offers practical implications for AI design, bridging the gap between theory and application. It goes beyond theoretical frameworks to identify specific appropriate trust-building mechanisms within AI systems, enhancing the field's practical relevance. The inclusion of empirical validation through human-subject studies adds robustness to the findings, emphasizing their real-world significance. Additionally, the thesis considers societal implications, addressing ethical considerations, biases in AI systems, and the broader societal impact of trust dynamics in human-AI interactions.

### 7.3.2 SOCIETAL

Aside from the scientific contributions, this thesis is also relevant for a broader societal audience. As this thesis explores the design of AI systems which humans can appropriately

trust, the main societal contributions are for AI systems designers and UX researchers. Also, the results of the use cases taken in this thesis are helpful for AI ethicists, policy makers and law enforcement agencies.

***AI systems designers and UX researchers*** This thesis offers practical guidance for AI system designers and UX researchers in enhancing the design and user experience of AI systems. By seamlessly aligning with key elements of the design cycle, we offer actionable insights for AI system designers and UX researchers at various stages of the design process. The formal definition of appropriate trust establishes a solid foundation during the initial phases, influencing the ideation and conceptualization of trust-related features. This definition informs early design decisions, setting the tone for how trust will be communicated and perceived throughout the user interface.

As the design cycle progresses, the comprehensive mapping of trust concepts aids researchers in identifying user preferences, refining interface elements, and ensuring that the design resonates with users' expectations. The emphasis on value similarity in the user study guides designers in implementing personalized features that enhance trust, integrating user feedback iteratively.

Additionally, the introduction of measurable constructs for appropriate trust and considerations of AI system integrity and explanations offers practical elements for the evaluation and refinement stages of the design cycle. Designers can use these constructs as measurable metrics to assess the effectiveness of appropriate trust-building features and refine their designs based on empirical insights gathered from user interactions.

Finally, addressing ethical implications and biases throughout the thesis influences the design cycle at every stage, prompting designers to incorporate ethical considerations from the outset and continually reassess the societal impact of their designs.

7

***AI Ethicists, Policy Makers and Law Enforcement Agencies*** The findings of Chapter 4 offer practical insights for AI ethicists grappling with the ethical dimensions of trust in artificial intelligence. Specifically, the identification of a positive correlation between perceived value similarity and trust in AI agents provides ethicists with a tangible dimension to consider when evaluating the ethical implications of AI design and deployment. Ethicists can use these findings to advocate for the integration of value-aligned design principles in AI systems, promoting the ethical imperative of building trustworthy technologies. Overall, the practical utility for AI ethicists lies in leveraging these insights to inform ethical frameworks, advocate for responsible AI design, and contribute to the ongoing discourse on fostering trust in artificial intelligence in an ethically sound manner.

Organizations operating at the intersection of AI and food nutrition can leverage the findings of Chapter 5 to enhance the practical implementation of AI systems in their domain. The formal definition introduced for measuring appropriate trust provides these organizations with a quantitative tool to assess and improve the effectiveness of their AI agents in cultivating trust among users. This formal method enables organizations to establish standardized metrics, facilitating a comprehensive evaluation of the impact their AI systems have on user trust. Furthermore, the insights from the user study offer practical guidance for refining the explanations provided by AI agents in food and nutrition-related tasks. By incorporating this approach, organizations enhance the trustworthiness of their

AI agents, especially in scenarios where accurate and reliable information about food and nutrition is critical.

Finally, the results of Chapter 6 hold substantial practical implications for law enforcement agencies such as police. First, by gaining clarity on how explanations impact trust, police departments can refine the design of AI systems to instil confidence and trust among the public. Second, depending on the effectiveness of each explanation type, law enforcement can strategically choose and optimize the format of explanations provided by predictive policing systems to enhance user understanding and trust. Third, police departments can use the obtained results to refine the accuracy of their AI models, address perceived contestability issues, and adjust task stakes to align with the community's expectations. Finally, by addressing concerns related to trust, explanation formats, and system parameters, law enforcement agencies can enhance transparency, reduce scepticism, and encourage collaboration between the community and the predictive policing systems.

## 7.4 FUTURE WORK

The first evident step of future work in this thesis is to employ the findings of various chapters in this thesis by the AI system designers and UX researchers in real world settings. AI system designers and UX researchers working within large organizations can test the findings on a larger scale. Going forward, future work can aim to refine the formalization and implement a method to evaluate the definition of appropriate trust. In particular, one can conduct user studies to evaluate our notions regarding beliefs about beliefs in an experimental setting. This can both help us understand how trust beliefs are formed in humans, and how agents can appropriately use these beliefs to improve teamwork. Another interesting direction for future research would be to study if the trust of the human in AI agent *a* is affected by another AI agent *b* working in a similar context.

As to the systematic review reported in Chapter 3 our mapping to concepts related to appropriate trust based on beliefs, desires, and intentions is only one of many possible ways to organize such concepts under an umbrella. Future research can focus on the development of a clear and concise mapping of definitions of appropriate trust and related concepts from a multidisciplinary perspective. Based on the identified research gaps in our study, future work should aim for (a) a clear definition of appropriate trust, (b) defining, distinguishing, treating and measuring concepts related to appropriate trust as independent concepts, and (c) focus on integrity and benevolence of the AI systems. Finally, future work should incorporate recent developments in the field of building appropriate trust with a special focus on Large Language Models.

Following up on Chapters 4 & 5, we have utilized situation vignettes to craft our explanations. In our work, custom-built explanations to highlight different principles related to integrity were better suited to our user study. More research on e.g. style of writing, length, etc. would be relevant for future studies to better understand how explanations focusing on integrity can be designed. Furthermore, our results suggest the importance of value similarity for trust which opens up the question of how we design agents to align with human values, or personalize them. Also, given the significant role of integrity-based explanations in building appropriate trust from Chapter 5, future work can focus on exploring how different types of integrity-based explanations can be tailored to various contexts and user-needs to further optimize trust calibration and decision-making.

This could involve personalizing the level of detail, focusing on specific integrity aspects (honesty, fairness, transparency) that resonate most with the user in a given scenario, or even exploring interactive explanations where users can delve deeper into aspects they find most critical for trust.

Similarly, following up on Chapter 6 future work can delve deeper into understanding why hybrid explanations resonated with expert users but failed to cultivate appropriate trust across the board. One avenue could be exploring the specific informational needs of different user groups. Did lay users misinterpret the hybrid explanations, or did they lack crucial details to truly assess the AI's capabilities? Additionally, research could investigate alternative explanation formats or interactive elements that cater to both lay and expert users, promoting a more nuanced understanding of the AI's strengths and limitations.

## 7.5 TAKE-HOME MESSAGE

Trust serves as a cornerstone of society, enabling cooperation, fostering relationships, and facilitating progress<sup>1</sup>. We trust people instinctively for social connection and because past experiences (conscious and subconscious) tell us trusting is generally helpful. Without trust, the intricate web of social interactions that sustains society would unravel, leading to chaos and instability. However, more trust is not necessarily better. There are hazards in both over-trusting and under-trusting. Hence, the need for designing for appropriate trust in AI is especially crucial because AI systems can be unpredictable at times.

In this thesis, we explored how we can design for appropriate trust in AI with the help of three lenses as mentioned previously. The key takeaways are:

1. An AI agent's perception of its trustworthiness towards a human and the human's actual trust in the agent should be aligned to build appropriate trust (Chapter 2).
2. There is a research gap in the literature focusing on integrity and benevolence which serves as two of the main three pillars for building trust (Chapter 2).
3. A comprehensive understanding of fostering appropriate trust in AI was lacking due to the diversity of perspectives arising from various backgrounds that influence it and the lack of single definitions of appropriate trust and related concepts (Chapter 3).
4. Value similarity between a human and an AI agent positively affects the human trust in the AI agent (Chapter 4).
5. Being explicit about integrity in communication in the form of explanations can help in achieving appropriate trust (Chapter 5)
6. An increase in a human's subjective trust in an AI agent does not necessarily mean an increase in the level of appropriate trust in that AI agent (Chapter 6).

---

<sup>1</sup><https://www.belfercenter.org/publication/ai-and-trust>

# V

## APPENDICES



A large, bold, black serif letter 'A' is centered on the page.

# EFFECT OF VALUE SIMILARITY ON TRUST IN HUMAN-AGENT INTERACTION

## **SUPPLEMENTARY**

The raw data set of this study along with the processed data files are available at <https://doi.org/10.4121/14518380>.

## A.1 ALGORITHM 1

---

**Algorithm 1:** Resolve conflicts in value profiles

---

**Input:**  $n$  = number of values in each group,  $i$  &  $j = 0$  and, number of groups ( $g$ ) = 0;

**Result:** Corrected value profile without conflicts

```

while  $n > 2$  &  $g < 5$  do
    combinations = fact( $n$ ) / (fact(2) * fact( $n - 2$ ));
    for ( $i = (combination-1)$ ;  $i \geq 1$ ;  $i-$ ) do
        for ( $j = 0$ ;  $j \leq (i-1)$ ;  $j++$ ) do
            if ( $List[j] == List[j+1]$ ) then
                user input to select a value;
                if ( $Selected\ Value == List[j]$ ) then
                    List [ $j$ ] -= 0.05;
                else
                    List [ $j+1$ ] -= 0.05;
                end
            end
        end
        end
        g++;
    end
end

```

---

## A.2 VSQ AND HCTS

**Value Similarity Questionnaire - [341]**

Scale: Totally Agree - Agree - Neutral - Disagree - Totally Disagree

- Do you think the Agent X acts as you would do in this scenario?
- Do you think Agent X thinks like you?
- Do you think Agent X shares your values?

**Human Computer Trust Scale - [134]:**

Scale: Totally Agree - Agree - Neutral - Disagree - Totally Disagree

- It is risky to interact with Agent A in this scenario. - *Willingness*
- Agent X will do its best to help you if you need help. - *Benevolence*
- If you take Agent X help, you would be able to depend on it. - *Trust*
- You can rely on Agent X in this scenario. - *Trust*
- You can trust the information presented to you by Agent X in this scenario.- *Trust*

# B

## INTEGRITY BASED EXPLANATIONS FOR FOSTERING APPROPRIATE TRUST IN AI

### B.1 APPENDIX 1

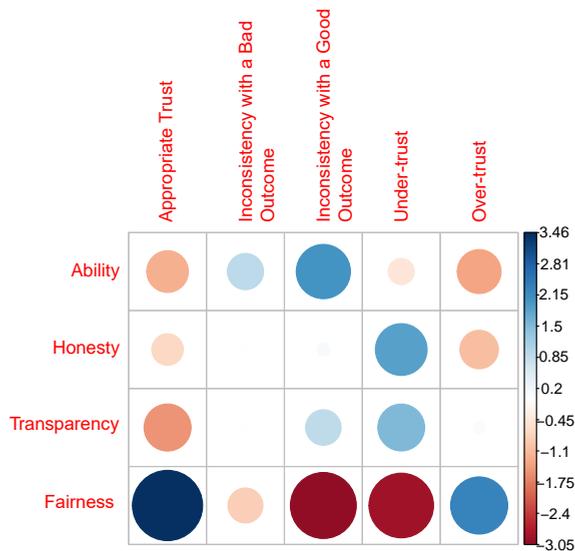


Figure B.1: A correlation plot between trust categories and integrity conditions. Positive residuals are in blue and specify an attraction (positive association). Negative residuals are in red implying a repulsion (negative association). The relative contribution of each cell to the total Chi-square score provides an indication of the nature of the dependency between trust categories and conditions.

B

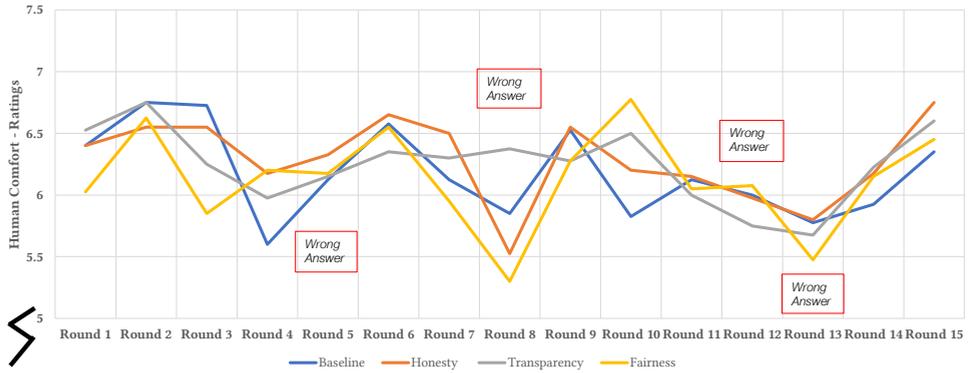


Figure B.2: Illustration of mean responses for changes in human comfort in decision-making ratings over 15 rounds. The red coloured boxes represents when the AI agent provided a wrong answer *i.e.*, round 5, 8, 12 and 13.

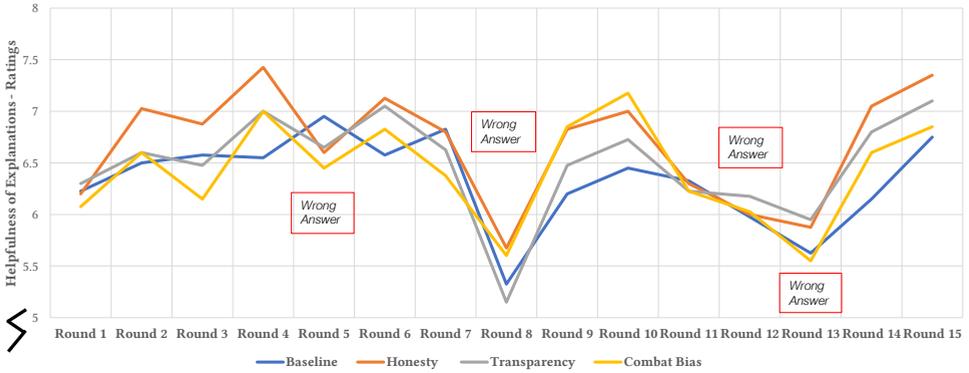
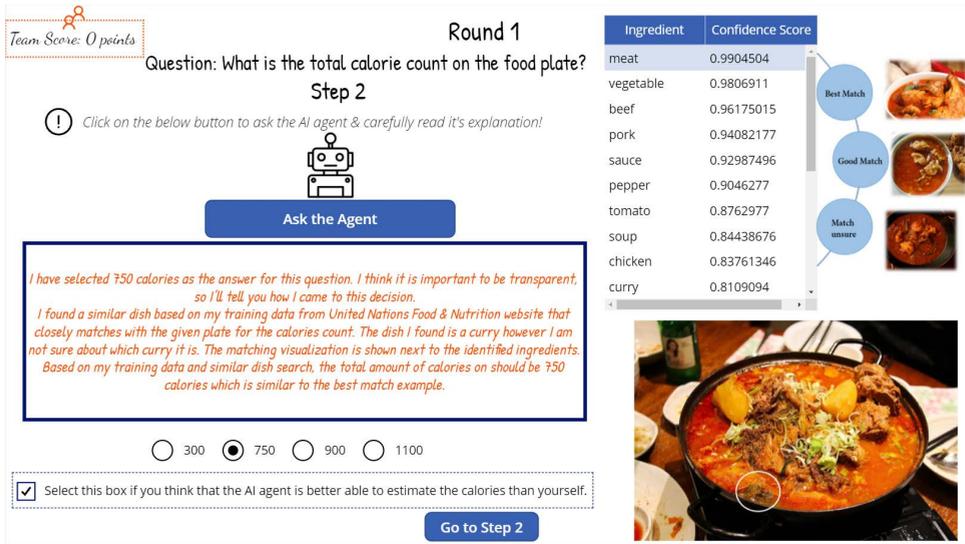


Figure B.3: Illustration of mean responses for changes in helpfulness of explanations ratings over 15 rounds. The red coloured boxes represents when the AI agent provided a wrong answer *i.e.*, round 5, 8, 12 and 13.



**B**

Figure B.4: Screenshot of transparency condition of the user study. This condition provided visualization of confidence scores in terms of best, good and an unsure match (refer top right corner).

## B.2 APPENDIX 2

The code of the developed experiment can be accessed at <https://doi.org/10.4121/bb0d42f4-a98d-4ae6-8043-2d3756b035ad>

## B.3 APPENDIX 3

| Model                        | Marginal R <sup>2</sup> | Conditional R <sup>2</sup> |
|------------------------------|-------------------------|----------------------------|
| Appropriate Trust            | 0.021                   | 0.082                      |
| Inconsistency (Bad outcome)  | 0.011                   | 0.098                      |
| Inconsistency (Good outcome) | 0.014                   | 0.032                      |
| Under trust                  | 0.028                   | 0.199                      |
| Over trust                   | 0.010                   | 0.848                      |

Table B.1: Marginal and Conditional R<sup>2</sup> values for Regression model of RQ1

|                      | Independent variables      | Coefficient |      | t value | Pr(> t ) | Significance |
|----------------------|----------------------------|-------------|------|---------|----------|--------------|
|                      |                            | $\beta$     | SE   |         |          |              |
| <b>Human Comfort</b> |                            |             |      |         |          |              |
| (Intercept)          |                            | 4.08        | 0.27 | 15.31   | <0.001   | ***          |
| Participants         | Round                      | 0.01        | 0.01 | 0.26    | 0.792    |              |
|                      | Fairness about bias        | -0.14       | 0.23 | -0.58   | 0.562    |              |
|                      | Honesty                    | -0.04       | 0.24 | -0.16   | 0.870    |              |
|                      | Transparency               | 0.00        | 0.23 | 0.02    | 0.985    |              |
|                      | Correct/Incorrect Answer   | -0.16       | 0.08 | -2.08   | 0.561    |              |
|                      | Lag Correct/Incorrect      | -0.28       | 0.08 | -1.71   | 0.069    |              |
|                      | Trust Score                | 0.01        | 0.00 | 1.40    | 0.162    |              |
|                      | Explanation Help           | 0.35        | 0.02 | 17.36   | <0.001   | ***          |
|                      | Marginal R <sup>2</sup>    |             |      |         | 0.150    |              |
|                      | Conditional R <sup>2</sup> |             |      |         | 0.394    |              |

Table B.2: Results of LMER analysis for RQ3 - Helpfulness of Explanations (\*:  $p < .05$ , \*\*:  $p < .01$ , \*\*\*:  $p < .001$ )

| Model   | AIC   | BIC   |
|---|-------|-------|
| Baseline  | 18617 | 18662 |
| Baseline+Lag  | 18566 | 18618 |
| Baseline+Lag+Interactions                             | 18565 | 18639 |
| Baseline+Lag+Interactions+Helpfulness of Explanations | 18481 | 18561 |

Table B.3: AIC &amp; BIC Statistics for the Regression Models of RQ2

| Model                     | Baseline   |            | Baseline+Lag |            |
|---------------------------|------------|------------|--------------|------------|
|                           | Chi Square | Pr(>Chisq) | Chi Square   | Pr(>Chisq) |
| Baseline+Lag+Interactions | 62.032     | <0.001     | 9.701        | 0.045      |

Table B.4: Regression Models Comparisons of RQ2

| Model   | AIC    | BIC    |
|---|--------|--------|
| <b>Appropriate Trust</b>                          |        |        |
| Baseline (Correct/Incorrect Answer+Lag)           | 3037.5 | 3077.5 |
| Baseline+Covariate 1 (Care about eating)          | 3039   | 3084.7 |
| Baseline+Covariate 2 (Propensity to Trust)        | 3039.3 | 3085   |
| Baseline+Covariate 3 (Usefulness of Explanations) | 3039.1 | 3084.8 |
| Baseline+Covariate 4 (Human Comfort)              | 3039.3 | 3085   |
| <b>Inconsistency with a bad outcome</b>           |        |        |
| Baseline (Correct/Incorrect Answer+Lag)           | 1140.7 | 1180.7 |
| Baseline+Covariate 1 (Care about eating)          | 1142.6 | 1188.3 |
| Baseline+Covariate 2 (Propensity to Trust)        | 1142.5 | 1188.2 |
| Baseline+Covariate 3 (Usefulness of Explanations) | 1139.9 | 1185.6 |
| Baseline+Covariate 4 (Human Comfort)              | 1142.7 | 1188.4 |
| <b>Inconsistency with a good outcome</b>          |        |        |
| Baseline (Correct/Incorrect Answer+Lag)           | 2653   | 2693   |
| Baseline+Covariate 1 (Care about eating)          | 2651.7 | 2697.4 |
| Baseline+Covariate 2 (Propensity to Trust)        | 2653.6 | 2699.3 |
| Baseline+Covariate 3 (Usefulness of Explanations) | 2654.9 | 2700.6 |
| Baseline+Covariate 4 (Human Comfort)              | 2654   | 2699.7 |
| <b>Undertrust</b>                                 |        |        |
| Baseline (Correct/Incorrect Answer+Lag)           | 1671.1 | 1711.1 |
| Baseline+Covariate 1 (Care about eating)          | 1672   | 1717.7 |
| Baseline+Covariate 2 (Propensity to Trust)        | 1673.1 | 1718.8 |
| Baseline+Covariate 3 (Usefulness of Explanations) | 1671.5 | 1717.2 |
| Baseline+Covariate 4 (Human Comfort)              | 1670.3 | 1716.1 |
| <b>Overtrust</b>                                  |        |        |
| Baseline (Correct/Incorrect Answer+Lag)           | 822.8  | 862.8  |
| Baseline+Covariate 1 (Care about eating)          | 822    | 867.7  |
| Baseline+Covariate 2 (Propensity to Trust)        | 824.4  | 870.1  |
| Baseline+Covariate 3 (Usefulness of Explanations) | 824.6  | 870.3  |
| Baseline+Covariate 4 (Human Comfort)              | 824.2  | 869.9  |

Table B.5: AIC & BIC Statistics for the Regression Models of RQ1. AIC is best for prediction as it is asymptotically equivalent to cross-validation. BIC is best for explanation as it allows consistent estimation of the underlying data generating process.



## C

## C

# FOSTERING APPROPRIATE TRUST IN AI-BASED PREDICTIVE POLICING SYSTEMS:

## C.1 PRELIMINARY STUDY

We conducted a preliminary study ( $n = 3$ ) aimed at: understanding how predictive policing works, what can we learn from currently used predictive policing applications?, and how can we design explanations for an AI-based predictive policing system? We also sought to understand what to consider in designing such systems to foster appropriate trust. Prior work has already studied predictive policing from the perspective of fairness [147] and identified policy recommendations [237]. However, the interplay between explanations and its effect in building appropriateness of trust still needs to be explored, hence the need to perform this preliminary study. The design of our preliminary study and the instruments we used to capture participants' preferences can be found in our repository.

### C.1.1 METHOD OF THE PRELIMINARY STUDY

As part of our preliminary study, specifically, a senior engineer (previously) involved in predictive policing system development, a police officer who chose to stay anonymous and a senior researcher working on predictive policing were informally interviewed in 2023. The main aim of the informal interviews was to gain additional information not available in the current predictive policing literature about the current applications of AI-based predictive policing and their practical implementation.

### C.1.2 INSIGHTS FROM THE PRELIMINARY STUDY

Based on our informal interviews, the current practices in predictive policing are characterized mainly by building a predictive model based on three phases: 1) data collection and preparation, (2) modelling, and (3) mapping. The constructed model is then applied to any crime type with known strong indicators for risk, where relevant data can be collected in advance. The effectiveness of such a system typically considers the accuracy of predictions,

changes in crime rates before and after implementation, and cost-effectiveness compared to traditional methods. Here, traditional methods include notes or intel a particular police unit has based on the crime. These notes usually contain the pattern followed by the past offenders, instructions from the intelligence department and individual diary notes.

As to the design of explanations, we received recommendations that (a) explanation design should map with correct understanding of the working of the overall model and (b) the relationship between specific input and the resulting output (*Global and Local Explanations*). Based on this insight, we provided five examples of different types of explanations to the participants (Importance-based [329], Input-influence [33], Case-based [96], Counterfactual [383] and a combination of counterfactual and input-influence [329]) following Yurrita et al. [410] work to understand our participants preferences.

All the participants agreed that explaining why and how some output is produced (input-influence-based) combined with how the output relates to a prior result (case-based) is most suited to predictive policing. P1 said, "*Let us say if your model gives output that hotspot C near the central train station should be considered for City A with reasoning and relate that output to City B, where a similar hotspot selection helped the police catch the criminal. It is beneficial!*" Finally, we got an insight to include how weather information can influence the hotspot selection (a higher probability of crime during heavy rain or snow as there are fewer people on the street) and a possible escape route the offenders can use (often near to highways) to include in the explanations.

## C.2 PILOT USABILITY STUDY

Before starting the main study, we wanted to make sure that our design did not contain usability flaws. To do so, we conducted a pilot user study with five HCI researchers from our lab (3M:2F, aged between 24 and 32). They followed the same methodology as the main user study described above, but additionally, they were asked to answer a SUS questionnaire [41]. The resulting score of this questionnaire was  $\mu_{sus} = 86$ ,  $SD_{sus} = 2.23$ , by which we can conclude that the system has no significant usability flaws following Bangor et al. [23].

## C.3 ADDITIONAL FIGURES

C

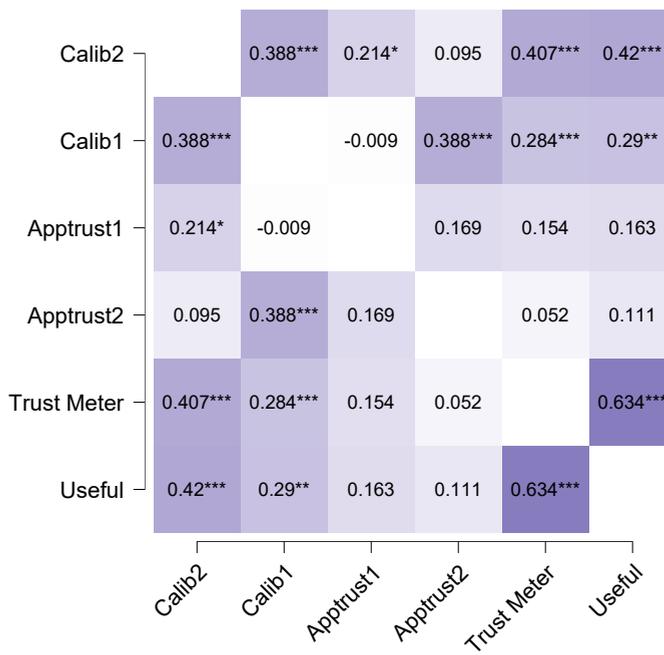


Figure C.1: Kendall tau correlation scores of various measures of appropriate trust from Study 1.

C

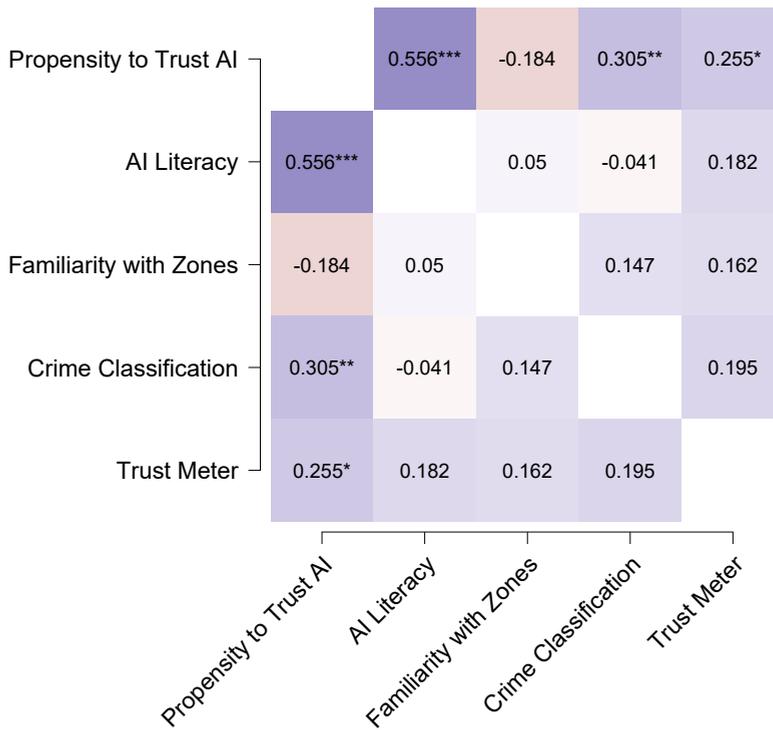


Figure C.2: Pearson's R heatmap of exploratory variables from Study 1.

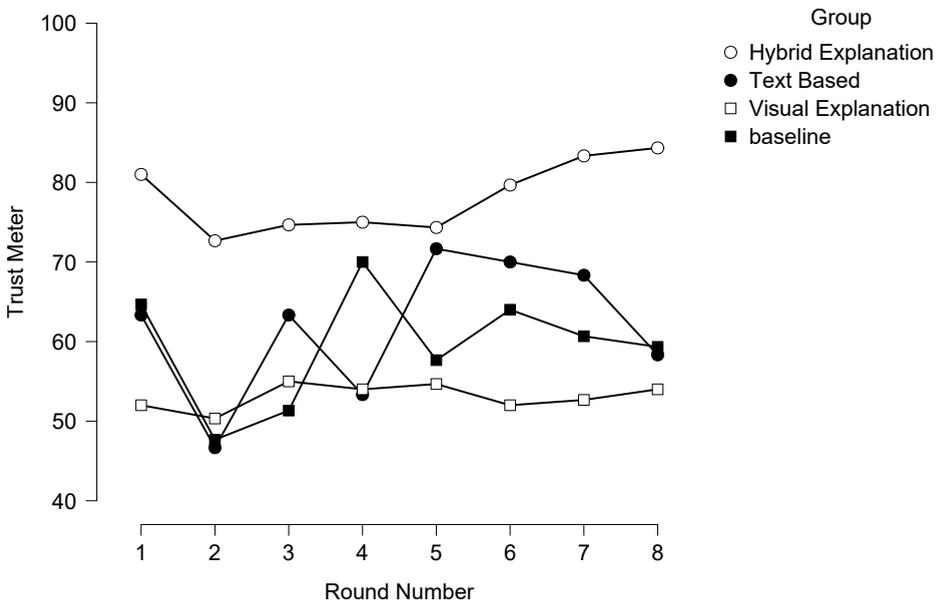


Figure C.3: An illustration of mean responses for changes in Global Trust Meter over 8 rounds.



# BIBLIOGRAPHY

## REFERENCES

- [1] *Ethically Aligned Design - A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 2, 2017*. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems., 2017.
- [2] Google pair. people + ai guidebook, May 2019. URL <https://pair.withgoogle.com/guidebook/>.
- [3] Alfarez Abdul-Rahman and Stephen Hailes. Supporting trust in virtual communities. In *Proceedings of the 33rd annual Hawaii international conference on system sciences*. IEEE, 2000.
- [4] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.
- [5] Kumar Akash, Neera Jain, and Teruhisa Misu. Toward adaptive trust calibration for level 2 driving automation. In *Proceedings of the 2020 international conference on multimodal interaction*, pages 538–547, 2020.
- [6] Arjun Akula, Shuai Wang, and Song-Chun Zhu. Cocox: Generating conceptual and counterfactual explanations via fault-lines. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2594–2601, 2020.
- [7] Yusuf Albayram, Theodore Jensen, Mohammad Maifi Hasan Khan, Md Abdullah Al Fahim, Ross Buck, and Emil Coman. Investigating the effects of (empty) promises on human-automation interaction and trust repair. In *Proceedings of the 8th International Conference on Human-Agent Interaction*, pages 6–14, 2020.
- [8] Basel Alhaji, Michael Prilla, and Andreas Rausch. Trust dynamics and verbal assurances in human robot physical collaboration. *Frontiers in Artificial Intelligence*, 4, 2021. ISSN 2624-8212. doi: 10.3389/frai.2021.703504. URL <https://www.frontiersin.org/articles/10.3389/frai.2021.703504>.
- [9] Kiana Alikhademi, Emma Drobina, Diandra Prioleau, Brianna Richardson, Duncan Purves, and Juan E Gilbert. A review of predictive policing from the perspective of fairness. *Artificial Intelligence and Law*, pages 1–17, 2022.
- [10] Fatemeh Alizadeh, Oleksandra Vereschak, Dominik Pins, Gunnar Stevens, Gilles Bailly, and Baptiste Caramiaux. Building appropriate trust in human-ai interactions. In *20th European Conference on Computer-Supported Cooperative Work (ECSCW 2022)*, volume 6, 2022.

- [11] James F Allen. Towards a general theory of action and time. *Artificial intelligence*, 23(2):123–154, 1984.
- [12] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. Guidelines for human-ai interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–13, 2019.
- [13] Lori Andrews and Hannah Bucher. Automating discrimination: Ai hiring practices and gender inequality. *Cardozo L. Rev.*, 44:145, 2022.
- [14] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58: 82–115, 2020.
- [15] Onur Asan, Alparslan Emrah Bayrak, Avishek Choudhury, et al. Artificial intelligence and human trust in healthcare: focus on clinicians. *Journal of medical Internet research*, 22(6):e15154, 2020.
- [16] Arvind Ashta and Heinz Herrmann. Artificial intelligence and fintech: An overview of opportunities and risks for banking, investments, and microfinance. *Strategic Change*, 30(3):211–222, 2021.
- [17] Giselle A Auger. Trust me, trust me not: An experimental analysis of the effect of transparency on organizations. *Journal of Public Relations Research*, 26(4):325–343, 2014.
- [18] Jackie Ayoub, Lilit Avetisyan, Mustapha Makki, and Feng Zhou. An investigation of drivers’ dynamic situational trust in conditionally automated driving. *IEEE Transactions on Human-Machine Systems*, 52(3):501–511, 2021.
- [19] Hebert Azevedo-Sa, Suresh Kumaar Jayaraman, X Jessie Yang, Lionel P Robert, and Dawn M Tilbury. Context-adaptive management of drivers’ trust in automated vehicles. *IEEE Robotics and Automation Letters*, 5(4):6908–6915, 2020.
- [20] Alan D Baddeley, Neil Thomson, and Mary Buchanan. Word length and the structure of short-term memory. *Journal of verbal learning and verbal behavior*, 14(6):575–589, 1975.
- [21] Annette Baier. Trust and antitrust. *Ethics*, 96(2):231–260, 1986.
- [22] Lisanne Bainbridge. Ironies of automation. In *Analysis, design and evaluation of man-machine systems*, pages 129–135. Elsevier, 1983.
- [23] Aaron Bangor, Philip T Kortum, and James T Miller. An empirical evaluation of the system usability scale. *Intl. Journal of Human-Computer Interaction*, 24(6):574–594, 2008.

- [24] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2021.
- [25] Gagan Bansal, Alison Marie Smith-Renner, Zana Buçinca, Tongshuang Wu, Kenneth Holstein, Jessica Hullman, and Simone Stumpf. Workshop on trust and reliance in ai-human teams (trait). In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, Chi Ea '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391566. doi: 10.1145/3491101.3503704. URL <https://doi.org/10.1145/3491101.3503704>.
- [26] Gabriel Diniz Junqueira Barbosa, Dalai dos Santos Ribeiro, Marisa do Carmo Silva, Hélio Lopes, and Simone Diniz Junqueira Barbosa. Investigating the relationships between class probabilities and users' appropriate trust in computer vision classifications of ambiguous images. *Journal of Computer Languages*, 72:101149, 2022.
- [27] David Barnard. Vulnerability and trustworthiness: Polestars of professionalism in healthcare. *Cambridge Quarterly of Healthcare Ethics*, 25(2):288–300, 2016.
- [28] Jeff A Bauhs and Nancy J Cooke. Is knowing more really better? effects of system development information in human-expert system interactions. In *Conference Companion on Human Factors in Computing Systems*, pages 99–100, 1994.
- [29] Max H Bazerman and Don A Moore. *Judgment in managerial decision making*. John Wiley & Sons, 2012.
- [30] Tom L Beauchamp. Moral prejudices: Essays on ethics. *The Hastings Center Report*, 25(4):36–37, 1995.
- [31] Izak Benbasat and Weiquan Wang. Trust in and adoption of online recommendation agents. *Journal of the association for information systems*, 6:4, 2005.
- [32] Rajeev Bhattacharya, Timothy M Devinney, and Madan M Pillutla. A formal model of trust based on outcomes. *Academy of management review*, 23(3):459–472, 1998.
- [33] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 'it's reducing a human being to a percentage' perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 Chi conference on human factors in computing systems*, pages 1–14, 2018.
- [34] Jens Blechert, Adrian Meule, Niko A. Busch, and Kathrin Ohla. Food-pics: an image database for experimental research on eating and appetite. *Frontiers in Psychology*, 5: 617, 2014. ISSN 1664-1078.
- [35] Philip Bobko, Leanne Hirshfield, Lucca Eloy, Cara Spencer, Emily Doherty, Jack Driscoll, and Hannah Obolsky. Human-agent teaming and trust calibration: a theoretical framework, configurable testbed, empirical illustration, and implications for the development of adaptive systems. *Theoretical Issues in Ergonomics Science*, pages 1–25, 2022.

- [36] Tibor Bosse, Catholijn M Jonker, Jan Treur, and Dmytro Tykhonov. Formal analysis of trust dynamics in human and software agent experiments. In *International Workshop on Cooperative Information Agents*, pages 343–359. Springer, 2007.
- [37] Ryan Boyd, Steven Wilson, James Pennebaker, Michal Kosinski, David Stillwell, and Rada Mihalcea. Values in words: Using language to evaluate and understand personal values. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9, 2015.
- [38] Michael Bratman. *Intention, Plans, and Practical Reason*. Cambridge: Cambridge, MA: Harvard University Press, 1987.
- [39] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.
- [40] Virginia Braun and Victoria Clarke. One size fits all? what counts as quality practice in (reflexive) thematic analysis? *Qualitative research in psychology*, pages 1–25, 2020.
- [41] John Brooke. Sus: a “quick and dirty” usability. *Usability evaluation in industry*, 189(3):189–194, 1996.
- [42] Amy Bruckman. Research ethics and hci. *Ways of Knowing in HCI*, pages 449–468, 2014.
- [43] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. To trust or to think: Cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), apr 2021. doi: 10.1145/3449287. URL <https://doi.org/10.1145/3449287>.
- [44] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(Cscw1):1–21, 2021.
- [45] Justin B Bullock. Artificial intelligence, discretion, and bureaucracy. *The American Review of Public Administration*, 49(7):751–761, 2019.
- [46] Meghan Madhavi Burke. Shraddha: A special kind of trust - healing arts centre, 2016. URL <http://healingartscentre.net/shraddha-a-special-kind-of-trust/>.
- [47] Chris Burnett, Timothy J. Norman, and Katia Sycara. Stereotypical trust and bias in dynamic multiagent systems. *ACM Transactions on Intelligent Systems and Technology*, 4, 3 2013. ISSN 21576904. doi: 10.1145/2438653.2438661.
- [48] Margaret Burnett. Explaining ai: fairly? well? In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 1–2, 2020.
- [49] Adrian Bussone, Simone Stumpf, and Dympna O’Sullivan. The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics*, pages 160–169. IEEE, 2015.

- [50] John K Butler Jr. Toward understanding and measuring conditions of trust: Evolution of a conditions of trust inventory. *Journal of Management*, 17(3):643–663, 1991.
- [51] Francesca Cabiddu, Ludovica Moi, Gerardo Patriotta, and David G Allen. Why do users trust algorithms? a review and conceptualization of initial trust and trust over time. *European management journal*, 40(5):685–706, 2022.
- [52] Carrie J Cai, Jonas Jongejan, and Jess Holbrook. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th international conference on intelligent user interfaces*, pages 258–262, 2019.
- [53] Davide Calvaresi, Kevin Appoggetti, Luca Lustrissimini, Mauro Marinoni, Paolo Sernani, Aldo Franco Dragoni, and Michael Schumacher. Multi-agent systems’ negotiation protocols for cyber-physical systems: Results from a systematic literature review. *Icaart (1)*, pages 224–235, 2018.
- [54] Christiano Castelfranchi and Rino Falcone. *Trust theory: A socio-cognitive and computational model*, volume 18. John Wiley & Sons, 2010.
- [55] Cristiano Castelfranchi and Rino Falcone. Principles of trust for mas: Cognitive anatomy, social importance, and quantification. In *Proceedings International Conference on Multi Agent Systems (Cat. No. 98EX160)*, pages 72–79. Ieee, 1998.
- [56] Cristiano Castelfranchi and Rino Falcone. Trust is much more than subjective probability: Mental components and sources of trust. In *Proceedings of the 33rd annual Hawaii international conference on system sciences*, pages 10–pp. IEEE, 2000.
- [57] Cristiano Castelfranchi and Rino Falcone. *Trust & Self-Organising Socio-technical Systems*. Springer International Publishing, 2010. doi: 10.1007/978-3-319-29201-4\_8.
- [58] Cristiano Castelfranchi and Rino Falcone. Trust & self-organising socio-technical systems. In *Trustworthy Open Self-Organising Systems*, pages 209–229. Springer, 2016.
- [59] Alain Chavaillaz, David Wastell, and Jürgen Sauer. System reliability, performance and trust in adaptable automation. *Applied Ergonomics*, 52:333–342, 2016.
- [60] Gilad Chen and John E Mathieu. Goal orientation dispositions and performance trajectories: The roles of supplementary and complementary situational inducements. *Organizational behavior and human decision processes*, 106(1):21–38, 2008.
- [61] Jessie YC Chen and Michael J Barnes. Human–agent teaming for multirobot control: A review of human factors issues. *IEEE Transactions on Human-Machine Systems*, 44(1):13–29, 2014.
- [62] Jessie YC Chen, Shan G Lakhmani, Kimberly Stowers, Anthony R Selkowitz, Julia L Wright, and Michael Barnes. Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical issues in ergonomics science*, 19(3):259–282, 2018.

- [63] Jing Chen, Scott Mishler, Bin Hu, Ninghui Li, and Robert W Proctor. The description-experience gap in the effect of warning reliability on user trust and performance in a phishing-detection context. *International Journal of Human-Computer Studies*, 119: 35–47, 2018.
- [64] Jing Chen, Scott Mishler, and Bin Hu. Automation error type and methods of communicating automation reliability affect trust and performance: An empirical study in the cyber domain. *IEEE Transactions on Human-Machine Systems*, 51(5): 463–473, 2021.
- [65] Kinzang Chhogyal, Abhaya Nayak, Aditya Ghose, and Hoa K. Dam. A value-based trust assessment model for multi-agent systems. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 194–200. International Joint Conferences on Artificial Intelligence Organization, 7 2019. doi: 10.24963/ijcai.2019/28. URL <https://doi.org/10.24963/ijcai.2019/28>.
- [66] Erin K Chiou and John D Lee. Trusting automation: Designing for responsivity and resilience. *Human factors*, 65(1):137–165, 2023.
- [67] Jin-Hee Cho, Kevin Chan, and Sibel Adali. A survey on trust modeling. *ACM Computing Surveys (CSUR)*, 48(2):1–40, 2015.
- [68] Sanghyun Choo and Chang S Nam. Detecting human trust calibration in automation: a convolutional neural network approach. *IEEE Transactions on Human-Machine Systems*, 52(4):774–783, 2022.
- [69] Lara Christoforakos, Alessio Gallucci, Tinatini Surmava-Große, Daniel Ullrich, and Sarah Diefenbach. Can robots earn our trust the same way humans do? a systematic exploration of competence, warmth, and anthropomorphism as determinants of trust development in hri. *Frontiers in Robotics and AI*, 8:640444, 2021.
- [70] Andy Cockburn and Carl Gutwin. A predictive model of human performance with scrolling and hierarchical lists. *Human-Computer Interaction*, 24(3):273–314, 2009.
- [71] Marvin S Cohen, Raja Parasuraman, and Jared T Freeman. Trust in decision aids: A model and its training implications. In *Proceedings of the 1998 Command and Control Research and Technology Symposium*, pages 1–37. CCRP Washington, DC, 1998.
- [72] Marvin S Cohen, Raja Parasuraman, and Jared T Freeman. Trust in decision aids: A model and its training implications. In *in Proc. Command and Control Research and Technology Symp*, 1998.
- [73] Robin Cohen, Mike Schaekermann, Sihao Liu, and Michael Cormier. Trusted ai and the contribution of trust modeling in multiagent systems. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1644–1648. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

- [74] Michael G Collins and Ion Juvina. Trust miscalibration is sometimes necessary: An empirical study and a computational model. *Frontiers in Psychology*, 12, 2021.
- [75] EU Commission. Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. proposal for a regulation of the european parliament and of the council, 2021.
- [76] Sven Coppers, Davy Vanacken, and Kris Luyten. Fortniot: Intelligible predictions to improve user understanding of smart home behavior. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(4):1–24, 2020.
- [77] Stephen Cranefield, Michael Winikoff, Virginia Dignum, and Frank Dignum. No pizza for you: Value-based plan selection in bdi agents. In *IJCAI*, pages 178–184, 2017.
- [78] Caterina Cruciani, Anna Moretti, and Paolo Pellizzari. Dynamic patterns in similarity-based cooperation: An agent-based investigation. *Journal of Economic Interaction and Coordination*, 12(1):121–141, 2017.
- [79] Oliver Daniels-Koch and Rachel Freedman. The expertise problem: Learning from specialized feedback. *arXiv preprint arXiv:2211.06519*, 2022.
- [80] David Danks. The value of trustworthy ai. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, Aies '19, page 521–522, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450363242. doi: 10.1145/3306618.3314228. URL <https://doi-org.tudelft.idm.oclc.org/10.1145/3306618.3314228>.
- [81] David Danks. The value of trustworthy ai. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 521–522, 2019.
- [82] Aritra Dasgupta, Joon-Yong Lee, Ryan Wilson, Robert A Lafrance, Nick Cramer, Kristin Cook, and Samuel Payne. Familiarity vs trust: A comparative study of domain scientists' trust in visual analytics and conventional analysis methods. *IEEE transactions on visualization and computer graphics*, 23(1):271–280, 2016.
- [83] Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. In *Ethics of data and analytics*, pages 296–299. Auerbach Publications, 2018.
- [84] Ewart de Visser and Raja Parasuraman. Adaptive aiding of human-robot teaming: Effects of imperfect automation on performance, trust, and workload. *Journal of Cognitive Engineering and Decision Making*, 5(2):209–231, 2011.
- [85] Ewart de Visser, Brian Kidwell, John Payne, Li Lu, James Parker, Nathan Brooks, Timur Chabuk, Sarah Spriggs, Amos Freedy, Paul Scerri, and Raja Parasuraman. Best of both worlds: Design and evaluation of an adaptive delegation interface. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 57(1): 255–259, 2013. doi: 10.1177/1541931213571056. URL <https://doi.org/10.1177/1541931213571056>.

- [86] Ewart J de Visser, Frank Krueger, Patrick McKnight, Steven Scheid, Melissa Smith, Stephanie Chalk, and Raja Parasuraman. The world is not enough: Trust in cognitive agents. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 56, pages 263–267. Sage Publications Sage CA: Los Angeles, CA, 2012.
- [87] Ewart J. de Visser, Marvin Cohen, Amos Freedy, and Raja Parasuraman. A design methodology for trust cue calibration in cognitive agents. In Randall Shumaker and Stephanie Lackey, editors, *Virtual, Augmented and Mixed Reality. Designing and Developing Virtual and Augmented Environments*, pages 251–262, Cham, 2014. Springer International Publishing. ISBN 978-3-319-07458-0.
- [88] Ewart J De Visser, Samuel S Monfort, Ryan McKendrick, Melissa AB Smith, Patrick E McKnight, Frank Krueger, and Raja Parasuraman. Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, 22(3):331, 2016.
- [89] Ewart J De Visser, Marieke MM Peeters, Malte F Jung, Spencer Kohn, Tyler H Shaw, Richard Pak, and Mark A Neerincx. Towards a theory of longitudinal trust calibration in human–robot teams. *International journal of social robotics*, 12(2):459–478, 2020.
- [90] Francien Dechesne, Gennaro Di Tosto, Virginia Dignum, and Frank Dignum. No smoking here: Values, norms and culture in multi-agent systems. *Artificial intelligence and law*, 21(1):79–107, 2013.
- [91] Louis Deslauriers, Logan S McCarty, Kelly Miller, Kristina Callaghan, and Greg Kestin. Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *Proceedings of the National Academy of Sciences*, 116(39):19251–19257, 2019.
- [92] Morton Deutsch. Trust and suspicion. *Journal of conflict resolution*, 2(4):265–279, 1958.
- [93] S Kate Devitt. Trustworthiness of autonomous systems. In *Foundations of trusted autonomy*, pages 161–184. Springer, Cham, 2018.
- [94] Thomas G Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning*, 40(2):139–157, 2000.
- [95] Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan. Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th international conference on intelligent user interfaces*, pages 275–285, 2019.
- [96] Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan. Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th international conference on intelligent user interfaces*, pages 275–285, 2019.

- [97] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.
- [98] Pierre P Duez, Michael J Zuliani, and Greg A Jamieson. Trust by design: information requirements for appropriate trust in automation. In *Proceedings of the 2006 conference of the Center for Advanced Studies on Collaborative research*, pages 9–es, 2006.
- [99] Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, and Linda G. Pierce. The role of trust in automation reliance. Jun 2003.
- [100] Upol Ehsan, Philipp Wintersberger, Q Vera Liao, Martina Mara, Marc Streit, Sandra Wachter, Andreas Riener, and Mark O Riedl. Operationalizing human-centered perspectives in explainable ai. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2021.
- [101] Shmuel Noah Eisenstadt, Shemuel Noah Aizenshtadt, Luis Roniger, et al. *Patrons, clients and friends: Interpersonal relations and the structure of trust in society*. Cambridge University Press, 1984.
- [102] Fredrick Ekman, Mikael Johansson, and Jana Sochor. Creating appropriate trust in automated vehicle systems: A framework for hmi design. *IEEE Transactions on Human-Machine Systems*, 48(1):95–101, 2017.
- [103] MICA R. ENDSLEY and DAVID B. KABER. Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics*, 42(3):462–492, 1999. doi: 10.1080/001401399185595. Pmid: 10048306.
- [104] Thorsten M Erle and Michael K Zürn. Illusory trust: Kanizsa shapes incidentally increase trust and willingness to invest. *Journal of Behavioral Decision Making*, 33(5):671–682, 2020.
- [105] Anthony M Evans, Ursula Athenstaedt, and Joachim I Krueger. The development of trust and altruism during childhood. *Journal of economic psychology*, 36:82–95, 2013.
- [106] Rino Falcone and Cristiano Castelfranchi. Social trust: A cognitive approach. *Trust and deception in virtual societies*, pages 55–90, 2001.
- [107] Rino Falcone, Michele Piunti, Matteo Venanzi, and Cristiano Castelfranchi. From manifesta to krypta: The relevance of categories for trusting others. *ACM Transactions on Intelligent Systems and Technology*, 4, 3 2013. ISSN 21576904. doi: 10.1145/2438653.2438662.
- [108] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. G\* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, 39(2):175–191, 2007.
- [109] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. Statistical power analyses using g\* power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, 41(4):1149–1160, 2009.

- [110] Edward A. Feigenbaum. Computer professionals against abm: Organization of computer experts calls abm project a dangerous mistake, 1971. URL <https://exhibits-lb.stanford.edu/cs/catalog/pc764bb9418>.
- [111] Heike Felzmann, Eduard Fosch-Villaronga, Christoph Lutz, and Aurelia Tamò-Larrieux. Towards transparency by design for artificial intelligence. *Science and Engineering Ethics*, 26(6):3333–3361, 2020.
- [112] Andrew Guthrie Ferguson. The legal risks of big data policing. *Crim. Just.*, 33:4, 2018.
- [113] Andrea Ferrario and Michele Loi. How explainability contributes to trust in ai. In *2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22*, page 1457–1466, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533202. URL <https://doi.org/10.1145/3531146.3533202>.
- [114] Carolina Ferreira Gomes Centeio Jorge, Siddharth Mehrotra, Myrthe L. Tielman, and Catholijn M. Jonker. Trust should correspond to trustworthiness: a formalization of appropriate mutual trust in human-agent teams. *Proceedings of the 22nd International Workshop on Trust in Agent Societies, London, UK*, 2021.
- [115] W Holmes Finch, Jocelyn E Bolin, and Ken Kelley. *Multilevel modeling using R*. CRC Press, 2019.
- [116] Gavin D Flood. *An introduction to Hinduism*. Cambridge University Press, 1996.
- [117] Michael W Floyd, Michael Drinkwater, and David W Aha. How much do you trust me? learning a case-based model of inverse trust. In *International Conference on Case-Based Reasoning*, pages 125–139. Springer, 2014.
- [118] M Lance Frazier, Paul D Johnson, and Stav Fainshmidt. Development and validation of a propensity to trust scale. *Journal of Trust Research*, 3(2):76–97, 2013.
- [119] Amos Freedy, Ewart DeVisser, Gershon Weltman, and Nicole Coeyman. Measurement of trust in human-robot collaboration. In *2007 international symposium on collaborative technologies and systems*, pages 106–114. IEEE, 2007.
- [120] Markus Freitag and Paul C Bauer. Personality traits and the propensity to trust friends and strangers. *The Social Science Journal*, 53(4):467–476, 2016.
- [121] Batya Friedman, Peter H Kahn, and Alan Borning. Value sensitive design and information systems. *The handbook of information and computer ethics*, pages 69–101, 2008.
- [122] Catherine O Fritz, Peter E Morris, and Jennifer J Richler. Effect size estimates: current use, calculations, and interpretation. *Journal of experimental psychology: General*, 141(1):2, 2012.

- [123] Michael Georgeff, Barney Pell, Martha Pollack, Milind Tambe, and Michael Wooldridge. The belief-desire-intention model of agency. In *Intelligent Agents V: Agents Theories, Architectures, and Languages: 5th International Workshop, ATAL'98 Paris, France, July 4–7, 1998 Proceedings 5*, pages 1–10. Springer, 1999.
- [124] Monica M Gerber and Jonathan Jackson. Justifying violence: legitimacy, ideology and public support for police use of force. *Psychology, crime & law*, 23(1):79–95, 2017.
- [125] Dominik Gerstner. Predictive policing in the context of residential burglary: An empirical illustration on the basis of a pilot project in baden-württemberg, germany. *European Journal for Security Research*, 3(2):115–138, 2018.
- [126] Bhavya Ghai, Q Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. Explainable active learning (xal) toward ai explanations as interfaces for machine teachers. *Proceedings of the ACM on Human-Computer Interaction*, 4(Cscw3):1–28, 2021.
- [127] Anthony Giddens. *The consequences of modernity*. John Wiley & Sons, 2013.
- [128] Felix Gille, Anna Jobin, and Marcello Ienca. What we talk about when we talk about trust: Theory of trust for ai in healthcare. *Intelligence-Based Medicine*, 1:100001, 2020.
- [129] Ella Glikson and Anita Williams Woolley. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, (14, 2), 2020.
- [130] Juan Angel Gonzalez-Aguirre, Ricardo Osorio-Oliveros, Karen L Rodríguez-Hernández, Javier Lizárraga-Iturralde, Ruben Morales Menendez, Ricardo A Ramírez-Mendoza, Mauricio Adolfo Ramírez-Moreno, and Jorge de Jesus Lozoya-Santos. Service robots: Trends and technology. *Applied Sciences*, 11(22):10702, 2021.
- [131] Jane Goudge and Lucy Gilson. How can trust be investigated? drawing lessons from past experience. *Social science & medicine*, 61(7):1439–1451, 2005.
- [132] Gregory M Gremillion, Jason S Metcalfe, Amar R Marathe, Victor J Paul, James Christensen, Kim Drnec, Benjamin Haynes, and Corey Atwater. Analysis of trust in autonomy for convoy operations. In *Micro-and nanotechnology sensors, systems, and applications viii*, volume 9836, pages 356–365. Spie, 2016.
- [133] N. Griffiths. Task delegation using experience-based multi-dimensional trust. In *AAMAS '05*, 2005.
- [134] Siddharth Gulati, Sonia Sousa, and David Lamas. Design, development and evaluation of a human-computer trust scale. *Behaviour & Information Technology*, 38(10):1004–1015, 2019.
- [135] David Gunning, Eric Vorm, Jennifer Yunyan Wang, and Matt Turek. Darpa's explainable ai (xai) program: A retrospective. *Applied AI Letters*, 2(4):e61, 2021. doi: 10.1002/ail2.61.

- [136] Yaohui Guo and X Jessie Yang. Modeling and predicting trust dynamics in human-robot teaming: A bayesian inference approach. *International Journal of Social Robotics*, 13(8):1899–1909, 2021.
- [137] Banavar Guru. What It Will Take for Us to Trust AI. *Harvard Business Review*, November 2016. ISSN 0017-8012. URL <https://hbr.org/2016/11/what-it-will-take-for-us-to-trust-ai>.
- [138] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. A meta-analysis of factors affecting trust in human-robot interaction. *Human factors*, 53(5):517–527, 2011.
- [139] Wim Hardyns and Anneleen Rummens. Predictive policing as a new tool for law enforcement? recent developments and challenges. *European journal on criminal policy and research*, 24:201–218, 2018.
- [140] Katherine Hawley. Trust, distrust and commitment. *Noûs*, 48(1):1–20, 2014.
- [141] Katherine Hawley. Trustworthy groups and organizations. *The philosophy of trust*, pages 230–250, 2017.
- [142] Gaole He, Agathe Balayn, Stefan Buijsman, Jie Yang, and Ujwal Gadiraju. It is like finding a polar bear in the savannah! concept-level ai explanations with analogical inference from commonsense knowledge. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, pages 89–101, 2022.
- [143] David R Heise. *Understanding events: Affect and the construction of social action*. Cambridge University Press New York, 1979.
- [144] Tove Helldin, Göran Falkman, Maria Riveiro, and Staffan Davidsson. Presenting system uncertainty in automotive uis for supporting trust calibration in autonomous driving. In *Proceedings of the 5th international conference on automotive user interfaces and interactive vehicular applications*, pages 210–217, 2013.
- [145] Sarita Herse, Jonathan Vitale, Benjamin Johnston, and Mary-Anne Williams. Using trust to determine user decision making & task outcome during a human-agent collaborative task. In *Proceedings of the 2021 ACM/IEEE international conference on human-robot interaction*, pages 73–82, 2021.
- [146] Andreas Herzig, Emiliano Lorini, Jomi F. Hübner, and Laurent Vercouter. A logic of trust and reputation. *Logic Journal of the IGPL*, 18:214–244, 12 2009. ISSN 13670751. doi: 10.1093/jigpal/jzp077.
- [147] Zoë Hobson, Julia A Yesberg, Ben Bradford, and Jonathan Jackson. Artificial fairness? trust in algorithmic police decision-making. *Journal of experimental criminology*, pages 1–25, 2021.
- [148] Kevin Anthony Hoff and Masooda Bashir. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, 57(3):407–434, 2015.

- [149] Robert R Hoffman. A taxonomy of emergent trusting in the human-machine relationship. *Cognitive Systems Engineering*, pages 137–164, 2017.
- [150] Heike Hofmann. Exploring categorical data: interactive mosaic plots. *Metrika*, 51(1): 11–26, 2000.
- [151] Fred Hohman, Arjun Srinivasan, and Steven M Drucker. Telegam: Combining visualization and verbalization for interpretable machine learning. In *2019 IEEE Visualization Conference (VIS)*, pages 151–155. IEEE, 2019.
- [152] Daniel Holliday, Stephanie Wilson, and Simone Stumpf. User trust in intelligent systems: A journey over time. In *Proceedings of the 21st international conference on intelligent user interfaces*, pages 164–168, 2016.
- [153] Linda Childers Hon and James E Grunig. Guidelines for measuring relationships in public relations. 1999.
- [154] Chong Sun Hong and Tae Gyu Oh. Correlation plot for a contingency table. *Communications for Statistical Applications and Methods*, 28(3):295–305, 2021.
- [155] Ming Hou, Simon Banbury, Brad Cain, Scott Fang, Hannah Willoughby, Liam Foley, Edward Tunstel, and Imre J Rudas. Impacts homeostasis trust management system: Optimizing trust in human-ai teams. *ACM Computing Surveys*.
- [156] Marc W Howard and Michael J Kahana. Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(4):923, 1999.
- [157] Sandy H Huang, Kush Bhatia, Pieter Abbeel, and Anca D Dragan. Establishing appropriate trust via critical states. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3929–3936. IEEE, 2018.
- [158] YL Huang and LT Fan. A fuzzy-logic-based approach to building efficient fuzzy rule-based expert systems. *Computers & chemical engineering*, 17(2):181–192, 1993.
- [159] Leo WJC Huberts. Integrity: What it is and why it is important. *Public Integrity*, 20 (sup1):S18–S32, 2018.
- [160] Judith E Hupcey, Janice Penrod, Janice M Morse, and Carl Mitcham. An exploration and advancement of the concept of trust. *Journal of advanced nursing*, 36(2):282–293, 2001.
- [161] Ronald Hurst and Leslie R Hurst. *Pilot error: The human factors*. Jason Aronson, 1982.
- [162] Aya Hussein, Sondoss Elsawah, and Hussein A Abbass. Trust mediating reliability-reliance relationship in supervisory control of human-swarm interactions. *Human Factors*, 62(8):1237–1248, 2020.

- [163] IEEE. Ethically aligned design - a vision for prioritizing human well-being with autonomous and intelligent systems, Dec 2017. URL [https://standards.ieee.org/wp-content/uploads/import/documents/other/ead\\_v2.pdf](https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf).
- [164] Brett Israelsen, Peggy Wu, Katharine Woodruff, Gianna Avdic-McIntire, Andrew Radlbeck, Angus McLean, Patrick" Dice" Highland, Thomas" Mach" Schnell, and Daniel" Animal" Javorsek. Introducing smrtt: A structural equation model of multi-modal real-time trust. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pages 126–130, 2021.
- [165] Brett W Israelsen and Nisar R Ahmed. “dave... i can assure you... that it’s going to be all right...” a definition, case for, and survey of algorithmic assurances in human-autonomy trust relationships. *ACM Computing Surveys (CSUR)*, 51(6):1–37, 2019.
- [166] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 624–635, 2021.
- [167] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 624–635, 2021.
- [168] Christian Janiesch, Patrick Zschech, and Kai Heinrich. Machine learning and deep learning. *Electronic Markets*, 31(3):685–695, 2021.
- [169] T. H. Jeavons. *Ethics in nonprofit management*, page 108–119. Routledge, 2001.
- [170] Theodore Jensen, Yusuf Albayram, Mohammad Maifi Hasan Khan, Ross Buck, Emil Coman, and Md Abdullah Al Fahim. Initial trustworthiness perceptions of a drone system based on performance and process information. pages 229–237, 2018.
- [171] Theodore Jensen, Yusuf Albayram, Mohammad M.H. Khan, Ross Buck, Emil Coman, and Md A. Al Fahim. Initial trustworthiness perceptions of a drone system based on performance and process information. In *Proceedings of 6th International Conference on Human-Agent Interaction*, 2018.
- [172] Theodore Jensen, Mohammad Maifi Hasan Khan, and Yusuf Albayram. The role of behavioral anthropomorphism in human-automation trust calibration. In *Artificial Intelligence in HCI: First International Conference, AI-HCI 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings*, pages 33–53. Springer, 2020.
- [173] Theodore Jensen, Mohammad Maifi Hasan Khan, Md Abdullah Al Fahim, and Yusuf Albayram. Trust and anthropomorphism in tandem: the interrelated nature of automated agent appearance and reliability in trustworthiness perceptions. In *Designing interactive systems conference 2021*, pages 1470–1480, 2021.

- [174] Wolfgang Jentner, Rita Sevastjanova, Florian Stoffel, Daniel A Keim, Jürgen Bernard, and Mennatallah El-Assady. Minions, sheep, and fruits: metaphorical narratives to explain artificial intelligence and build trust. In *Workshop on Visualization for AI Explainability at IEEE*, 2018.
- [175] Barbara Jodlbauer and Eva Jonas. Forecasting clients’ reactions: How does the perception of strategic behavior influence the acceptance of advice? *International Journal of Forecasting*, 27(1):121–133, 2011.
- [176] Craig J Johnson, Mustafa Demir, Nathan J McNeese, Jamie C Gorman, Alexandra T Wolff, and Nancy J Cooke. The impact of training on human–autonomy team communications and trust calibration. *Human factors*, page 00187208211047323, 2021.
- [177] Devon Johnson and Kent Grayson. Cognitive and affective trust in service relationships. *Journal of Business research*, 58(4):500–507, 2005.
- [178] Kevin B Johnson, Wei-Qi Wei, Dilhan Weeraratne, Mark E Frisse, Karl Misulis, Kyu Rhee, Juan Zhao, and Jane L Snowdon. Precision medicine, ai, and the future of personalized health care. *Clinical and translational science*, 14(1):86–93, 2021.
- [179] Christopher Jones. Law enforcement use of facial recognition: bias, disparate impacts on people of color, and the need for federal legislation. *NCJL & Tech.*, 22:777, 2020.
- [180] Karen Jones. The politics of credibility. In *A Mind of One’s Own*, pages 154–176. Routledge, 2018.
- [181] Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.
- [182] C Centeio Jorge, Siddharth Mehrotra, ML Tielman, and CM Jonker. Trust should correspond to trustworthiness: a formalization of appropriate mutual trust in human-agent teams. In *22nd International Trust Workshop 2021*, 2021.
- [183] Wiard Jorritsma, Fokie Cnossen, and Peter MA van Ooijen. Improving the radiologist–cad interaction: designing for appropriate trust. *Clinical radiology*, 70(2):115–122, 2015.
- [184] Audun Jøsang and Stéphane Lo Presti. Analysing the relationship between risk and trust. In *International conference on trust management*, pages 135–145. Springer, 2004.
- [185] Poornima Kaniarasu, Aaron Steinfeld, Munjal Desai, and Holly Yanco. Robot confidence and trust alignment. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 155–156. Ieee, 2013.
- [186] Shivani Kapania, Oliver Siy, Gabe Clapper, Azhagu Meena SP, and Nithya Sambasivan. ”because ai is 100% right and safe”: User attitudes and sources of ai authority in india. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI ’22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3517533. URL <https://doi-org.tudelft.idm.oclc.org/10.1145/3491102.3517533>.

- [187] Lord Kelvin. If you cannot measure it, you cannot improve it. *Accessed May, 19, 2010.*
- [188] Arnon Keren. Trust and belief: a preemptive reasons account. *Synthese*, 191(12): 2593–2615, 2014.
- [189] Mohammad T Khasawneh, Shannon R Bowling, Xiaochun Jiang, Anand K Gramopadhye, and Brian J Melloy. A model for predicting human trust in automated systems. *Origins*, 5, 2003.
- [190] Siddartha Khastgir, Stewart Birrell, Gunwant Dhadyalla, and Paul Jennings. Calibrating trust through knowledge: Introducing the concept of informed safety for automation in vehicles. *Transportation research part C: emerging technologies*, 96: 290–303, 2018.
- [191] Siddartha Khastgir, Stewart Birrell, Gunwant Dhadyalla, and Paul Jennings. Calibrating trust through knowledge: Introducing the concept of informed safety for automation in vehicles. *Transportation Research Part C: Emerging Technologies*, 96:290–303, 2018. ISSN 0968-090x. doi: <https://doi.org/10.1016/j.trc.2018.07.001>. URL <https://www.sciencedirect.com/science/article/pii/S0968090X18309252>.
- [192] Sara Kiesler and Jennifer Goetz. Mental models of robotic assistants. In *CHI'02 extended abstracts on Human Factors in Computing Systems*, pages 576–577, 2002.
- [193] Been Kim, Caleb M Chacha, and Julie A Shah. Inferring team task plans from human meetings: A generative modeling approach with logic-based prior. *Journal of Artificial Intelligence Research*, 52:361–398, 2015.
- [194] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. In *Proceedings of the European conference on computer vision (ECCV)*, pages 563–578, 2018.
- [195] Wonjoon Kim, Nayoung Kim, Joseph B Lyons, and Chang S Nam. Factors affecting trust in high-vulnerability human-robot interaction contexts: A structural equation modelling approach. *Applied ergonomics*, 85:103056, 2020.
- [196] HE Knee and JC Schryver. Operator role definition and human system integration. Technical report, Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States), 1989.
- [197] Bran Knowles and John T Richards. The sanction of authority: Promoting public trust in ai. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 262–271, 2021.
- [198] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019.

- [199] Barbara Kordy, Ludovic Piètre-Cambacédès, and Patrick Schweitzer. Dag-based attack and defense modeling: Don't miss the forest for the attack trees. *Computer science review*, 13:1–38, 2014.
- [200] Roderick M Kramer and Tom R Tyler. *Trust in organizations: Frontiers of theory and research*. Sage Publications, 1995.
- [201] Johannes Kraus, David Scholz, Dina Stiegemeier, and Martin Baumann. The more you know: trust dynamics and calibration in highly automated driving and the effects of take-overs, system malfunction, and system transparency. *Human factors*, 62(5):718–736, 2020.
- [202] Johannes Maria Kraus, Yannick Forster, Sebastian Hergeth, and Martin Baumann. Two routes to trust calibration: effects of reliability and brand information on trust in automation. *International Journal of Mobile Human Computer Interaction (IJMHCI)*, 11(3):1–17, 2019.
- [203] Alexandra Kuznetsova, Per B Brockhoff, and Rune HB Christensen. Imertest package: tests in linear mixed effects models. *Journal of statistical software*, 82:1–26, 2017.
- [204] Justine Lacey, Mark Howden, Christopher Cvitanovic, and R. M. Colvin. Understanding and managing trust at the climate science–policy interface. *Nature Climate Change*, 8(1):22–28, Jan 2018. ISSN 1758-6798. doi: 10.1038/s41558-017-0010-z. URL <https://doi.org/10.1038/s41558-017-0010-z>.
- [205] Vivian Lai and Chenhao Tan. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 29–38, 2019.
- [206] Vivian Lai and Chenhao Tan. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT\* '19*, page 29–38, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287590. URL <https://doi.org/10.1145/3287560.3287590>.
- [207] Vivian Lai, Han Liu, and Chenhao Tan. " why is' chicago' deceptive?" towards building model-driven tutorials for humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.
- [208] Himabindu Lakkaraju and Osbert Bastani. " how do i fool you?" manipulating user trust via misleading black box explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 79–85, 2020.
- [209] Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16*, page 1675–1684, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939874. URL <https://doi.org/10.1145/2939672.2939874>.

- [210] Kathryn Ann Lambe, Gary O'Reilly, Brendan D Kelly, and Sarah Curristan. Dual-process cognitive interventions to enhance diagnostic reasoning: a systematic review. *BMJ quality & safety*, 25(10):808–820, 2016.
- [211] Retno Larasati, Anna De Liddo, and Enrico Motta. Meaningful explanation effect on user's trust in an ai medical system: Designing explanations for non-expert users. *ACM Trans. Interact. Intell. Syst.*, 13(4), dec 2023. ISSN 2160-6455. doi: 10.1145/3631614. URL <https://doi.org/10.1145/3631614>.
- [212] Christian Lebiere, Leslie M Blaha, Corey K Fallon, and Brett Jefferson. Adaptive cognitive mechanisms to maintain calibrated trust and reliance in automation. *Frontiers in Robotics and AI*, 8:652776, 2021.
- [213] Dong-Jin Lee, Moonkyu Lee, and Jaebeom Suh. Benevolence in the importer-exporter relationship: Moderating role of value similarity and cultural familiarity. *International Marketing Review*, 2007.
- [214] J. D. Lee and Katrina A. See. Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of Human Factors and Ergonomics Society*, 46:50 – 80, 2004.
- [215] John D Lee and Neville Moray. Trust, self-confidence, and operators' adaptation to automation. *International journal of human-computer studies*, 40(1):153–184, 1994.
- [216] John D Lee and Katrina A See. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1):50–80, 2004.
- [217] John D Lee and Katrina A See. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1):50–80, 2004.
- [218] Min Hun Lee and Chong Jun Chew. Understanding the effect of counterfactual explanations on trust and reliance on ai for human-ai collaborative clinical decision making. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2), oct 2023. doi: 10.1145/3610218. URL <https://doi.org/10.1145/3610218>.
- [219] Scott LeeTiernan, Edward Cutrell, Mary Czerwinski, and Hunter G Hoffman. Effective notification systems depend on user trust. In *INTERACT*, pages 684–685, 2001.
- [220] Jie Leng and Jixia Wu. Integrity perceptions and behavior triggered by the hand-over-chest gesture: A semiotic perspective. *Language*, 3, 2019.
- [221] Roy J Lewicki and Chad Brinsfield. Trust research: measuring trust beliefs and behaviours. In *Handbook of research methods on trust*. Edward Elgar Publishing, 2015.
- [222] Roy J Lewicki, Barbara B Bunker, et al. Developing and maintaining trust in work relationships. *Trust in organizations: Frontiers of theory and research*, 114:139, 1996.

- [223] Q.Vera Liao and S. Shyam Sundar. Designing for responsible trust in ai systems: A communication perspective. In *2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22*, page 1257–1268, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533182. URL <https://doi.org/10.1145/3531146.3533182>.
- [224] Enrico Liscio, Michiel van der Meer, Luciano C Siebert, Catholijn M Jonker, Niek Mouter, and Pradeep K Murukannaiah. Axies: Identifying and evaluating context-specific values. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, pages 799–808, 2021.
- [225] Han Liu, Vivian Lai, and Chenhao Tan. Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making. *Proc. ACM Hum.-Comput. Interact.*, 5(Cscw2), October 2021. doi: 10.1145/3479552. URL <https://doi.org/10.1145/3479552>.
- [226] Han Liu, Vivian Lai, and Chenhao Tan. Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–45, 2021.
- [227] Yidu Lu and Nadine Sarter. Feedback on system or operator performance: Which is more useful for the timely detection of changes in reliability, trust calibration and appropriate automation usage? In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 63, pages 312–316. SAGE Publications Sage CA: Los Angeles, CA, 2019.
- [228] Henrietta Lyons, Senuri Wijenayake, Tim Miller, and Eduardo Velloso. What’s the appeal? perceptions of review processes for algorithmic decisions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, CHI '22*, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3517606. URL <https://doi-org.tudelft.idm.oclc.org/10.1145/3491102.3517606>.
- [229] Stefanie M. Faas, Johannes Kraus, Alexander Schoenhals, and Martin Baumann. Calibrating pedestrians’ trust in automated vehicles: does an intent display in an external hmi support trust calibration and safe crossing behavior? In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2021.
- [230] Shuai Ma, Ying Lei, Xinru Wang, Chengbo Zheng, Chuhan Shi, Ming Yin, and Xiaojuan Ma. Who should i trust: Ai or myself? leveraging human and ai correctness likelihood to promote appropriate trust in ai-assisted decision-making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23*, page 1–19, New York, NY, USA, Apr 2023. Association for Computing Machinery. ISBN 978-1-4503-9421-5. doi: 10.1145/3544548.3581058. URL <https://dl.acm.org/doi/10.1145/3544548.3581058>.
- [231] Erina L MacGeorge and Lyn M Van Swol. *The Oxford handbook of advice*. Oxford University Press, 2018.

- [232] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.
- [233] Poornima Madhavan and Douglas A Wiegmann. Similarities and differences between human–human and human–automation trust: an integrative review. *Theoretical Issues in Ergonomics Science*, 8(4):277–301, 2007.
- [234] Alexander Maedche, Christine Legner, Alexander Benlian, Benedikt Berger, Henner Gimpel, Thomas Hess, Oliver Hinz, Stefan Morana, and Matthias Söllner. Ai-based digital assistants: Opportunities, threats, and research perspectives. *Business & Information Systems Engineering*, 61:535–544, 2019.
- [235] Kit-Kay Mak, Yi-Hang Wong, and Mallikarjuna Rao Pichika. Artificial intelligence in drug discovery and development. *Drug Discovery and Evaluation: Safety and Pharmacokinetic Assays*, pages 1–38, 2023.
- [236] Gary Marcus and Ernest Davis. *Rebooting AI: Building artificial intelligence we can trust*. Vintage, 2019.
- [237] Vidushi Marda and Shivangi Narayan. Data in new delhi’s predictive policing system. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 317–324, 2020.
- [238] Stephen Marsh and Mark R Dibben. Trust, untrust, distrust and mistrust—an exploration of the dark (er) side. In *International conference on trust management*, pages 17–33. Springer, 2005.
- [239] Ronald Scott Marshall. Building trust early: the influence of first and second order expectations on trust in international channels of distribution. *International Business Review*, 12(4):421–443, 2003.
- [240] Brendan Max. Soundthinking’s black-box gunshot detection method: Untested and unvetted tech flourishes in the criminal justice system. *Stan. Tech. L. Rev.*, 26:193, 2022.
- [241] Roger C Mayer, James H Davis, and F David Schoorman. An integrative model of organizational trust. *Source: The Academy of Management Review*, 20:709–734, 1995.
- [242] Roger C Mayer, James H Davis, and F David Schoorman. An integrative model of organizational trust. *Academy of management review*, 20(3):709–734, 1995.
- [243] Maranda McBride and Shona Morgan. Trust calibration for automated decision aids. *Institute for Homeland Security Solutions*, pages 1–11, 2010.
- [244] John McDaniel and Ken Pease. *Predictive policing and artificial intelligence*. Routledge, 2021.

- [245] Patricia L McDermott and Ronna N ten Brink. Practical guidance for evaluating calibrated trust. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 63, pages 362–366. SAGE Publications Sage CA: Los Angeles, CA, 2019.
- [246] Lynne McFall. Integrity. *Ethics*, 98:5–20, 1987.
- [247] John M McGuirl and Nadine B Sarter. Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human factors*, 48(4):656–665, 2006.
- [248] D Harrison McKnight, Vivek Choudhury, and Charles Kacmar. The impact of initial consumer trust on intentions to transact with a web site: a trust building model. *The journal of strategic information systems*, 11(3-4):297–323, 2002.
- [249] David L McLain and Katarina Hackman. Trust, risk, and decision-making in organizational change. *Public Administration Quarterly*, pages 152–176, 1999.
- [250] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.
- [251] Siddharth Mehrotra. Modelling trust in human-ai interaction. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1826–1828, 2021.
- [252] Siddharth Mehrotra, Carolina Centeio Jorge, Catholijn M Jonker, and Myrthe L Tielman. Integrity based explanations for fostering appropriate trust in ai agents. *ACM Transactions on Interactive Intelligent Systems*.
- [253] Siddharth Mehrotra, Catholijn M. Jonker, and Myrthe L. Tielman. More similar values, more trust? - the effect of value similarity on trust in human-agent interaction. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, page 777–783, New York, NY, USA, 2021. ACM. ISBN 9781450384735.
- [254] Siddharth Mehrotra, Catholijn M Jonker, and Myrthe L Tielman. More similar values, more trust?-the effect of value similarity on trust in human-agent interaction. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 777–783, 2021.
- [255] Siddharth Mehrotra, Chadha Degachi, Oleksandra Vereschak, Catholijn M Jonker, and Myrthe L Tielman. A systematic review on fostering appropriate trust in human-ai interaction. *arXiv preprint arXiv:2311.06305*, 2023.
- [256] Siddharth Mehrotra, Carolina Centeio Jorge, Catholijn M. Jonker, and Myrthe L. Tielman. Integrity based explanations for fostering appropriate trust in ai agents. *ACM Trans. Interact. Intell. Syst.*, jul 2023. ISSN 2160-6455. doi: 10.1145/3610578. URL <https://doi.org/10.1145/3610578>. Just Accepted.

- [257] Albert Meijer and Martijn Wessels. Predictive policing: Review of benefits and drawbacks. *International Journal of Public Administration*, 42(12):1031–1039, 2019.
- [258] Albert Meijer, Lukas Lorenz, and Martijn Wessels. Algorithmization of bureaucratic organizations: Using a practice lens to study how context shapes predictive policing systems. *Public Administration Review*, 81(5):837–846, 2021.
- [259] Rijk Mercur, Virginia Dignum, and Catholijn Jonker. The value of values and norms in social simulation. *Journal of Artificial Societies and Social Simulation*, 22(1), 2019.
- [260] Stephanie M Merritt. Affective processes in human–automation interactions. *Human Factors*, 53(4):356–370, 2011.
- [261] Tim Miller. Are we measuring trust correctly in explainability, interpretability, and transparency research? *arXiv preprint arXiv:2209.00651*, 2022.
- [262] Tim Miller. Are we measuring trust correctly in explainability, interpretability, and transparency research? *TRAIT 2022: Trust and Reliance in AI-Human Teams*, page 11, 2022.
- [263] Tim Miller. Explainable ai is dead, long live explainable ai! hypothesis-driven decision support using evaluative ai. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 333–342, 2023.
- [264] Alexander G Mirnig, Philipp Wintersberger, Christine Sutter, and Jürgen Ziegler. A framework for analyzing and calibrating trust in automated vehicles. In *Adjunct proceedings of the 8th international conference on automotive user interfaces and interactive vehicular applications*, pages 33–38, 2016.
- [265] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. A survey of evaluation methods and measures for interpretable machine learning. *ArXiv*, abs/1811.11839, 2018. URL <https://api.semanticscholar.org/CorpusID:54087635>.
- [266] Xiaomin Mou. Artificial intelligence: investment trends and selected industry uses. *International Finance Corporation*, 8, 2019.
- [267] Bonnie M. Muir. Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, 27(5):527–539, 1987. ISSN 0020-7373. doi: [https://doi.org/10.1016/S0020-7373\(87\)80013-5](https://doi.org/10.1016/S0020-7373(87)80013-5). URL <https://www.sciencedirect.com/science/article/pii/S0020737387800135>.
- [268] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.
- [269] Mohammad Naiseh, Reem S Al-Mansoori, Dena Al-Thani, Nan Jiang, and Raian Ali. Nudging through friction: An approach for calibrating trust in explainable ai. In *2021 8th International Conference on Behavioral and Social Computing (BESC)*, pages 1–5. Ieee, 2021.

- [270] Mohammad Naiseh, Dena Al-Thani, Nan Jiang, and Raian Ali. Explainable recommendation: When design meets trust calibration. *World Wide Web*, 24(5): 1857–1884, September 2021. ISSN 1386-145x. doi: 10.1007/s11280-021-00916-0. URL <https://doi.org/10.1007/s11280-021-00916-0>.
- [271] Alex Najibi. Racial discrimination in face recognition technology. *Science in the News*, 24, 2020.
- [272] Birthe Nettet, David A Robb, José Lopes, and Helen Hastie. Transparency in hri: Trust and decision making in the face of robot errors. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pages 313–317, 2021.
- [273] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015.
- [274] David J Niedober, Nhut T Ho, Gina Masesquesmay, Kolina Koltai, Mark Skoog, Artemio Cacanindin, Walter Johnson, and Joseph B Lyons. Influence of cultural, organizational and automation factors on human-automation trust: A case study of auto-gcas engineers and developmental history. In *Human-Computer Interaction. Applications and Services: 16th International Conference, HCI International 2014, Heraklion, Crete, Greece, June 22-27, 2014, Proceedings, Part III* 16, pages 473–484. Springer, 2014.
- [275] Nils J Nilsson. *Understanding beliefs*. MIT Press, 2014.
- [276] Bart Nooteboom. Trust and innovation. *Handbook of advances in trust research*, 106, 2013.
- [277] Ingrid Nunes and Dietmar Jannach. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction*, 27:393–444, 2017.
- [278] Rui Ogawa, Sung Park, and Hiroyuki Umemuro. How humans develop trust in communication robots: A phased model based on interpersonal trust. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 606–607. IEEE, 2019.
- [279] Kazuo Okamura and Seiji Yamada. Adaptive trust calibration for human-ai collaboration. *PLoS ONE*, 15(2), 2020.
- [280] Onora O’neill. *Autonomy and trust in bioethics*. Cambridge University Press, 2002.
- [281] Serena Oosterloo, Gerwin van Schie, Jo Bates, Paul Clough, Robert Jäschke, Jahna Otterbacher, et al. The politics and biases of the “crime anticipation system” of the dutch police. In *Proceedings of the international workshop on bias in information, algorithms, and systems*, volume 2103, pages 30–41. CEUR WS, 2018.
- [282] Scott Ososky, David Schuster, Elizabeth Phillips, and Florian G Jentsch. Building appropriate trust in human-robot teams. In *2013 AAAI spring symposium series*, 2013.

- [283] Elinor Ostrom. A behavioral approach to the rational choice theory of collective action: Presidential address, american political science association, 1997. *American political science review*, 92(1):1–22, 1998.
- [284] Margit E Oswald and Corina T Ulshöfer. Cooperation and distrust—a contradiction? *Social Dilemmas, Institutions, and the Evolution of Cooperation*, page 357, 2017.
- [285] Hafsa Ouchra, Abdessamad Belangour, and Allae Erraissi. An overview of geospatial artificial intelligence technologies for city planning and development. In *2023 Fifth International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, pages 1–7. IEEE, 2023.
- [286] Mourad Ouzzani, Hossam Hammady, Zbys Fedorowicz, and Ahmed Elmagarmid. Rayyan—a web and mobile app for systematic reviews. *Systematic reviews*, 5:1–10, 2016.
- [287] Andrés Páez. The pragmatic turn in explainable artificial intelligence (xai). *Minds and Machines*, 29(3):441–459, 2019.
- [288] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *International journal of surgery*, 88:105906, 2021.
- [289] Marine Pagliari, Valérien Chambon, and Bruno Berberian. What is new with artificial intelligence? human–agent interactions through the lens of social agency. *Frontiers in Psychology*, 13:954444, 2022.
- [290] Google PAIR. People + ai guidebook. [pair.withgoogle.com/guidebook](https://pair.withgoogle.com/guidebook), 2019. [Accessed 14-Feb-2023].
- [291] Michael E Palanski and Francis J Yammarino. Integrity and leadership:: clearing the conceptual confusion. *European Management Journal*, 25(3):171–184, 2007.
- [292] Raja Parasuraman and Evan A Byrne. Automation and human performance in aviation. *Principles and practice of aviation psychology*, pages 311–356, 2003.
- [293] Raja Parasuraman and Christopher A Miller. Trust and etiquette in high-criticality automated systems. *Communications of the ACM*, 47(4):51–55, 2004.
- [294] Raja Parasuraman and Victor Riley. Humans and automation: Use, misuse, disuse, abuse. *Human factors*, 39(2):230–253, 1997.
- [295] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8779–8788, 2018.
- [296] Young Moon Park, Gwang-Won Kim, and Jin-Man Sohn. A logic based expert system (lbes) for fault diagnosis of power system. *IEEE Transactions on power systems*, 12(1): 363–369, 1997.

- [297] Andisheh Partovi, Ingrid Zukerman, Kai Zhan, Nora Hamacher, and Jakob Hohwy. Relationship between device performance, trust and user behaviour in a care-taking scenario. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, pages 61–69, 2019.
- [298] Edsel A Peña and Elizabeth H Slate. Global validation of linear model assumptions. *Journal of the American Statistical Association*, 101(473):341–354, 2006.
- [299] Christopher J Peters. Foolish consistency: On equality, integrity, and justice in stare decisis. *Yale Lj*, 105:2031, 1995.
- [300] Ivens Portugal, Paulo Alencar, and Donald Cowan. The use of machine learning algorithms in recommender systems: A systematic review. *Expert Systems with Applications*, 97:205–227, 2018.
- [301] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–52, 2021.
- [302] Sareh Pouryousefi and Jonathan Tallant. Empirical and philosophical reflections on trust. *Journal of the American Philosophical Association*, pages 1–21, 2022.
- [303] Andrew Prahla and Winnie Wen Pin Goh. “rogue machines” and crisis communication: When ai fails, how do companies publicly respond? *Public Relations Review*, 47(4): 102077, 2021.
- [304] David V Pynadath and Stacy C Marsella. Psychsim: Modeling theory of mind with decision-theoretic agents. In *IJCAI*, volume 5, pages 1181–1186, 2005.
- [305] Chakravarti Rajagopalachari. *Mahabharata*, volume 1. Diamond Pocket Books (P) Ltd., 1970.
- [306] Anand Srinivasa Rao and M. Georgeff. Bdi agents: From theory to practice. In *ICMAS*, 1995.
- [307] Bruce Ratner. The correlation coefficient: Its values range between+ 1/- 1, or do they? *Journal of targeting, measurement and analysis for marketing*, 17(2):139–142, 2009.
- [308] John K Rempel, John G Holmes, and Mark P Zanna. Trust in close relationships. *Journal of personality and social psychology*, 49(1):95, 1985.
- [309] Mireia Ribera and Àgata Lapedriza García. Can we do better explanations? a proposal of user-centered explainable ai. *CEUR Workshop Proceedings*, 2019.
- [310] Jens Riegelsberger, M Angela Sasse, and John D McCarthy. The mechanics of trust: A framework for research and design. *International Journal of Human-Computer Studies*, 62(3):381–422, 2005.

- [311] Denise Christine Rieser and Orlando Bernhard. Measuring trust: the simpler the better? In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 2940–2946, 2016.
- [312] Maria Riveiro and Serge Thill. “that’s (not) the output i expected!” on the role of end user expectations in creating explanations of ai systems. *Artificial Intelligence*, 298:103507, 2021.
- [313] Vincent Robbmond, Oana Inel, and Ujwal Gadiraju. Understanding the role of explanation modality in ai-assisted decision-making. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, pages 223–233, 2022.
- [314] Paul Robinette, Wenchen Li, Robert Allen, Ayanna M Howard, and Alan R Wagner. Overtrust of robots in emergency evacuation scenarios. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 101–108. IEEE, 2016.
- [315] Heather M Roff and David Danks. “trust but verify”: The difficulty of trusting autonomous weapons systems. *Journal of Military Ethics*, 17(1):2–20, 2018.
- [316] Denise M Rousseau, Sim B Sitkin, Ronald S Burt, and Colin Camerer. Not so different after all: A cross-discipline view of trust. *Academy of management review*, 23(3): 393–404, 1998.
- [317] Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Pearson, 2016.
- [318] Mark Ryan. In ai we trust: Ethics, artificial intelligence, and reliability. *Science and Engineering Ethics*, pages 1–19, 2020.
- [319] Matthew NO Sadiku, Sarhan M Musa, and A Ajayi-Majebi. *A Primer on Multiple Intelligences*. Springer, 2021.
- [320] E. Salas, Dana E. Sims, and C. Burke. Is there a “big five” in teamwork? *Small Group Research*, 36:555 – 599, 2005.
- [321] Siby Samuel, William J Horrey, and Donald L Fisher. A predictive model of driver response in an autonomous environment. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 59, pages 1671–1675. SAGE Publications Sage CA: Los Angeles, CA, 2015.
- [322] David Sanders and Alexander Gegov. Ai tools for use in assembly automation and some examples of recent applications. *Assembly Automation*, 33(2):184–194, 2013.
- [323] Tracy Sanders, Kristin E Oleson, Deborah R Billings, Jessie YC Chen, and Peter A Hancock. A model of human-robot trust: Theoretical model development. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 55, pages 1432–1436. SAGE Publications Sage CA: Los Angeles, CA, 2011.

- [324] Kristin E Schaefer, Jessie YC Chen, James L Szalma, and Peter A Hancock. A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human factors*, 58(3):377–400, 2016.
- [325] Kristin E. Schaefer, Brandon S. Perelman, Gregory M. Gremillion, Amar R. Marathe, and Jason S. Metcalfe. A roadmap for developing team trust metrics for human-autonomy teams. In *Trust in Human-Robot Interaction*. Academic Press, 2021. ISBN 978-0-12-819472-0. doi: <https://doi.org/10.1016/B978-0-12-819472-0.00012-5>. URL <https://www.sciencedirect.com/science/article/pii/B9780128194720000125>.
- [326] John Schaubroeck, Simon SK Lam, and Ann Chunyan Peng. Cognition-based and affect-based trust as mediators of leader behavior influences on team performance. *Journal of applied psychology*, 96(4):863, 2011.
- [327] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. Appropriate reliance on ai advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, Iui '23, page 410–422, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701061. doi: 10.1145/3581641.3584066. URL <https://doi.org/10.1145/3581641.3584066>.
- [328] Nadine Schlicker and Markus Langer. Towards warranted trust: A model on the relation between actual and perceived system trustworthiness. In *Proceedings of Mensch und Computer 2021*, pages 325–329. 2021.
- [329] Jakob Schoeffer, Niklas Kuehl, and Yvette Machowski. “there is not enough information”: On the effects of explanations on perceptions of informational fairness and trustworthiness in automated decision-making. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1616–1628, 2022.
- [330] F. Schoorman, Roger Mayer, and J. Davis. An integrative model of organizational trust: Past, present, and future. *Academy of Management Review*, 32:344–354, 2007.
- [331] Shalom H Schwartz. An overview of the schwartz theory of basic values. *Online readings in Psychology and Culture*, 2(1):2307–0919, 2012.
- [332] Friso Selten, Marcel Robeer, and Stephan Grimmelikhuijsen. ‘just like i thought’: Street-level bureaucrats trust ai recommendations if they confirm their professional judgment. *Public Administration Review*, 83(2):263–278, 2023.
- [333] Kamran Shafi. A machine competence based analytical model to study trust calibration in supervised autonomous systems. In *2017 Ninth International Conference on Advanced Computational Intelligence (ICACI)*, pages 245–252. Ieee, 2017.
- [334] Daniel Shapiro and Ross Shachter. User-agent value alignment. In *Proc. of The 18th Nat. Conf. on Artif. Intell. AAAI*, 2002.

- [335] Gagan Deep Sharma, Anshita Yadav, and Ritika Chopra. Artificial intelligence and effective governance: A review, critique and research agenda. *Sustainable Futures*, 2, 2020.
- [336] TB Sheridan. Handbook of human factors. In *Supervisory Control*. Wiley-Interscience, 1987.
- [337] Thomas B Sheridan. Extending three existing models to analysis of trust in automation: signal detection, statistical parameter estimation, and model-based control. *Human factors*, 61(7):1162–1170, 2019.
- [338] R Jay Shively, Joel Lachter, Summer L Brandt, Michael Matessa, Vernol Battiste, and Walter W Johnson. Why human-autonomy teaming? In *Advances in Neuroergonomics and Cognitive Engineering: Proceedings of the AHFE 2017 International Conference on Neuroergonomics and Cognitive Engineering, July 17–21, 2017, The Westin Bonaventure Hotel, Los Angeles, California, USA 8*, pages 3–11. Springer, 2018.
- [339] Maayan Shvo, Jakob Buhmann, and Mubbasir Kapadia. Towards modeling the interplay of personality, motivation, emotion, and mood in social agents. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 2195–2197, 2019.
- [340] Keng Siau and Weiyu Wang. Building trust in artificial intelligence, machine learning, and robotics. *Cutter business technology journal*, 31(2):47–53, 2018.
- [341] Michael Siegrist, George Cvetkovich, and Claudia Roth. Salient value similarity, social trust, and risk/benefit perception. *Risk analysis*, 20(3):353–362, 2000.
- [342] Sehajbir Singh and Baljit Singh Saini. Autonomous cars: Recent developments, challenges, and possible solutions. In *IOP conference series: Materials science and engineering*, volume 1022, page 012028. IOP Publishing, 2021.
- [343] Sim B Sitkin and Nancy L Roth. Explaining the limited effectiveness of legalistic “remedies” for trust/distrust. *Organization science*, 4(3):367–392, 1993.
- [344] Venkatesh Sivaraman, Leigh A Bukowski, Joel Levin, Jeremy M. Kahn, and Adam Perer. Ignore, trust, or negotiate: Understanding clinician acceptance of ai-based treatment recommendations in health care. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI ’23, page 1–18, New York, NY, USA, Apr 2023. Association for Computing Machinery. ISBN 978-1-4503-9421-5. doi: 10.1145/3544548.3581075. URL <https://dl.acm.org/doi/10.1145/3544548.3581075>.
- [345] Robert C Solomon and Fernando Flores. *Building trust: In business, politics, relationships, and life*. Oxford University Press, 2003.
- [346] Robert D. Sorkin. FORUM: Why are people turning off our alarms? *The Journal of the Acoustical Society of America*, 84(3):1107–1108, 09 1988. ISSN 0001-4966. doi: 10.1121/1.397232. URL <https://doi.org/10.1121/1.397232>.

- [347] Robert D Sorkin, Barry H Kantowitz, and Susan C Kantowitz. Likelihood alarm displays. *Human Factors*, 30(4):445–459, 1988.
- [348] Randall D Spain, Ernesto A Bustamante, and James P Bliss. Towards an empirically developed scale for system trust: Take two. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 52, pages 1335–1339. SAGE Publications Sage CA: Los Angeles, CA, 2008.
- [349] Fabian Sperrle, Mennatallah El-Assady, Grace Guo, Duen Horng Chau, Alex Ender, and Daniel Keim. Should we trust (x) ai? design dimensions for structured experimental evaluations. *arXiv preprint arXiv:2009.06433*, 2020.
- [350] Randall Steeb and Steven C Johnston. A computer-based interactive system for group decisionmaking. *IEEE Transactions on Systems, Man, and Cybernetics*, 11(8): 544–552, 1981.
- [351] Mary Steffel, Elanor F Williams, and Jaclyn Permann-Graham. Passing the buck: Delegating choices to others to avoid responsibility and blame. *Organizational Behavior and Human Decision Processes*, 135:32–44, 2016.
- [352] Micha Strack and Carsten Gennerich. Personal and situational values predict ethical reasoning. *Europe’s Journal of Psychology*, 7(3):419–442, 2011.
- [353] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. Visual, textual or hybrid: the effect of user expertise on different explanations. In *26th International Conference on Intelligent User Interfaces*, pages 109–119, 2021.
- [354] Jonathan Tallant. Commitment in cases of trust and distrust. *Thought: A Journal of Philosophy*, 6(4):261–267, 2017.
- [355] Jiliang Tang, Xia Hu, and Huan Liu. Is distrust the negation of trust? the value of distrust in social media. In *Proceedings of the 25th ACM conference on Hypertext and social media*, pages 148–157, 2014.
- [356] Gabriele Taylor and Raimond Gaita. Integrity. *Proceedings of the Aristotelian Society, Supplementary Volumes*, 55:143–176, 1981.
- [357] Randy L Teach and Edward H Shortliffe. An analysis of physician attitudes regarding computer-based clinical consultation systems. *Computers and Biomedical Research*, 14(6):542–558, 1981.
- [358] Scott Thiebes, Sebastian Lins, and Ali Sunyaev. Trustworthy artificial intelligence. *Electronic Markets*, 31:447–464, 2021.
- [359] Huaglory Tianfield. Formalized analysis of structural characteristics of large complex systems. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 31(6):559–572, 2001.
- [360] Myrthe L Tielman, Catholijn M Jonker, and M Birna Van Riemsdijk. Deriving norms from actions, values and context. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 2223–2225, 2019.

- [361] Suzanne Tolmeijer, Astrid Weiss, Marc Hanheide, Felix Lindner, Thomas M Powers, Clare Dixon, and Myrthe L Tielman. Taxonomy of trust-relevant failures and mitigation strategies. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 3–12, 2020.
- [362] Suzanne Tolmeijer, Ujwal Gadiraju, Ramya Ghantasala, Akshit Gupta, and Abraham Bernstein. Second chance for a first impression? trust development in intelligent system interaction. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, pages 77–87, 2021.
- [363] Suzanne Tolmeijer, Markus Christen, Serhiy Kandul, Markus Kneer, and Abraham Bernstein. Capable but amoral? comparing ai and human expert collaboration in ethical decision making. In *CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2022.
- [364] Suzanne Tolmeijer, Markus Christen, Serhiy Kandul, Markus Kneer, and Abraham Bernstein. Capable but amoral? comparing ai and human expert collaboration in ethical decision making. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, Chi '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3517732. URL <https://doi.org/10.1145/3491102.3517732>.
- [365] Richard Tomsett, Alun Preece, Dave Braines, Federico Cerutti, Supriyo Chakraborty, Mani Srivastava, Gavin Pearson, and Lance Kaplan. Rapid trust calibration through interpretable and uncertainty-aware ai. *Patterns*, 1(4):100049, 2020.
- [366] Ehsan Toreini, Mhairi Aitken, Kovila Coopamootoo, Karen Elliott, Carlos Gonzalez Zelaya, and Aad Van Moorsel. The relationship between trust in ai and trustworthy machine learning technologies. In *Proceedings of the 2020 ACM conference on Fairness, Accountability, and Transparency (FAccT '20)*, pages 272–283, 2020.
- [367] Abdullah Aman Tutul, Ehsanul Haque Nirjhar, and Theodora Chaspari. Investigating trust in human-machine learning collaboration: A pilot study on estimating public anxiety from speech. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 288–296, 2021.
- [368] Takane Ueno, Yuto Sawa, Yeongdae Kim, Jacqueline Urakami, Hiroki Oura, and Katie Seaborn. Trust in human-ai interaction: Scoping out models, measures, and methods. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–7, 2022.
- [369] Anna-Sophie Ulfert, Eleni Georganta, Carolina Centeio Jorge, Siddharth Mehrotra, and Myrthe Tielman. Shaping a multidisciplinary understanding of team trust in human-ai teams: a theoretical framework. *European Journal of Work and Organizational Psychology*, pages 1–14, 2023.
- [370] Joana Urbano, Ana Paula Rocha, and Eugénio Oliveira. Computing confidence values: Does trust dynamics matter? In *Portuguese Conference on Artificial Intelligence*, pages 520–531. Springer, 2009.

- [371] Joana Urbano, Ana Paula Rocha, and Eugénio Oliveira. A socio-cognitive perspective of trust. In *Agreement Technologies*, pages 419–429. Springer, 2013.
- [372] Joana Urbano, Ana Paula Rocha, and Eugénio Oliveira. The impact of benevolence in computational trust. In *Agreement Technologies*, pages 210–224. Springer, 2013.
- [373] David Callisto Valentine, Iskander Smit, and Euiyoung Kim. Designing for calibrated trust: Exploring the challenges in calibrating trust between users and autonomous vehicles. *Proceedings of the Design Society*, 1:1143–1152, 2021.
- [374] Kees Van Dongen and Peter-Paul van Maanen. Under-reliance on the decision aid: A difference in calibration and attribution between self and aid. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 50, pages 225–229. SAGE Publications Sage CA: Los Angeles, CA, 2006.
- [375] M Birna Van Riemsdijk, Catholijn M Jonker, and Victor Lesser. Creating socially adaptive electronic partners: Interaction, reasoning and ethical challenges. In *Proceedings of the 2015 international conference on autonomous agents and multiagent systems*, pages 1201–1206, 2015.
- [376] Mascha Van’t Wout and Alan G Sanfey. Friend or foe: The effect of implicit trustworthiness judgments in social decision-making. *Cognition*, 108(3):796–803, 2008.
- [377] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. Explanations can reduce overreliance on ai systems during decision-making. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW1), apr 2023. doi: 10.1145/3579605. URL <https://doi.org/10.1145/3579605>.
- [378] Jerry J Vaske, James D Absher, and Alan D Bright. Salient value similarity, social trust and attitudes toward wildland fire management strategies. *Human Ecology Review*, pages 223–232, 2007.
- [379] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. How to evaluate trust in ai-assisted decision making? a survey of empirical methodologies. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–39, 2021.
- [380] Giulia Vilone and Luca Longo. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106, 2021.
- [381] Samuele Vinanzi, Massimiliano Patacchiola, Antonio Chella, and Angelo Cangelosi. Would a robot trust you? developmental robotics model of trust and theory of mind. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374, 4 2019. ISSN 14712970. doi: 10.1098/rstb.2018.0032.
- [382] Ewart J De Visser, Marieke, M M Peeters, Malte, F Jung, Spencer Kohn, Tyler, H Shaw, Richard Pak, and Mark A Neerincx. Towards a theory of longitudinal trust calibration in human-robot teams. *International Journal of Social Robotics*, 12:459–478, 2020. doi: 10.1007/s12369-019-00596-x. URL <https://doi.org/10.1007/s12369-019-00596-x>.

- [383] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.
- [384] Alan R Wagner, Jason Borenstein, and Ayanna Howard. Overtrust in the robotic age. *Communications of the ACM*, 61(9):22–24, 2018.
- [385] Alan R Wagner, Paul Robinette, and Ayanna Howard. Modeling the human-robot trust phenomenon: A conceptual framework based on risk. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(4):1–24, 2018.
- [386] Francesco Walker, Anika Boelhouwer, Tom Alkim, Willem B Verwey, and Marieke H Martens. Changes in trust after driving level 2 automated cars. *Journal of advanced transportation*, 2018, 2018.
- [387] Connie R Wanberg and Paul M Muchinsky. A typology of career decision status: Validity extension of the vocational decision status model. *Journal of Counseling Psychology*, 39(1):71, 1992.
- [388] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. Designing theory-driven user-centric explainable ai. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–15, 2019.
- [389] Lu Wang, Greg A Jamieson, and Justin G Hollands. Improving reliability awareness to support appropriate trust and reliance on individual combat identification systems. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 52, pages 292–296. SAGE Publications Sage CA: Los Angeles, CA, 2008.
- [390] Lu Wang, Greg A Jamieson, and Justin G Hollands. Trust and reliance on an automated combat identification system. *Human factors*, 51(3):281–291, 2009.
- [391] Ning Wang, David V Pynadath, and Susan G Hill. Trust calibration within a human-robot team: Comparing automatically generated explanations. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 109–116, 2016. doi: 10.1109/hri.2016.7451741.
- [392] Ning Wang, David V Pynadath, and Susan G Hill. The impact of pomdp-generated explanations on trust and performance in human-robot teams. In *Proceedings of the 2016 international conference on autonomous agents & multiagent systems*, pages 997–1005, 2016.
- [393] Ning Wang, David V Pynadath, and Susan G Hill. Trust calibration within a human-robot team: Comparing automatically generated explanations. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 109–116. IEEE, 2016.
- [394] Xinru Wang and Ming Yin. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*, pages 318–328, 2021.

- [395] Xinru Wang and Ming Yin. Effects of explanations in ai-assisted decision making: Principles and comparisons. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2022.
- [396] Traci B Warrington, Helen M Caldwell, et al. Building trust to develop competitive advantage in e-business relationships. *Competitiveness Review: An International Business Journal*, 2000.
- [397] Lawrence R Wheelless and Janis Grotz. The measurement of trust and its relationship to self-disclosure. *Human Communication Research*, 3(3):250–257, 1977.
- [398] EL Wiener. Complacency: Is the term useful for air safety. In *Proceedings of the 26th corporate aviation safety seminar*, volume 117, pages 116–125, 1981.
- [399] William Wilson. Suggestions to foster effective consultation within conservation. *Environments*, 32(2):71, 2004.
- [400] Michael Winikoff. Towards trusting autonomous systems. In *International Workshop on Engineering Multi-Agent Systems*, pages 3–20. Springer, 2017.
- [401] Philipp Wintersberger, Dmitrijs Dmitrenko, Clemens Schartmüller, Anna-Katharina Frison, Emanuela Maggioni, Marianna Obrist, and Andreas Riener. S(c)entinel: Monitoring automated vehicles with olfactory reliability displays. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, Iui '19, page 538–546, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362726. doi: 10.1145/3301275.3302332. URL <https://doi.org/10.1145/3301275.3302332>.
- [402] Magdalena Wischniewski, Nicole Krämer, and Emmanuel Müller. Measuring and understanding trust calibrations for automated systems: A survey of the state-of-the-art and future directions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2023.
- [403] Bogdan Wojciszke. Parallels between competence-versus morality-related traits and individualistic versus collectivistic values. *European Journal of Social Psychology*, 27(3):245–256, 1997.
- [404] Jingjun David Xu, Ronald T Cenfetelli, and Karl Aquino. Do different kinds of trust matter? an examination of the three trusting beliefs on satisfaction and purchase behavior in the buyer–seller context. *The Journal of Strategic Information Systems*, 25:15–31, 2016.
- [405] Toshio Yamagishi, Satoshi Akutsu, Kisuk Cho, Yumi Inoue, Yang Li, and Yoshie Matsumoto. Two-component model of general trust: Predicting behavioral trust from attitudinal trust. *Social Cognition*, 33(5):436–458, 2015.
- [406] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L Arendt. How do visual explanations foster end users' appropriate trust in machine learning? In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 189–201, 2020.

- [407] Qian Yang, Yuexing Hao, Kexin Quan, Stephen Yang, Yiran Zhao, Volodymyr Kuleshov, and Fei Wang. Harnessing biomedical literature to calibrate clinicians' trust in ai decision support systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581393. URL <https://doi.org/10.1145/3544548.3581393>.
- [408] X. Jessie Yang, Vaibhav V. Unhelkar, Kevin Li, and Julie A. Shah. Evaluating effects of user experience and system transparency on trust in automation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, Hri '17, page 408–416, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450343367. doi: 10.1145/2909824.3020230. URL <https://doi.org/10.1145/2909824.3020230>.
- [409] Beste F Yuksel, Penny Collisson, and Mary Czerwinski. Brains or beauty: How to engender trust in user-agent interactions. *ACM Transactions on Internet Technology (TOIT)*, 17(1):1–20, 2017.
- [410] Mireia Yurrita, Tim Draws, Agathe Balayn, Dave Murray-Rust, Nava Tintarev, and Alessandro Bozzon. Disentangling fairness perceptions in algorithmic decision-making: The effects of explanations, human oversight, and contestability. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581161. URL <https://doi-org.tudelft.idm.oclc.org/10.1145/3544548.3581161>.
- [411] Lotfi Asker Zadeh. The role of fuzzy logic in the management of uncertainty in expert systems. *Fuzzy sets and systems*, 11(1-3):199–227, 1983.
- [412] Zahra Zahedi, Sarath Sreedharan, and Subbarao Kambhampati. A mental model based theory of trust. *arXiv preprint arXiv:2301.12569*, 2023.
- [413] Qiaoning Zhang, Matthew L Lee, and Scott Carter. You complete me: Human-ai teams and complementary expertise. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, Chi '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3517791. URL <https://doi.org/10.1145/3491102.3517791>.
- [414] Qiaoning Zhang, Matthew L Lee, and Scott Carter. You complete me: Human-ai teams and complementary expertise. In *CHI Conference on Human Factors in Computing Systems*, pages 1–28, 2022.
- [415] Richong Zhang and Yongyi Mao. Trust prediction via belief propagation. *ACM Transactions on Information Systems (TOIS)*, 32(3):1–27, 2014.
- [416] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Fat\* '20, page 295–305, New York, NY, USA, 2020. Association for Com-

- puting Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372852. URL <https://doi.org/10.1145/3351095.3372852>.
- [417] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 295–305, 2020.
- [418] Zelun Tony Zhang, Yuanting Liu, and Heinrich Hussmann. Forward reasoning decision support: Toward a more complete view of the human-ai interaction design space. In *CHIItaly 2021: 14th Biannual Conference of the Italian SIGCHI Chapter*, pages 1–5, 2021.
- [419] Yuhui Zhong, Bharat Bhargava, Yi Lu, and Pelin Angin. A computational dynamic trust model for user authorization. *IEEE Transactions on Dependable and Secure Computing*, 12(1):1–15, 2014.
- [420] Bing Zhu, André Habisch, and John Thøgersen. The importance of cultural values and trust for innovation—a european study. *International Journal of Innovation Management*, 22(02):1850017, 2018.



---

## SIKS DISSERTATIONS

- 2016 01 Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines  
02 Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through  
decision support: prescribing a better pill to swallow  
03 Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge  
Worker Support  
04 Laurens Rietveld (VUA), Publishing and Consuming Linked Data  
05 Evgeny Sherkhonov (UvA), Expanded Acyclic Queries: Containment and an  
Application in Explaining Missing Answers  
06 Michel Wilson (TUD), Robust scheduling in an uncertain environment  
07 Jeroen de Man (VUA), Measuring and modeling negative emotions for virtual  
training  
08 Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social  
Networks from Unstructured Data  
09 Archana Nottamkandath (VUA), Trusting Crowdsourced Information on Cultural  
Artefacts  
10 George Karafotias (VUA), Parameter Control for Evolutionary Algorithms  
11 Anne Schuth (UvA), Search Engines that Learn from Their Users  
12 Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent  
Systems  
13 Nana Baah Gyan (VUA), The Web, Speech Technologies and Rural Development  
in West Africa - An ICT4D Approach  
14 Ravi Khadka (UU), Revisiting Legacy Software System Modernization  
15 Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algo-  
rithms and Experiments  
16 Guangliang Li (UvA), Socially Intelligent Autonomous Agents that Learn from  
Human Reward  
17 Berend Weel (VUA), Towards Embodied Evolution of Robot Organisms  
18 Albert Meroño Peñuela (VUA), Refining Statistical Data on the Web  
19 Julia Efremova (TU/e), Mining Social Structures from Genealogical Data  
20 Daan Odijk (UvA), Context & Semantics in News & Web Search  
21 Alejandro Moreno Celleri (UT), From Traditional to Interactive Playspaces:  
Automatic Analysis of Player Behavior in the Interactive Tag Playground  
22 Grace Lewis (VUA), Software Architecture Strategies for Cyber-Foraging Sys-  
tems  
23 Fei Cai (UvA), Query Auto Completion in Information Retrieval  
24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An  
Iterative and data model independent approach

- 25 Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior
  - 26 Dilhan Thilakarathne (VUA), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains
  - 27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media
  - 28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control
  - 29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning
  - 30 Ruud Mattheij (TiU), The Eyes Have It
  - 31 Mohammad Khelghati (UT), Deep web content monitoring
  - 32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations
  - 33 Peter Bloem (UvA), Single Sample Statistics, exercises in learning from just one example
  - 34 Dennis Schunselaar (TU/e), Configurable Process Trees: Elicitation, Analysis, and Enactment
  - 35 Zhaochun Ren (UvA), Monitoring Social Media: Summarization, Classification and Recommendation
  - 36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies
  - 37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry
  - 38 Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design
  - 39 Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect
  - 40 Christian Detweiler (TUD), Accounting for Values in Design
  - 41 Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance
  - 42 Spyros Martzoukos (UvA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora
  - 43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice
  - 44 Thibault Sellam (UvA), Automatic Assistants for Database Exploration
  - 45 Bram van de Laar (UT), Experiencing Brain-Computer Interface Control
  - 46 Jorge Gallego Perez (UT), Robots to Make you Happy
  - 47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks
  - 48 Tanja Buttler (TUD), Collecting Lessons Learned
  - 49 Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis
  - 50 Yan Wang (TiU), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains
-

- 2017 01 Jan-Jaap Oerlemans (UL), Investigating Cybercrime  
02 Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation  
03 Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines  
04 Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store  
05 Mahdieh Shadi (UvA), Collaboration Behavior  
06 Damir Vandic (EUR), Intelligent Information Systems for Web Product Search  
07 Roel Bertens (UU), Insight in Information: from Abstract to Anomaly  
08 Rob Konijn (VUA), Detecting Interesting Differences: Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery  
09 Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text  
10 Robby van Delden (UT), (Steering) Interactive Play Behavior  
11 Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment  
12 Sander Leemans (TU/e), Robust Process Mining with Guarantees  
13 Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology  
14 Shoshannah Tekofsky (TiU), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior  
15 Peter Berck (RUN), Memory-Based Text Correction  
16 Aleksandr Chuklin (UvA), Understanding and Modeling Users of Modern Search Engines  
17 Daniel Dimov (UL), Crowdsourced Online Dispute Resolution  
18 Ridho Reinanda (UvA), Entity Associations for Search  
19 Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval  
20 Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility  
21 Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)  
22 Sara Magliacane (VUA), Logics for causal inference under uncertainty  
23 David Graus (UvA), Entities of Interest – Discovery in Digital Traces  
24 Chang Wang (TUD), Use of Affordances for Efficient Robot Learning  
25 Veruska Zamborlini (VUA), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search  
26 Merel Jung (UT), Socially intelligent robots that understand and respond to human touch  
27 Michiel Joesse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors  
28 John Klein (VUA), Architecture Practices for Complex Contexts  
29 Adel Alhuraibi (TiU), From IT-Business Strategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"

- 
- 30 Wilma Latuny (TiU), The Power of Facial Expressions  
31 Ben Ruijl (UL), Advances in computational methods for QFT calculations  
32 Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives  
33 Brigit van Loggem (OU), Towards a Design Rationale for Software Documenta-  
tion: A Model of Computer-Mediated Activity  
34 Maren Scheffel (OU), The Evaluation Framework for Learning Analytics  
35 Martine de Vos (VUA), Interpreting natural science spreadsheets  
36 Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-  
throughput Imaging  
37 Alejandro Montes Garcia (TU/e), WiBAF: A Within Browser Adaptation Frame-  
work that Enables Control over Privacy  
38 Alex Kayal (TUD), Normative Social Applications  
39 Sara Ahmadi (RUN), Exploiting properties of the human auditory system and  
compressive sensing methods to increase noise robustness in ASR  
40 Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Hu-  
man Control in Relation to Emotions, Desires and Social Support For applications  
in human-aware support systems  
41 Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental  
Processes and a Smart Environment to Provide Support for a Healthy Lifestyle  
42 Elena Sokolova (RUN), Causal discovery from mixed and missing data with  
applications on ADHD datasets  
43 Maaïke de Boer (RUN), Semantic Mapping in Video Retrieval  
44 Garm Lucassen (UU), Understanding User Stories - Computational Linguistics  
in Agile Requirements Engineering  
45 Bas Testerink (UU), Decentralized Runtime Norm Enforcement  
46 Jan Schneider (OU), Sensor-based Learning Support  
47 Jie Yang (TUD), Crowd Knowledge Creation Acceleration  
48 Angel Suarez (OU), Collaborative inquiry-based learning
- 
- 2018 01 Han van der Aa (VUA), Comparing and Aligning Process Representations  
02 Felix Mannhardt (TU/e), Multi-perspective Process Mining  
03 Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling,  
Model-Driven Development of Context-Aware Applications, and Behavior Pre-  
diction  
04 Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in  
Data-Centric Engineering Tasks  
05 Hugo Huurdeman (UvA), Supporting the Complex Dynamics of the Information  
Seeking Process  
06 Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-  
Technical Systems  
07 Jieting Luo (UU), A formal account of opportunism in multi-agent systems  
08 Rick Smetsers (RUN), Advances in Model Learning for Software Systems  
09 Xu Xie (TUD), Data Assimilation in Discrete Event Simulations  
10 Julienka Mollee (VUA), Moving forward: supporting physical activity behavior  
change through intelligent technology

- 11 Mahdi Sargolzaei (UvA), Enabling Framework for Service-oriented Collaborative Networks
  - 12 Xixi Lu (TU/e), Using behavioral context in process mining
  - 13 Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future
  - 14 Bart Joosten (TiU), Detecting Social Signals with Spatiotemporal Gabor Filters
  - 15 Naser Davarzani (UM), Biomarker discovery in heart failure
  - 16 Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children
  - 17 Jianpeng Zhang (TU/e), On Graph Sample Clustering
  - 18 Henriette Nakad (UL), De Notaris en Private Rechtspraak
  - 19 Minh Duc Pham (VUA), Emergent relational schemas for RDF
  - 20 Manxia Liu (RUN), Time and Bayesian Networks
  - 21 Aad Slootmaker (OU), EMERGO: a generic platform for authoring and playing scenario-based serious games
  - 22 Eric Fernandes de Mello Araújo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks
  - 23 Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis
  - 24 Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots
  - 25 Riste Gligorov (VUA), Serious Games in Audio-Visual Collections
  - 26 Roelof Anne Jelle de Vries (UT), Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology
  - 27 Maikel Leemans (TU/e), Hierarchical Process Mining for Scalable Software Analysis
  - 28 Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel
  - 29 Yu Gu (TiU), Emotion Recognition from Mandarin Speech
  - 30 Wouter Beek (VUA), The "K" in "semantic web" stands for "knowledge": scaling semantics to the web
- 
- 2019 01 Rob van Eijk (UL), Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification
  - 02 Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty
  - 03 Eduardo Gonzalez Lopez de Murillas (TU/e), Process Mining on Databases: Extracting Event Data from Real Life Data Sources
  - 04 Ridho Rahmadi (RUN), Finding stable causal structures from clinical data
  - 05 Sebastiaan van Zelst (TU/e), Process Mining with Streaming Data
  - 06 Chris Dijkshoorn (VUA), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets
  - 07 Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms
  - 08 Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes
  - 09 Fahimeh Alizadeh Moghaddam (UvA), Self-adaptation for energy efficiency in software systems

- 10 Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction
  - 11 Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs
  - 12 Jacqueline Heinerman (VUA), Better Together
  - 13 Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation
  - 14 Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses
  - 15 Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments
  - 16 Guangming Li (TU/e), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models
  - 17 Ali Hurriyetoglu (RUN), Extracting actionable information from microtexts
  - 18 Gerard Wagenaar (UU), Artefacts in Agile Team Communication
  - 19 Vincent Koeman (TUD), Tools for Developing Cognitive Agents
  - 20 Chide Groenouwe (UU), Fostering technically augmented human collective intelligence
  - 21 Cong Liu (TU/e), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection
  - 22 Martin van den Berg (VUA), Improving IT Decisions with Enterprise Architecture
  - 23 Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification
  - 24 Anca Dumitrache (VUA), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing
  - 25 Emiel van Miltenburg (VUA), Pragmatic factors in (automatic) image description
  - 26 Prince Singh (UT), An Integration Platform for Synchromodal Transport
  - 27 Alessandra Antonaci (OU), The Gamification Design Process applied to (Massive) Open Online Courses
  - 28 Esther Kuindersma (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations
  - 29 Daniel Formolo (VUA), Using virtual agents for simulation and training of social skills in safety-critical circumstances
  - 30 Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems
  - 31 Milan Jelisavcic (VUA), Alive and Kicking: Baby Steps in Robotics
  - 32 Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games
  - 33 Anil Yaman (TU/e), Evolution of Biologically Inspired Learning in Artificial Neural Networks
  - 34 Negar Ahmadi (TU/e), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES
  - 35 Lisa Facey-Shaw (OU), Gamification with digital badges in learning programming
  - 36 Kevin Ackermans (OU), Designing Video-Enhanced Rubrics to Master Complex Skills
  - 37 Jian Fang (TUD), Database Acceleration on FPGAs
  - 38 Akos Kadar (OU), Learning visually grounded and multilingual representations
-

- 2020 01 Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour
- 02 Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models
- 03 Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding
- 04 Maarten van Gompel (RUN), Context as Linguistic Bridges
- 05 Yulong Pei (TU/e), On local and global structure mining
- 06 Preethu Rose Anish (UT), Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support
- 07 Wim van der Vegt (OU), Towards a software architecture for reusable game components
- 08 Ali Mirsoleimani (UL), Structured Parallel Programming for Monte Carlo Tree Search
- 09 Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research
- 10 Alifah Syamsiyah (TU/e), In-database Preprocessing for Process Mining
- 11 Sepideh Mesbah (TUD), Semantic-Enhanced Training Data Augmentation Methods for Long-Tail Entity Recognition Models
- 12 Ward van Breda (VUA), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment
- 13 Marco Virgolin (CWI), Design and Application of Gene-pool Optimal Mixing Evolutionary Algorithms for Genetic Programming
- 14 Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases
- 15 Konstantinos Georgiadis (OU), Smart CAT: Machine Learning for Configurable Assessments in Serious Games
- 16 Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling
- 17 Daniele Di Mitri (OU), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences
- 18 Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems
- 19 Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems
- 20 Albert Hankel (VUA), Embedding Green ICT Maturity in Organisations
- 21 Karine da Silva Miras de Araujo (VUA), Where is the robot?: Life as it could be
- 22 Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar
- 23 Rianne Conijn (UT), The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging
- 24 Lenin da Nóbrega Medeiros (VUA/RUN), How are you feeling, human? Towards emotionally supportive chatbots
- 25 Xin Du (TU/e), The Uncertainty in Exceptional Model Mining
- 26 Krzysztof Leszek Sadowski (UU), GAMBIT: Genetic Algorithm for Model-Based mixed-Integer opTimization
- 27 Ekaterina Muravyeva (TUD), Personal data and informed consent in an educational context

- 
- 28 Bibeg Limbu (TUD), Multimodal interaction for deliberate practice: Training complex skills with augmented reality
  - 29 Ioan Gabriel Bucur (RUN), Being Bayesian about Causal Inference
  - 30 Bob Zadok Blok (UL), Creatief, Creatiever, Creatiefst
  - 31 Gongjin Lan (VUA), Learning better – From Baby to Better
  - 32 Jason Rhuggenaath (TU/e), Revenue management in online markets: pricing and online advertising
  - 33 Rick Gilsing (TU/e), Supporting service-dominant business model evaluation in the context of business model innovation
  - 34 Anna Bon (UM), Intervention or Collaboration? Redesigning Information and Communication Technologies for Development
  - 35 Siamak Farshidi (UU), Multi-Criteria Decision-Making in Software Production
- 
- 2021 01 Francisco Xavier Dos Santos Fonseca (TUD), Location-based Games for Social Interaction in Public Space
  - 02 Rijk Mercur (TUD), Simulating Human Routines: Integrating Social Practice Theory in Agent-Based Models
  - 03 Seyyed Hadi Hashemi (UvA), Modeling Users Interacting with Smart Devices
  - 04 Ioana Jivet (OU), The Dashboard That Loved Me: Designing adaptive learning analytics for self-regulated learning
  - 05 Davide Dell'Anna (UU), Data-Driven Supervision of Autonomous Systems
  - 06 Daniel Davison (UT), "Hey robot, what do you think?" How children learn with a social robot
  - 07 Armel Lefebvre (UU), Research data management for open science
  - 08 Nardie Fanchamps (OU), The Influence of Sense-Reason-Act Programming on Computational Thinking
  - 09 Cristina Zaga (UT), The Design of Robothings. Non-Anthropomorphic and Non-Verbal Robots to Promote Children's Collaboration Through Play
  - 10 Quinten Meertens (UvA), Misclassification Bias in Statistical Learning
  - 11 Anne van Rossum (UL), Nonparametric Bayesian Methods in Robotic Vision
  - 12 Lei Pi (UL), External Knowledge Absorption in Chinese SMEs
  - 13 Bob R. Schadenberg (UT), Robots for Autistic Children: Understanding and Facilitating Predictability for Engagement in Learning
  - 14 Negin Samaeemofrad (UL), Business Incubators: The Impact of Their Support
  - 15 Onat Ege Adali (TU/e), Transformation of Value Propositions into Resource Re-Configurations through the Business Services Paradigm
  - 16 Esam A. H. Ghaleb (UM), Bimodal emotion recognition from audio-visual cues
  - 17 Dario Dotti (UM), Human Behavior Understanding from motion and bodily cues using deep neural networks
  - 18 Remi Wieten (UU), Bridging the Gap Between Informal Sense-Making Tools and Formal Systems - Facilitating the Construction of Bayesian Networks and Argumentation Frameworks
  - 19 Roberto Verdecchia (VUA), Architectural Technical Debt: Identification and Management
  - 20 Masoud Mansoury (TU/e), Understanding and Mitigating Multi-Sided Exposure Bias in Recommender Systems

- 
- 21 Pedro Thiago Timbó Holanda (CWI), Progressive Indexes
  - 22 Sihang Qiu (TUD), Conversational Crowdsourcing
  - 23 Hugo Manuel Proença (UL), Robust rules for prediction and description
  - 24 Kaijie Zhu (TU/e), On Efficient Temporal Subgraph Query Processing
  - 25 Eoin Martino Grua (VUA), The Future of E-Health is Mobile: Combining AI and Self-Adaptation to Create Adaptive E-Health Mobile Applications
  - 26 Benno Kruit (CWI/VUA), Reading the Grid: Extending Knowledge Bases from Human-readable Tables
  - 27 Jelte van Waterschoot (UT), Personalized and Personal Conversations: Designing Agents Who Want to Connect With You
  - 28 Christoph Selig (UL), Understanding the Heterogeneity of Corporate Entrepreneurship Programs
- 
- 2022 01 Judith van Stegeren (UT), Flavor text generation for role-playing video games
  - 02 Paulo da Costa (TU/e), Data-driven Prognostics and Logistics Optimisation: A Deep Learning Journey
  - 03 Ali el Hassouni (VUA), A Model A Day Keeps The Doctor Away: Reinforcement Learning For Personalized Healthcare
  - 04 Ünal Aksu (UU), A Cross-Organizational Process Mining Framework
  - 05 Shiwei Liu (TU/e), Sparse Neural Network Training with In-Time Over-Parameterization
  - 06 Reza Refaei Afshar (TU/e), Machine Learning for Ad Publishers in Real Time Bidding
  - 07 Sambit Praharaaj (OU), Measuring the Unmeasurable? Towards Automatic Co-located Collaboration Analytics
  - 08 Maikel L. van Eck (TU/e), Process Mining for Smart Product Design
  - 09 Oana Andreea Inel (VUA), Understanding Events: A Diversity-driven Human-Machine Approach
  - 10 Felipe Moraes Gomes (TUD), Examining the Effectiveness of Collaborative Search Engines
  - 11 Mirjam de Haas (UT), Staying engaged in child-robot interaction, a quantitative approach to studying preschoolers' engagement with robots and tasks during second-language tutoring
  - 12 Guanyi Chen (UU), Computational Generation of Chinese Noun Phrases
  - 13 Xander Wilcke (VUA), Machine Learning on Multimodal Knowledge Graphs: Opportunities, Challenges, and Methods for Learning on Real-World Heterogeneous and Spatially-Oriented Knowledge
  - 14 Michiel Overeem (UU), Evolution of Low-Code Platforms
  - 15 Jelmer Jan Koorn (UU), Work in Process: Unearthing Meaning using Process Mining
  - 16 Pieter Gijbbers (TU/e), Systems for AutoML Research
  - 17 Laura van der Lubbe (VUA), Empowering vulnerable people with serious games and gamification
  - 18 Paris Mavromoustakos Blom (TiU), Player Affect Modelling and Video Game Personalisation
  - 19 Bilge Yigit Ozkan (UU), Cybersecurity Maturity Assessment and Standardisation

- 
- 20 Fakhra Jabeen (VUA), Dark Side of the Digital Media - Computational Analysis of Negative Human Behaviors on Social Media
- 21 Seethu Mariyam Christopher (UM), Intelligent Toys for Physical and Cognitive Assessments
- 22 Alexandra Sierra Rativa (TiU), Virtual Character Design and its potential to foster Empathy, Immersion, and Collaboration Skills in Video Games and Virtual Reality Simulations
- 23 Ilir Kola (TUD), Enabling Social Situation Awareness in Support Agents
- 24 Samaneh Heidari (UU), Agents with Social Norms and Values - A framework for agent based social simulations with social norms and personal values
- 25 Anna L.D. Latour (UL), Optimal decision-making under constraints and uncertainty
- 26 Anne Dirkson (UL), Knowledge Discovery from Patient Forums: Gaining novel medical insights from patient experiences
- 27 Christos Athanasiadis (UM), Emotion-aware cross-modal domain adaptation in video sequences
- 28 Onuralp Ulusoy (UU), Privacy in Collaborative Systems
- 29 Jan Kolkmeier (UT), From Head Transform to Mind Transplant: Social Interactions in Mixed Reality
- 30 Dean De Leo (CWI), Analysis of Dynamic Graphs on Sparse Arrays
- 31 Konstantinos Traganos (TU/e), Tackling Complexity in Smart Manufacturing with Advanced Manufacturing Process Management
- 32 Cezara Pastrav (UU), Social simulation for socio-ecological systems
- 33 Brinn Hekkelman (CWI/TUD), Fair Mechanisms for Smart Grid Congestion Management
- 34 Nimat Ullah (VUA), Mind Your Behaviour: Computational Modelling of Emotion & Desire Regulation for Behaviour Change
- 35 Mike E.U. Ligthart (VUA), Shaping the Child-Robot Relationship: Interaction Design Patterns for a Sustainable Interaction
- 
- 2023 01 Bojan Simoski (VUA), Untangling the Puzzle of Digital Health Interventions
- 02 Mariana Rachel Dias da Silva (TiU), Grounded or in flight? What our bodies can tell us about the whereabouts of our thoughts
- 03 Shabnam Najafian (TUD), User Modeling for Privacy-preserving Explanations in Group Recommendations
- 04 Gineke Wiggers (UL), The Relevance of Impact: bibliometric-enhanced legal information retrieval
- 05 Anton Bouter (CWI), Optimal Mixing Evolutionary Algorithms for Large-Scale Real-Valued Optimization, Including Real-World Medical Applications
- 06 António Pereira Barata (UL), Reliable and Fair Machine Learning for Risk Assessment
- 07 Tianjin Huang (TU/e), The Roles of Adversarial Examples on Trustworthiness of Deep Learning
- 08 Lu Yin (TU/e), Knowledge Elicitation using Psychometric Learning
- 09 Xu Wang (VUA), Scientific Dataset Recommendation with Semantic Techniques

- 10 Dennis J.N.J. Soemers (UM), Learning State-Action Features for General Game Playing
  - 11 Fawad Taj (VUA), Towards Motivating Machines: Computational Modeling of the Mechanism of Actions for Effective Digital Health Behavior Change Applications
  - 12 Tessel Bogaard (VUA), Using Metadata to Understand Search Behavior in Digital Libraries
  - 13 Injy Sarhan (UU), Open Information Extraction for Knowledge Representation
  - 14 Selma Čaušević (TUD), Energy resilience through self-organization
  - 15 Alvaro Henrique Chaim Correia (TU/e), Insights on Learning Tractable Probabilistic Graphical Models
  - 16 Peter Blomsma (TiU), Building Embodied Conversational Agents: Observations on human nonverbal behaviour as a resource for the development of artificial characters
  - 17 Meike Nauta (UT), Explainable AI and Interpretable Computer Vision – From Oversight to Insight
  - 18 Gustavo Penha (TUD), Designing and Diagnosing Models for Conversational Search and Recommendation
  - 19 George Aalbers (TiU), Digital Traces of the Mind: Using Smartphones to Capture Signals of Well-Being in Individuals
  - 20 Arkadiy Dushatskiy (TUD), Expensive Optimization with Model-Based Evolutionary Algorithms applied to Medical Image Segmentation using Deep Learning
  - 21 Gerrit Jan de Bruin (UL), Network Analysis Methods for Smart Inspection in the Transport Domain
  - 22 Alireza Shojaiifar (UU), Volitional Cybersecurity
  - 23 Theo Theunissen (UU), Documentation in Continuous Software Development
  - 24 Agathe Balayn (TUD), Practices Towards Hazardous Failure Diagnosis in Machine Learning
  - 25 Jurian Baas (UU), Entity Resolution on Historical Knowledge Graphs
  - 26 Loek Tonnaer (TU/e), Linearly Symmetry-Based Disentangled Representations and their Out-of-Distribution Behaviour
  - 27 Ghada Sokar (TU/e), Learning Continually Under Changing Data Distributions
  - 28 Floris den Hengst (VUA), Learning to Behave: Reinforcement Learning in Human Contexts
  - 29 Tim Draws (TUD), Understanding Viewpoint Biases in Web Search Results
- 
- 2024 01 Daphne Miedema (TU/e), On Learning SQL: Disentangling concepts in data systems education
  - 02 Emile van Krieken (VUA), Optimisation in Neurosymbolic Learning Systems
  - 03 Feri Wijayanto (RUN), Automated Model Selection for Rasch and Mediation Analysis
  - 04 Mike Huisman (UL), Understanding Deep Meta-Learning
  - 05 Yiyong Gou (UM), Aerial Robotic Operations: Multi-environment Cooperative Inspection & Construction Crack Autonomous Repair
  - 06 Azqa Nadeem (TUD), Understanding Adversary Behavior via XAI: Leveraging Sequence Clustering to Extract Threat Intelligence

- 07 Parisa Shayan (TiU), Modeling User Behavior in Learning Management Systems
- 08 Xin Zhou (UvA), From Empowering to Motivating: Enhancing Policy Enforcement through Process Design and Incentive Implementation
- 09 Giso Dal (UT), Probabilistic Inference Using Partitioned Bayesian Networks
- 10 Cristina-Iulia Bucur (VUA), Linkflows: Towards Genuine Semantic Publishing in Science
- 11 Mahmoud Shokrollahi-Far (TiU), Computational Reliability of Quranic Grammar
- 12 Peide Zhu (TUD), Towards Robust Automatic Question Generation For Learning
- 13 Enrico Liscio (TUD), Context-Specific Value Inference via Hybrid Intelligence
- 14 Larissa Capobianco Shimomura (TU/e), On Graph Generating Dependencies and their Applications in Data Profiling
- 15 Ting Liu (VUA), A Gut Feeling: Biomedical Knowledge Graphs for Interrelating the Gut Microbiome and Mental Health
- 16 Arthur Barbosa Câmara (TUD), Designing Search-as-Learning Systems
- 17 Razieh Alidoosti (VUA), Ethics-aware Software Architecture Design
- 18 Laurens Stoop (UU), Data Driven Understanding of Energy-Meteorological Variability and its Impact on Energy System Operations
- 19 Azadeh Mozafari Mehr (TU/e), Multi-perspective Conformance Checking: Identifying and Understanding Patterns of Anomalous Behavior
- 20 Ritsart Anne Plantenga (UL), Omgang met Regels
- 21 Federica Vinella (UU), Crowdsourcing User-Centered Teams
- 22 Zeynep Ozturk Yurt (TU/e), Beyond Routine: Extending BPM for Knowledge-Intensive Processes with Controllable Dynamic Contexts
- 23 Jie Luo (VUA), Lamarck's Revenge: Inheritance of Learned Traits Improves Robot Evolution
- 24 Nirmal Roy (TUD), Exploring the effects of interactive interfaces on user search behaviour
- 25 Alisa Rieger (TUD), Striving for Responsible Opinion Formation in Web Search on Debated Topics
- 26 Tim Gubner (CWI), Adaptively Generating Heterogeneous Execution Strategies using the VOILA Framework
- 27 Lincen Yang (UL), Information-theoretic Partition-based Models for Interpretable Machine Learning
- 28 Leon Helwerda (UL), Grip on Software: Understanding development progress of Scrum sprints and backlogs
- 29 David Wilson Romero Guzman (VUA), The Good, the Efficient and the Inductive Biases: Exploring Efficiency in Deep Learning Through the Use of Inductive Biases
- 30 Vijanti Ramautar (UU), Model-Driven Sustainability Accounting
- 31 Ziyu Li (TUD), On the Utility of Metadata to Optimize Machine Learning Workflows
- 32 Vinicius Stein Dani (UU), The Alpha and Omega of Process Mining
- 33 Siddharth Mehrotra (TUD), Designing for Appropriate Trust in Human-AI interaction

---

## ACKNOWLEDGMENTS

Embarking on a PhD journey abroad is like navigating a labyrinth of intellectual and personal growth. As an international candidate, I've grappled with academic challenges while reconciling my cultural background with a new worldview. This odyssey has transformed my character, beliefs, and understanding. Throughout, I've received invaluable support from numerous individuals, whose collective impact on my research and personal development has been immeasurable. While acknowledging each person's contribution fully would be impractical, I am deeply indebted to all who have played a part in my academic pursuit. As I write these acknowledgments, I'm filled with profound gratitude for my metamorphosis and the many individuals integral to this process. I invite you to join me in recognizing those who have supported me during this extraordinary adventure.

First and foremost, I extend my deepest gratitude to my promoter, Catholijn, and my daily supervisor, Myrthe, for granting me this invaluable PhD opportunity. The commencement of my academic journey amidst the unprecedented COVID-19 lockdown was far from the typical start any PhD candidate would envision. Despite these challenging circumstances, I am profoundly thankful to both of you for transforming this PhD experience into a rewarding and enriching journey.

Myrthe, you are an exceptional supervisor, excelling both professionally and interpersonally. Your support has been crucial, especially during challenging times. From the start, your commitment to monitoring my progress and offering guidance has been the super helpful of this PhD. Our demanding discussions have fostered both personal and professional growth. Your ability to challenge me intellectually while providing emotional support struck a perfect balance, pushing me to exceed expectations without feeling overwhelmed.

Catholijn, I have a lot of respect and admiration for you. Your support both academically and personally, has been a cornerstone of my PhD journey. Your generous spirit have made working with you an extraordinary privilege. The wealth of knowledge and wisdom I've gained under your guidance is immeasurable, and I aspire to apply these lessons and pay them forward throughout my career. Your mentorship has left an indelible mark on my life, a blissful imprint that will guide me for years to come.

I want to thank my external defense committee members: Prof. Roeser, Prof. Croon, Prof. Yolum, and Prof. Winikoff. Thank you for your feedback and especially for the excellent suggestions in improving this thesis.

I would also like to express my appreciation to Ujwal Gadiraju who has been a driving force for me in this PhD. Our countless discussions and common efforts to design user studies, formulate research questions, develop experimental setups, and writing all exciting findings have greatly contributed to the successful realization of this thesis.

I want to thank my PhD mentors, Roel and Elif, for their kindness and humbleness. Both of you patiently listened to me, understood my concerns and appreciations, and paid attention to every tiny detail I shared. Having your support during different phases of

my PhD has been invaluable in maintaining an excellent work-life balance and receiving outstanding career advice. Your guidance has significantly contributed to my growth both professionally and personally throughout this journey.

I want to thank all the collaborators I had the pleasure of working with over the last four years. Carolina, I'm grateful we met and had various opportunities to work together. Your support has been unwavering. Beyond academic support, you're a dear friend whom I will cherish throughout my life. Chadha, I enjoyed our discussions and am grateful for your support. Your patience and perseverance as a collaborator in this complex research are qualities I highly value. Miriea, you are an amazing person and have a happy soul, I loved our engaging discussions and it was lovely working with you. Eleni and Anna-Sophie, thank you for being excellent collaborators. Our insightful discussions, combining perspectives from different disciplines on the same topic, have been fruitful in broadening my vision for this research. Your contributions have added significant depth and interdisciplinary value to our work.

Eva and the amazing group - Maren, Lucas, Izabel, Navid, and Carina - who hosted me in Hamburg, thank you for having me there. I enjoyed our discussions both during meetings and lunch talks. Suzanne, my academic cousin, thank you for helping me arrange the Hamburg visit and for being a support line throughout this PhD; your support is invaluable. Oleksandra, you have been an amazing collaborator, and I have gained great research skills from you. Folkert, this last phase of my PhD would not have been possible without you. You've been so kind to help me with the experiment design alongside your regular work; thank you for all your support.

I want to thank my student, Alan with whom I got an opportunity to work together and all the TAs from the HCI course who helped me during last 3 years to manage the labs with 300+ students, you all are amazing. I also thank my colleague, Garrett for taking over my responsibilities in the last year. I have learned something from every one of you.

Finally, this PhD journey is incomplete without acknowledging the unwavering support of my colleagues from the Interactive Intelligence group. Along the all the great intellectual discussions that I had with so many of you, I'll cherish the friendship that I am taking with me. Thank you Ruben, Pei-Yu and Mo for all the lovely bi-weekly discussions and helping our friendship grow stronger day by day. Thank you Enrico (a friend with best advice and a tidy roommate), Nele (bayesian and salad expert), Masha (my go-to user study participant), Mani (your humor man), Sietze (its so much fun discussing anything and everything with you), Amir (our never ending coffee conversations), Michaël (the only one who can correctly understand Cricket), Emma (always a helpful friend), Morita (YouTube superstar), Laxmi, Paul, Deborah, Jinke, Sandy, Davide, Zuzanna, Rolf, Antonio, Stephanie, Edgar, Agnes, Yu-Wen, Joanna, Aleks, Ilir (my first PhD friend), Miguel, Fran, Merijn, Jasper, Marieke (Everything started with your support), Tina (thank you for the study design), and all the other present and past II members. A special thank you to all the faculty members of the group, Willem-Paul, Mark (thank you for the availability as reserve member), Frans, Luciano (a special thank you for presenting my poster), Catha, Pradeep and Luuk. Thanks also to the support staff, Ruud, Bart, Wouter, and especially Anita (this group is incomplete without you).

The Hybrid Intelligence consortium has become an integral part of my academic journey. Our numerous meet-ups fostered a sense of camaraderie and created lasting connections.

The quality time we spent together, particularly during our gatherings in Vlieland, has been invaluable. Thank you Tae (Procrastinator), Chirag (missing Vada pav?), Tiffany (It was great discussing life and PhD), Davide (the proactive one), Íñigo, Selene, Bram, Urja, Wijnand, Emre, Kata, Ludi, Sharvaree, J.D., Annet, Niklas, Maria, Cor, Loan, Merle, Putra, Nicole, Mark, Delaram, Anna, Andreas, Johanna and all other present and past HI members. Thank you Frank and Wendy for all the superb organization.

A special thanks to my WIS friends, Agathe (You know what you mean to me & Srishti), Tim (always to the correct point), Alisa (we started and we'll finish together), Esra (thank you for helping me with the last project), Lorenzo, Sara (ASPECT expert), and Shreyan. And thank you to Ashwin (HRI, thanks for helping with the mock), Sem (giving me helpful career tips), VP (insightful life discussions), Prof. Marianna (being a power house of positive energy) and Jim (AI lab Ideation colleague).

A special thank you to my closest friends in the Netherlands - Nikita, Shubham, Anurima, Sanket and Nakul. Also, my dear friends from Finland, Priyank and Richa. You've become like a second family to me. As we conclude this season, I couldn't have asked for better companions to help me make the best work-life balance. Ambika, you are my oldest friend. Thank you for visiting us with Jacob. Our journey is a long one! Pooja, Owais, Saurav and Nishit, I miss our time together. Leaving Germany was a hard decision, but I'm glad I met all of you and cherish our friendship. There are too many to name individually, but each of you has made my journey meaningful through your support and shared moments.

To my friends—Vasu, Anuj Sirji, Himendra, Kathan, Arpit, Mahima, Aman, Shubham—thank you from the bottom of my heart. The moments we've shared back at home have been more than just memories; they've been the foundation of a family-like bond that has carried me through this journey. If I've missed anyone's name, know that you are still cherished and appreciated. This journey wouldn't have been as meaningful or memorable without each of you.

I want to extend my deepest gratitude to my family in India. Mummy, Papa, Mummyji, Papaji, Didi, Jiju — your unconditional love and blessings have been my constant source of strength, making every challenge seem surmountable. To my brother and sister-in-law, your unwavering support and the joy brought by little Vivaan have brightened my most exhausting days. And to all my loved ones back home in India, your love and values have been my guiding light. The one whom I can rely on during both challenging and wonderful times, Kavya, my sister, bhai loves you!

Finally, Srishti, my life partner and lovely wife —words can hardly capture the depth of my gratitude. Your love, patience, and unwavering belief in me have been my anchor throughout this journey. You've stood by me through every high and low, offering not just support, but also the strength to keep going. You brought our beautiful daughter, Virika, into this world, filling our lives with even more joy and purpose. I hope that one day, when she reads this, she will be proud of what we've achieved together. Thank you for making our lives so incredibly rich with love and meaning.

Thanks to everyone who helped me finish this ride in style, even those I may have forgotten to mention!



# CURRICULUM VITÆ

## Siddharth MEHROTRA

17/10/1993      Born in Shahjahanpur, India.

### EDUCATION

2020–2024      **Ph.D. in Computer Science**  
Delft University of Technology, the Netherlands

2017–2020      **Master of Science in Media Informatics**  
RWTH Aachen University, Germany

2012–2015      **Bachelor of Science in Computer Science**  
Galgotias University, India

### EXPERIENCE

2023              **University of Hamburg**, Hamburg, Germany  
Visiting Researcher

2020              **Mazda**, Leverkusen, Germany  
Research Intern

2019              **University of Bonn**, Bonn, Germany  
Research Assistant

2018–2019      **Grohe**, Düsseldorf, Germany  
Working Student

2016–2017      **Siemens R&D**, Bangalore, India  
Research Intern

2012–2015      **Microsoft Student Partner**, India  
Member App Review Board



# LIST OF PUBLICATIONS

## IN REVIEW

- 1. **Siddharth Mehrotra**, Folkert van Delden, Eva Bittner, Ujwal Gadiraju, Catholijn Jonker, Myrthe Tielman. Fostering Appropriate Trust in AI-based Predictive Policing Systems: A Case-Study. Under review at *30th Annual ACM Conference on Intelligent User Interfaces (IUI '25)*; [🎤 Invited talk at *4th TAILOR Conference – Trustworthy AI from lab to market (2024)* in Lisbon, Portugal].

## 2024

- 1. **Siddharth Mehrotra**, Chadha Degachi, Oleksandra Vereschak, Catholijn Jonker, Myrthe Tielman. A Systematic Review on Fostering Appropriate Trust in Human-AI Interaction: Trends, Opportunities and Challenges in *ACM Journal of Responsible Computing (JRC) 2024* [This work was presented at *NWO ICT.Open conference (2024)* in Utrecht, Netherlands].
- 2. **Siddharth Mehrotra\***, Chadha Degachi\*, Mireia Yurrita Semperena, Evangelos Niforatos, Myrthe Tielman. Practising Appropriate Trust in Human-Centred AI Design. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, CHI '24 (Extended Abstracts)*, Honolulu, USA.
- 3. Carolina Centeio Jorge, Emma M. van Zoelen, Ruben Verhagen, **Siddharth Mehrotra**, Catholijn Jonker, and Myrthe Tielman. Appropriate context-dependent artificial trust in human-machine teamwork. Book chapter in *Putting AI in the Critical Loop*, pp. 41-60. Academic Press, 2024.

## 2023

- 1. **Siddharth Mehrotra**, Carolina Centeio Jorge, Catholijn M. Jonker, and Myrthe L. Tielman. Integrity-based Explanations for Fostering Appropriate Trust in AI Agents in *ACM Transactions on Interactive Intelligent Systems* 14, no. 1 (2024): 1-36.
- 2. **Siddharth Mehrotra**, Carolina Centeio Jorge, Catholijn M. Jonker, and Myrthe L. Tielman. "Building Appropriate Trust in AI: The Significance of Integrity-Centered Explanations." in *HFAI 2023 conference*, pp. 436-439. 2023. [Best Poster Award]
- 3. Anna-Sophie Ulfert, Eleni Georganta, Carolina Centeio Jorge, **Siddharth Mehrotra**, and Myrthe Tielman. Shaping a multidisciplinary understanding of team trust in human-AI teams: a theoretical framework in *European Journal of Work and Organizational Psychology (EJWOP)* (2023): 1-14.
- 4. **Siddharth Mehrotra**, Carolina Centeio Jorge, Catholijn M. Jonker, and Myrthe L. Tielman. 2023. Integrity Based Explanations for Fostering Appropriate Trust in AI Agents. In *35th*

*Benelux Conference on Artificial Intelligence and 32nd Belgian-Dutch Conference on Machine Learning, BNAIC '23, Delft, the Netherlands, 1-4.*

## 2022

1. Ruben Verhagen, **Siddharth Mehrotra**, Mark Neerinx, Catholijn Jonker, and Myrthe Tielman. Exploring Effectiveness of Explanations for Appropriate Trust: Lessons from Cognitive Psychology in *Workshop on TRust and EXPertise in Visualization co-located with IEEE Visualization Conference* Oklahoma City, USA (2021).

## 2021

1. **Siddharth Mehrotra**, Catholijn M. Jonker, and Myrthe L. Tielman. More similar values, more trust?-the effect of value similarity on trust in human-agent interaction in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pp. 777-783. 2021.
2. **Siddharth Mehrotra**, Modelling Trust in Human-AI Interaction: Doctoral Consortium Track. in *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '21*, Online, IFAAMAS.
3. Carolina Centeio Jorge, **Siddharth Mehrotra**, Myrthe L. Tielman, and Catholijn M. Jonker. "Trust should correspond to trustworthiness: a formalization of appropriate mutual trust in human-agent teams." in *Proceedings of the 22nd International Workshop on Trust in Agent Societies (TRUST 2021)* co-located with the *20th International Conferences on Autonomous Agents and Multiagent Systems (AAMAS 2021)*.

 Included in this Thesis.

 Won a best poster award/ Invited Talk.

\* Equal contribution.



