Human Perception of Trust in AI

ENFIELD Human-centric AI Workshop
Bucharest, September 9 2025

Sebastian Heil, Xavier Carpent, Steven Furnell

UNIVERSITY OF TECHNOLOGY
IN THE EUROPEAN CAPITAL OF CULTURE
CHEMNITZ

University of Nottingham
UK | CHINA | MALAYSIA

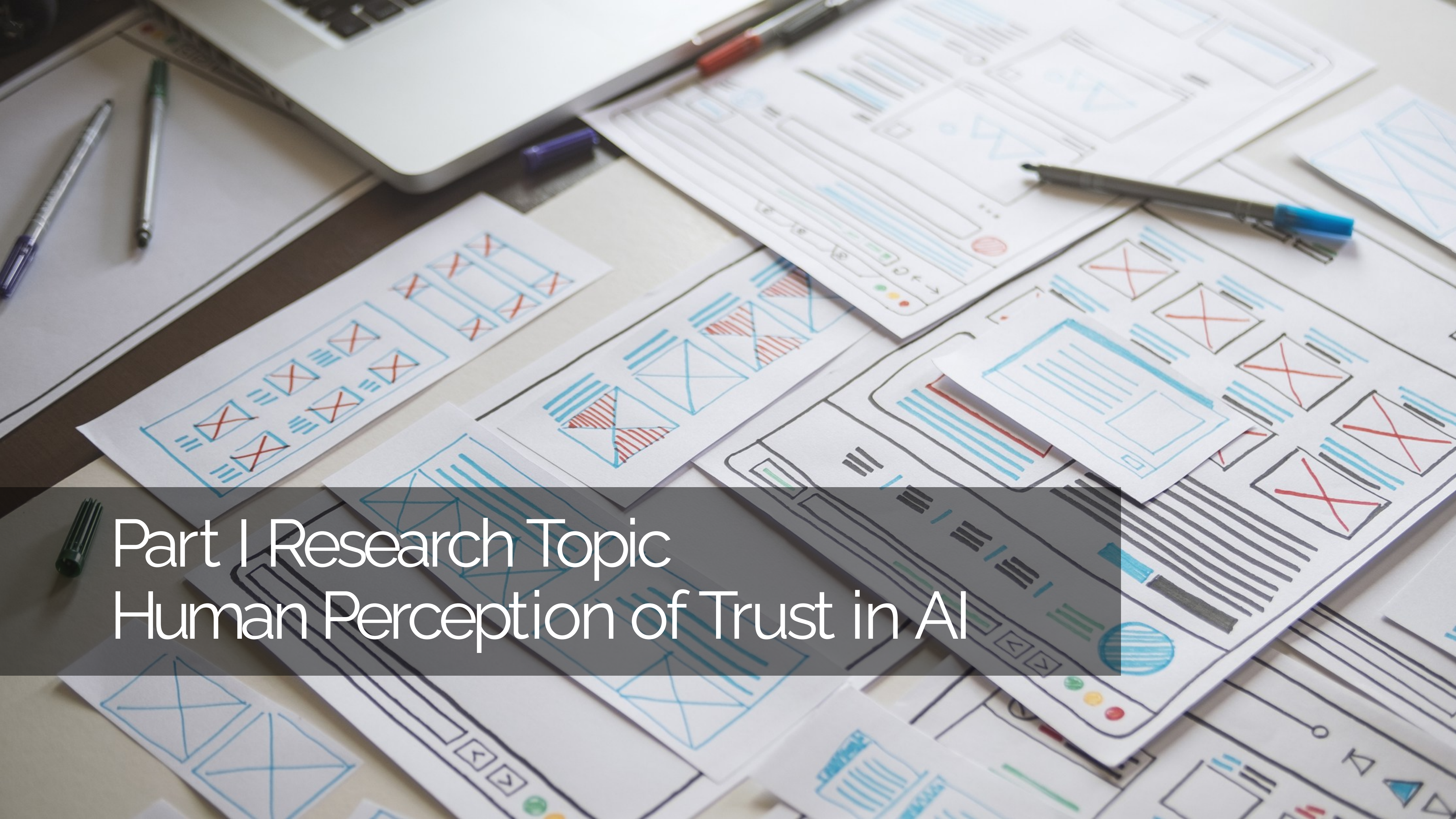
2025-09-09

Outline

RESEARCH TOPIC

THEORETICAL FOUNDATIONS

CURRENT RESEARCH



Part I Research Topic
Human Perception of Trust in AI

RT3 User Perception and Expectation

Scientific Challenge

Understanding how users *perceive AI* from the perspective of *trust and confidence* in the technology, which in turn includes multiple possible *perspectives* (including accuracy, reliability, safety, security).

Expected result

Factors that motivate or impact *user trust in AI* technologies and their relative *weightings in different contexts*.

Involved Partners

UoN, TUC, Telenor

Part II
Theoretical Foundations

About Trust and Trustworthiness

Trust is earned in drops and lost in buckets.

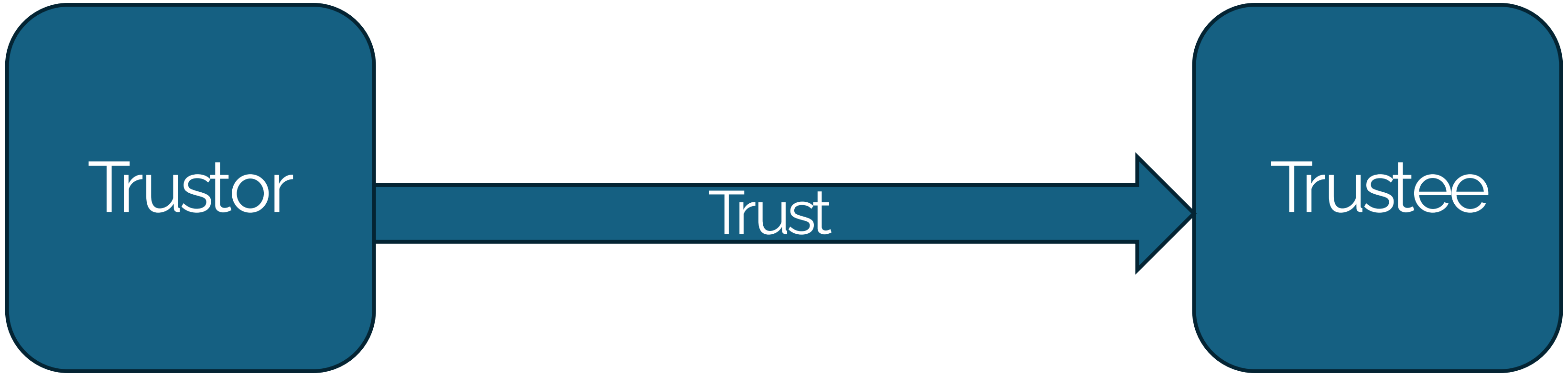
Kevin Kelly, founding executive editor of Wired magazine

Trust

Trust is viewed as:
(1) a set of specific beliefs dealing with benevolence, competence, integrity, and predictability (trusting beliefs);
(2) the willingness of one party to depend on another in a risky situation (trusting intention); or
(3) the combination of these elements.

Siau, K., & Wang, W. (2018). Building trust in artificial intelligence, machine learning, and robotics. Cutter Business Technology Journal, 31(2), 47–53.

Trust



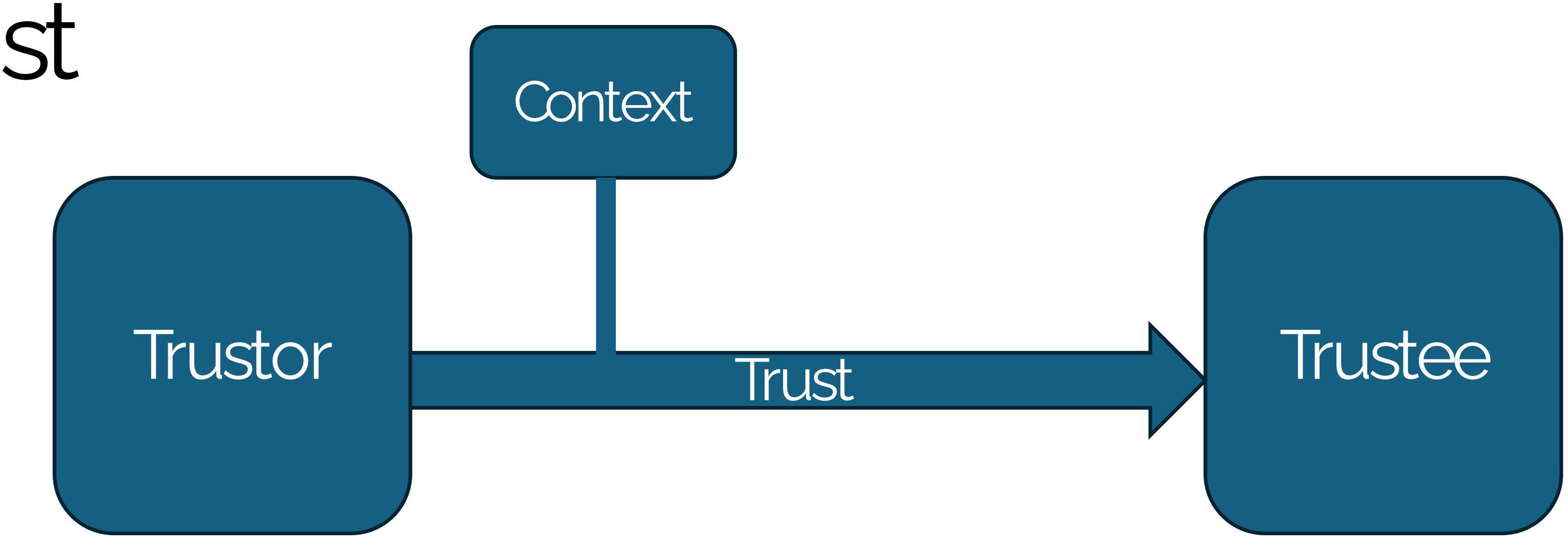
Trustor

Trust

Trustee

Mayer, R. C. ., Davis, J. H. ., & Schoorman, F. . D. (1995). An Integrative Model of Organizational Trust. Academy of Management Review, 20(3), 709–734.
Schoorman, F. D., Mayer, R. C., & Davis, J. H. (2007). An Integrative Model of Organizational Trust: Past, Present, and Future. Academy of Management Review, 32(2), 344–354.

Trust



Context

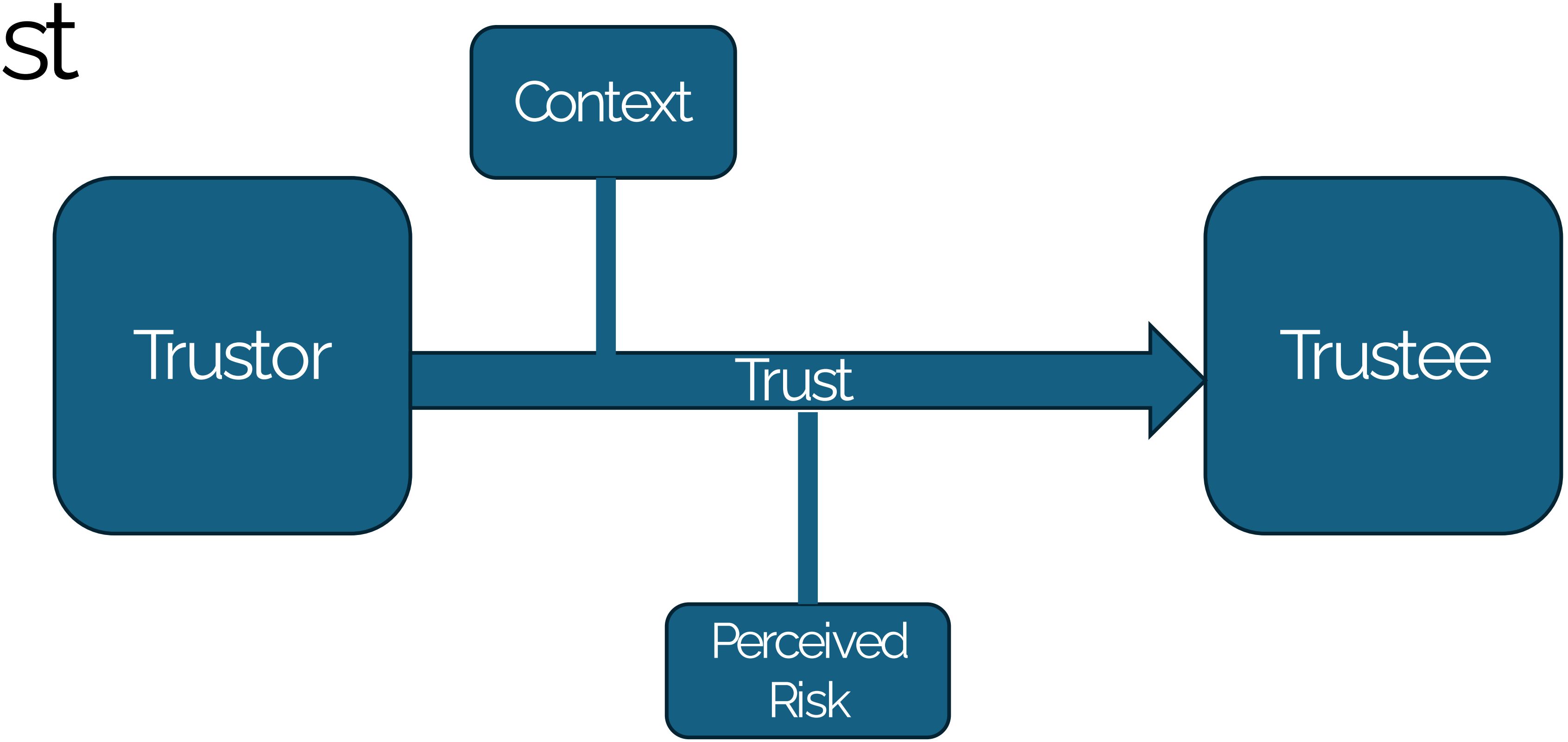
Trustor

Trust

Trustee

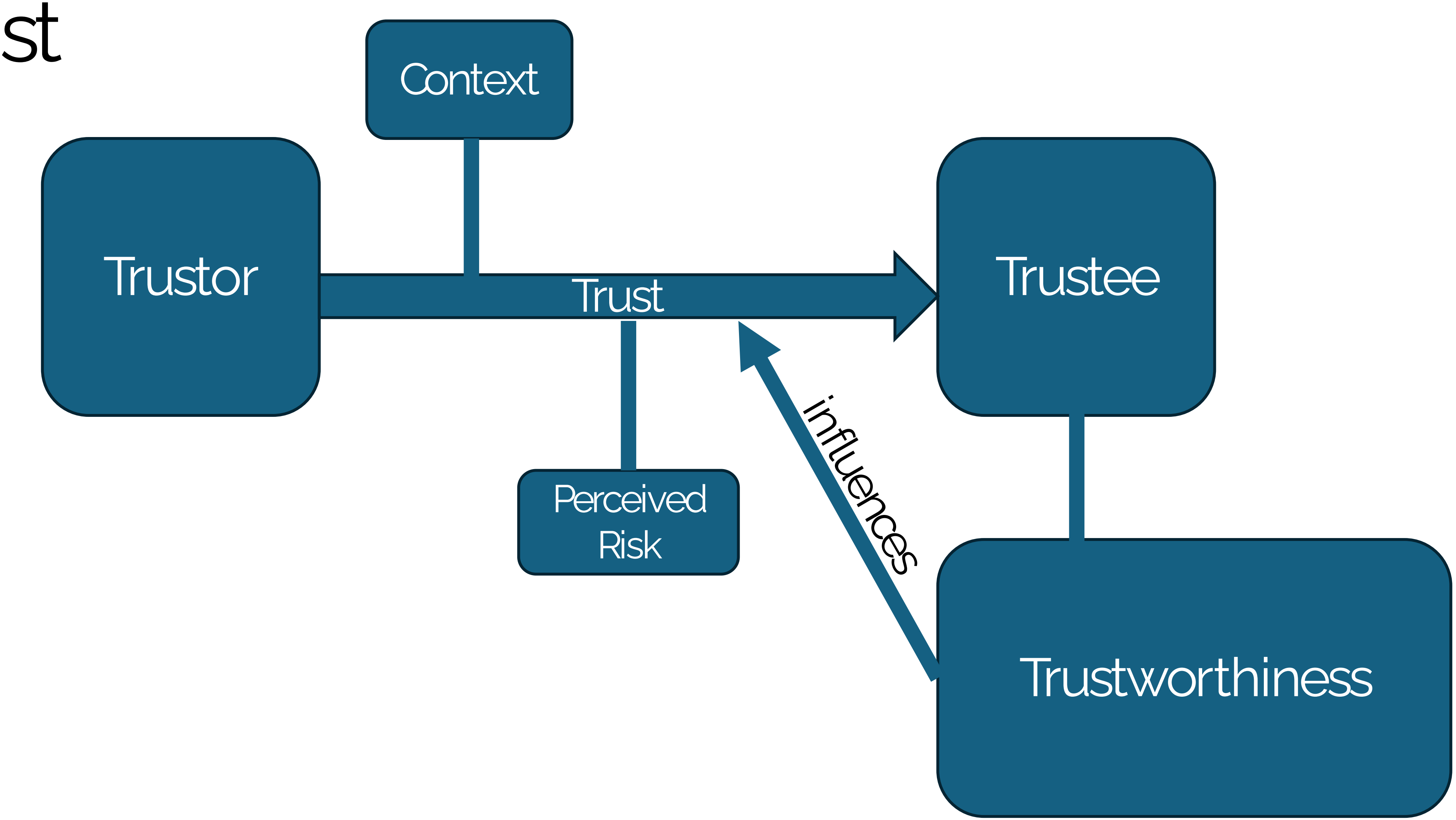
Mayer, R. C. ., Davis, J. H. ., & Schoorman, F. . D. (1995). An Integrative Model of Organizational Trust. Academy of Management Review, 20(3), 709–734.
Schoorman, F. D., Mayer, R. C., & Davis, J. H. (2007). An Integrative Model of Organizational Trust: Past, Present, and Future. Academy of Management Review, 32(2), 344–354.

Trust



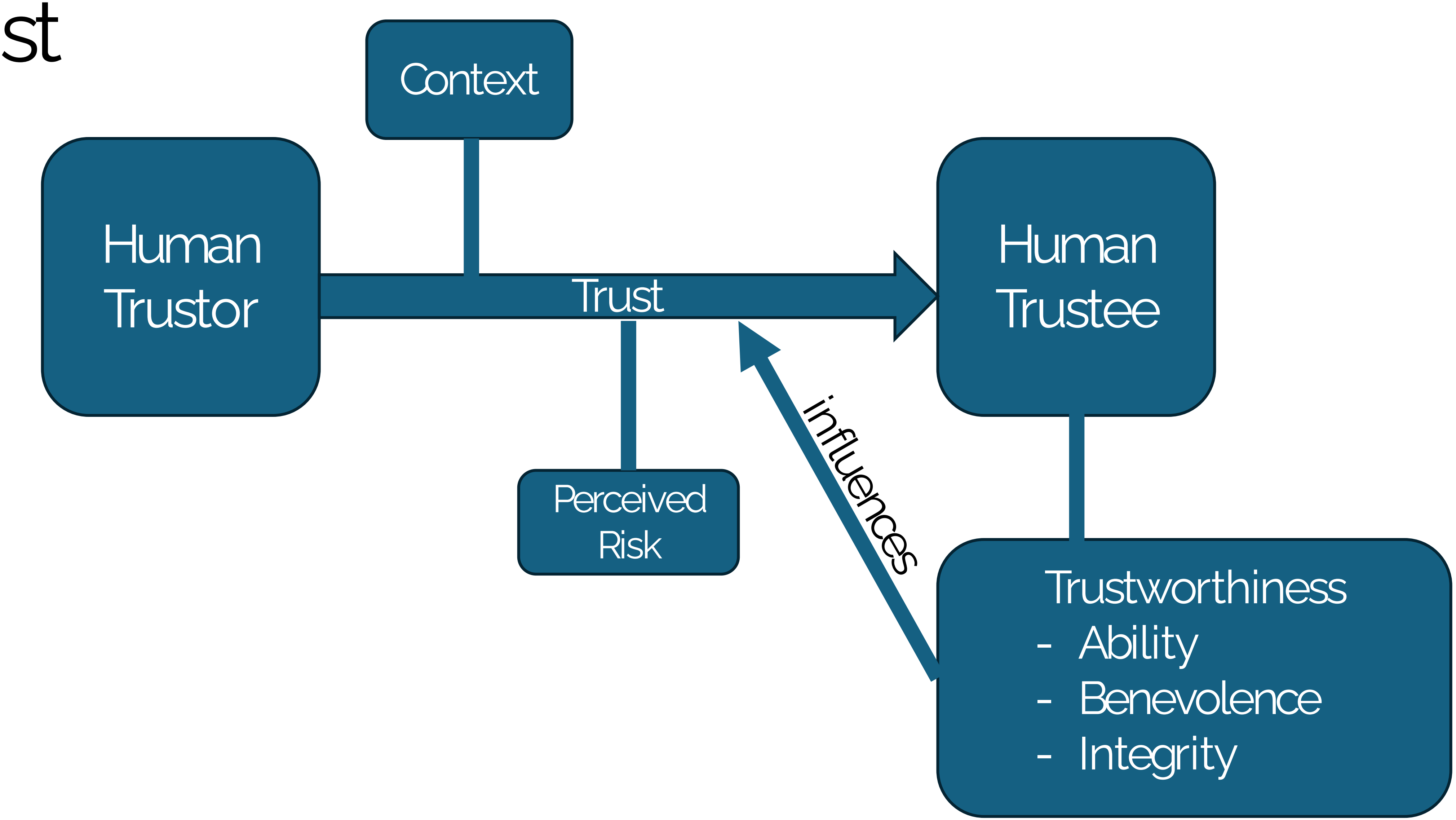
Mayer, R. C. ., Davis, J. H. ., & Schoorman, F. . D. (1995). An Integrative Model of Organizational Trust. Academy of Management Review, 20(3), 709–734.
Schoorman, F. D., Mayer, R. C., & Davis, J. H. (2007). An Integrative Model of Organizational Trust: Past, Present, and Future. Academy of Management Review, 32(2), 344–354.

Trust



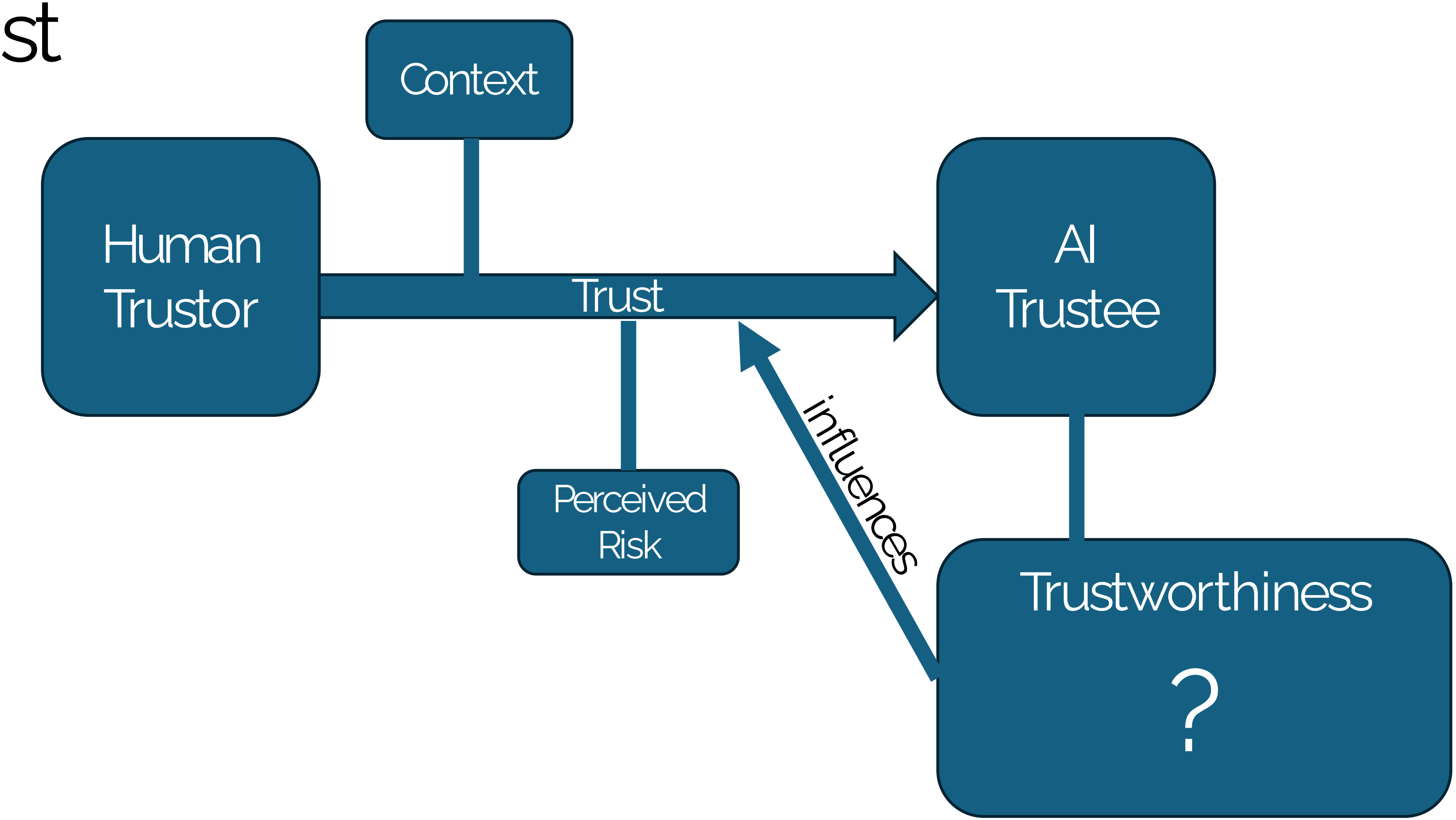
Mayer, R. C. ., Davis, J. H. ., & Schoorman, F. . D. (1995). An Integrative Model of Organizational Trust. Academy of Management Review, 20(3), 709–734.
Schoorman, F. D., Mayer, R. C., & Davis, J. H. (2007). An Integrative Model of Organizational Trust: Past, Present, and Future. Academy of Management Review, 32(2), 344–354.

Trust



Mayer, R. C. ., Davis, J. H. ., & Schoorman, F. . D. (1995). An Integrative Model of Organizational Trust. Academy of Management Review, 20(3), 709–734.
Schoorman, F. D., Mayer, R. C., & Davis, J. H. (2007). An Integrative Model of Organizational Trust: Past, Present, and Future. Academy of Management Review, 32(2), 344–354.

Trust



Mayer, R. C. ., Davis, J. H. ., & Schoorman, F. . D. (1995). An Integrative Model of Organizational Trust. Academy of Management Review, 20(3), 709–734.
Schoorman, F. D., Mayer, R. C., & Davis, J. H. (2007). An Integrative Model of Organizational Trust: Past, Present, and Future. Academy of Management Review, 32(2), 344–354.

Trustworthy AI

AI HLEG Report

European Perspective on TAI

Key Concepts

Guidelines

European Commission

INDEPENDENT

HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE

SET UP BY THE EUROPEAN COMMISSION

AI

ETHICS GUIDELINES FOR TRUSTWORTHY AI

European Commission (2019). Ethics Guidelines for Trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Key Requirements of Trustworthy AI

1. Human Agency & Oversight
2. Technical Robustness & Safety
3. Privacy & Data Governance
4. Transparency
5. Diversity, Non-Discrimination & Fairness
6. Environmental & Societal Well-Being
7. Accountability

European Commission (2019). Ethics Guidelines for Trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

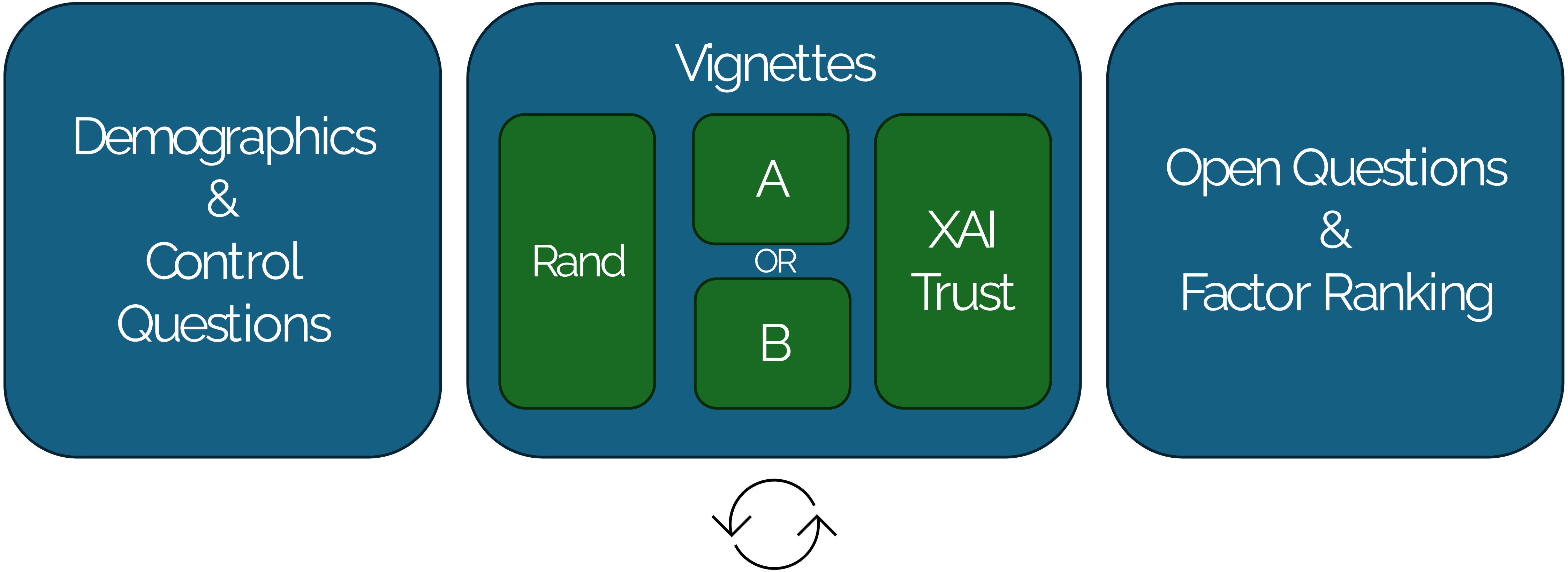
Part III
Current Research

Perceived Trust in ENFIELD Domains

RQ1	How is trust in the application of AI in the ENFIELD domains perceived by non-experts?
RQ2	Which factors influence the perceived trust?
Method	Vignette-based Survey

Perceived Trust in ENFIELD Domains

Survey Outline



Demographics
&
Control
Questions

Vignettes

Rand

A

OR

B

XAI
Trust

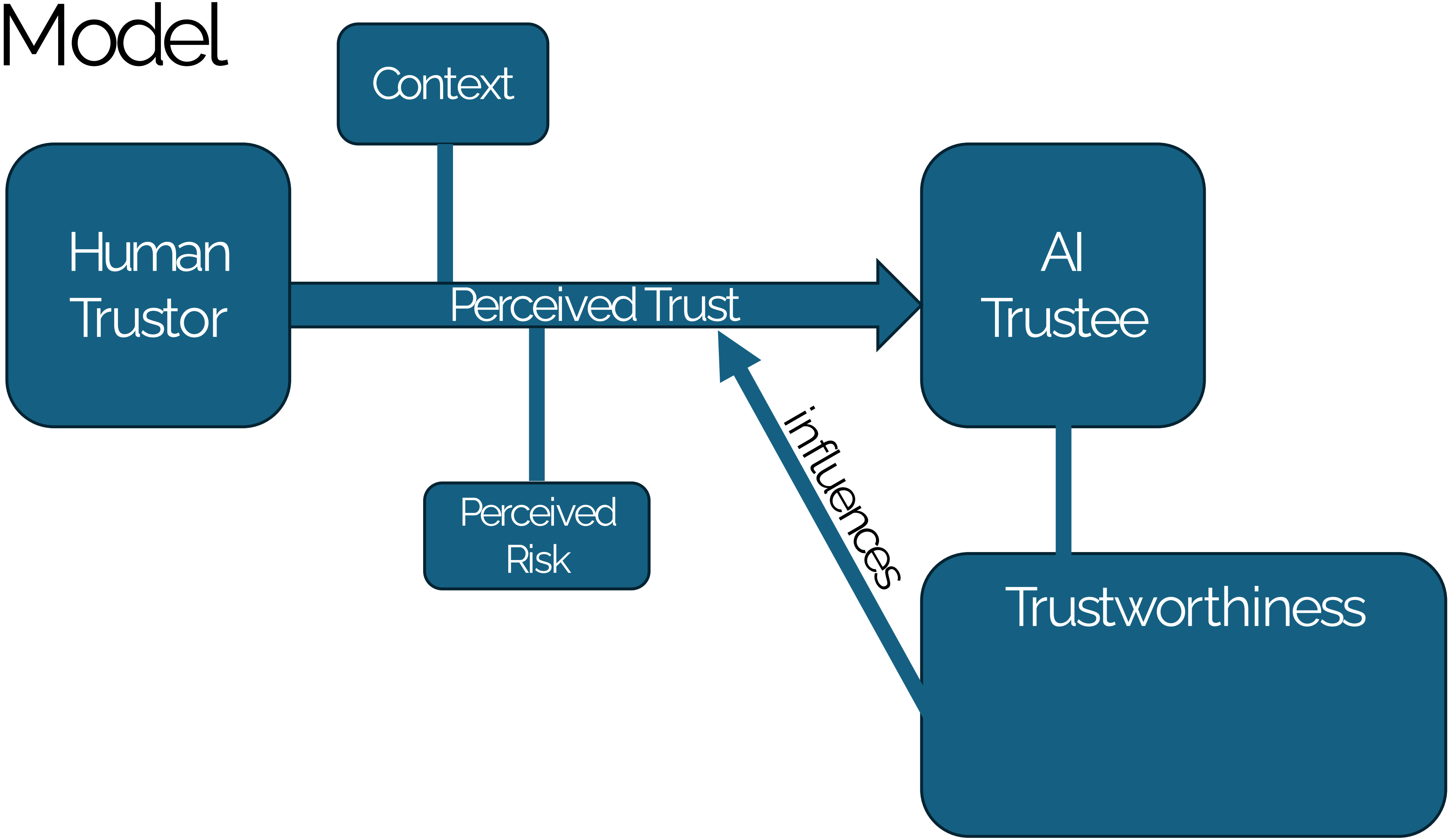
Open Questions
&
Factor Ranking

Vignette Sample

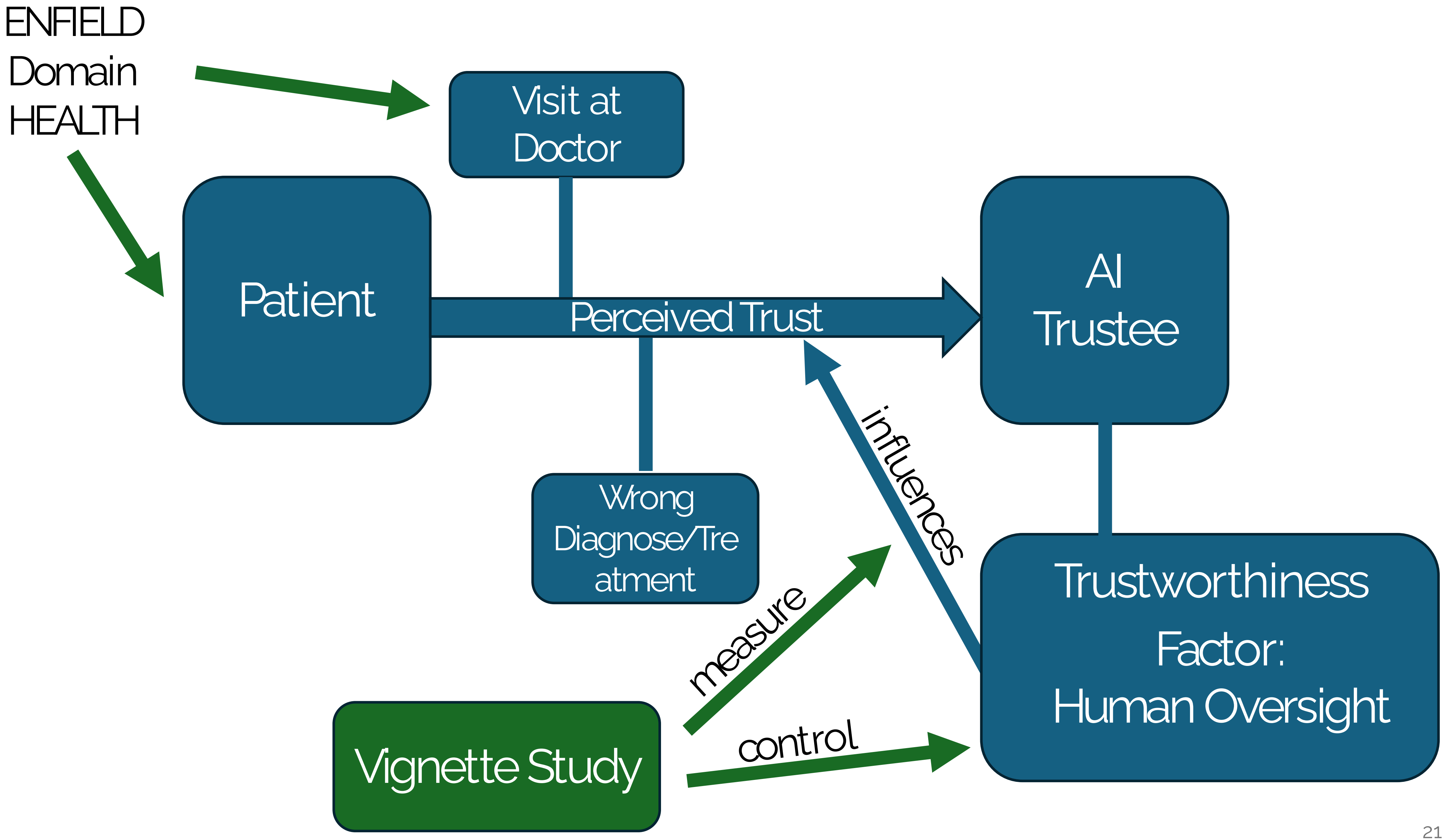
Domain: HEALTH Factor: Human Oversight

You are consulting a doctor after experiencing symptoms that concern you. Before your appointment, you learn that the doctor uses an AI system for diagnosis and creating treatment plans. On a regular basis, **[5/10]** % of decisions made by the AI are selected for review by experienced doctors. You would be informed in the event that your diagnosis and treatment have been found to require adaptation. During the consultation, the doctor explains that the AI has analyzed your medical records, lab results, and symptoms. Based on this analysis, the doctor informs you that you have been diagnosed with diabetes.

Trust Model



Mayer, R. C. ., Davis, J. H. ., & Schoorman, F. . D. (1995). An Integrative Model of Organizational Trust. Academy of Management Review, 20(3), 709–734.
Schoorman, F. D., Mayer, R. C., & Davis, J. H. (2007). An Integrative Model of Organizational Trust: Past, Present, and Future. Academy of Management Review, 32(2), 344–354.



ENFIELD
Domain
HEALTH

Visit at Doctor

Patient

Perceived Trust

AI Trustee

Wrong Diagnose/Tre atment

influences

measure

Trustworthiness Factor: Human Oversight

Vignette Study

control

Vignette Design Considerations

Domain: SPACE Factor: **Technical robustness and safety**

You are about to embark on a long flight and learn that the airline uses an AI-based system to plan the optimal route, specifically designed to avoid areas of turbulence. Weather conditions are monitored in real time to ensure a smooth flight, adjusting the path as necessary to avoid any disruptions. The system has been rigorously tested and **[includes a failsafe system that can take over in case of any technical problems with the AI/can adjust the route as needed to maintain the best possible flight experience].**

Vignette Design Considerations

Domain: SPACE Factor: Technical robustness and safety

You are about to embark on a long flight and learn that the airline uses an AI-based system to plan the optimal route, specifically designed to avoid areas of turbulence. Weather conditions are monitored in real time to ensure a smooth flight, adjusting the path as necessary to avoid any disruptions. The system has been rigorously tested and [includes a failsafe system that can take over in case of any technical problems with the AI/can adjust the route as needed to maintain the best possible flight experience].

Fallback mechanism needs to be technical, not human (confounder human oversight)
Fallback system needs to visibly not improve accuracy (confounder accuracy)
Keep same length and level of detail (confounder: transparency)

Vignette Design Considerations

Domain: SPACE Factor: Technical robustness and safety

You are about to embark on a long flight and learn that the airline uses an AI-based system to plan the optimal route, specifically designed to avoid areas of turbulence. Weather conditions are monitored in real time to ensure a smooth flight, adjusting the path as necessary to avoid any disruptions. The system has been rigorously tested and [includes a failsafe system that can take over in case of any technical problems with the AI/can adjust the route as needed to maintain the best possible flight experience].

Fallback mechanism needs to be technical, not human (confounder human oversight)
Fallback system needs to visibly not improve accuracy (confounder accuracy)
Keep same length and level of detail (confounder: transparency)

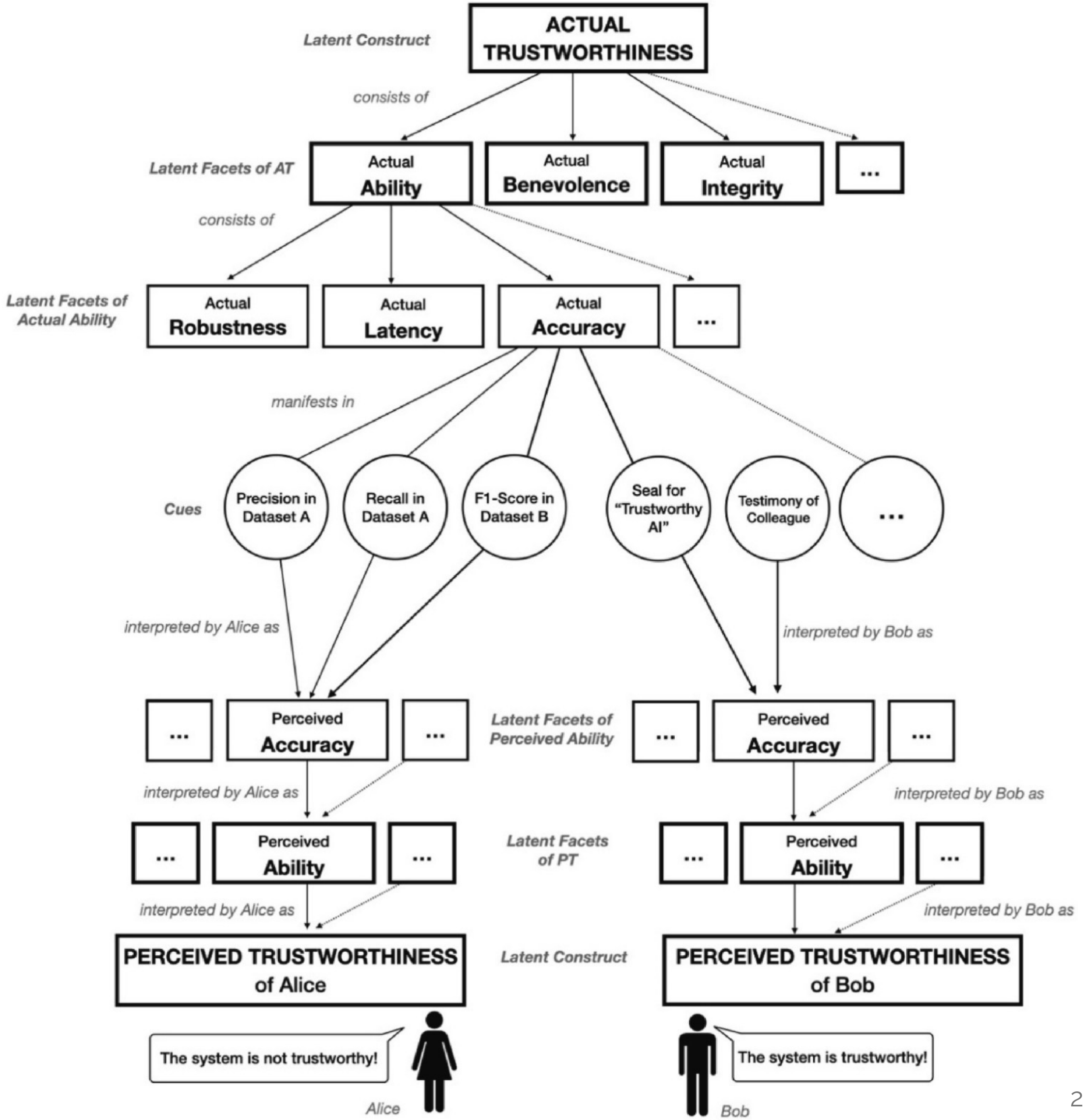
Vignette Design Considerations

Domain: SPACE Factor: Technical robustness and safety

You are about to embark on a long flight and learn that the airline uses an AI-based system to plan the optimal route, specifically designed to avoid areas of turbulence. Weather conditions are monitored in real time to ensure a smooth flight, adjusting the path as necessary to avoid any disruptions. The system has been rigorously tested and [includes a failsafe system that can take over in case of any technical problems with the AI/can adjust the route as needed to maintain the best possible flight experience].

Fallback mechanism needs to be technical, not human (confounder human oversight)
Fallback system needs to visibly not improve accuracy (confounder accuracy)
Keep same length and level of detail (confounder: transparency)

What are we measuring?



Schlicker, N.et al. (2025). How do we assess the trustworthiness of AI? Introducing the trustworthiness assessment model (TrAM). Computers in Human Behavior, 170, 108671.

XAI Trust Scale

8 items, 5-level Likert-scaled agreement

1. I am confident in the [tool]. I feel that it works well.
2. The outputs of the [tool] are very predictable.
3. The tool is very reliable. I can count on it to be correct all the time.
4. I feel safe that when I rely on the [tool] I will get the right answers.
5. The [tool] is efficient in that it works very quickly.
6. I am wary of the [tool].
7. The [tool] can perform the task better than a novice human user.

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2023). Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. Frontiers in Computer Science, 5.

Control Scale: Meta AI Literacy Scale

MAILS Short, 10 items, 0-10 rating scale

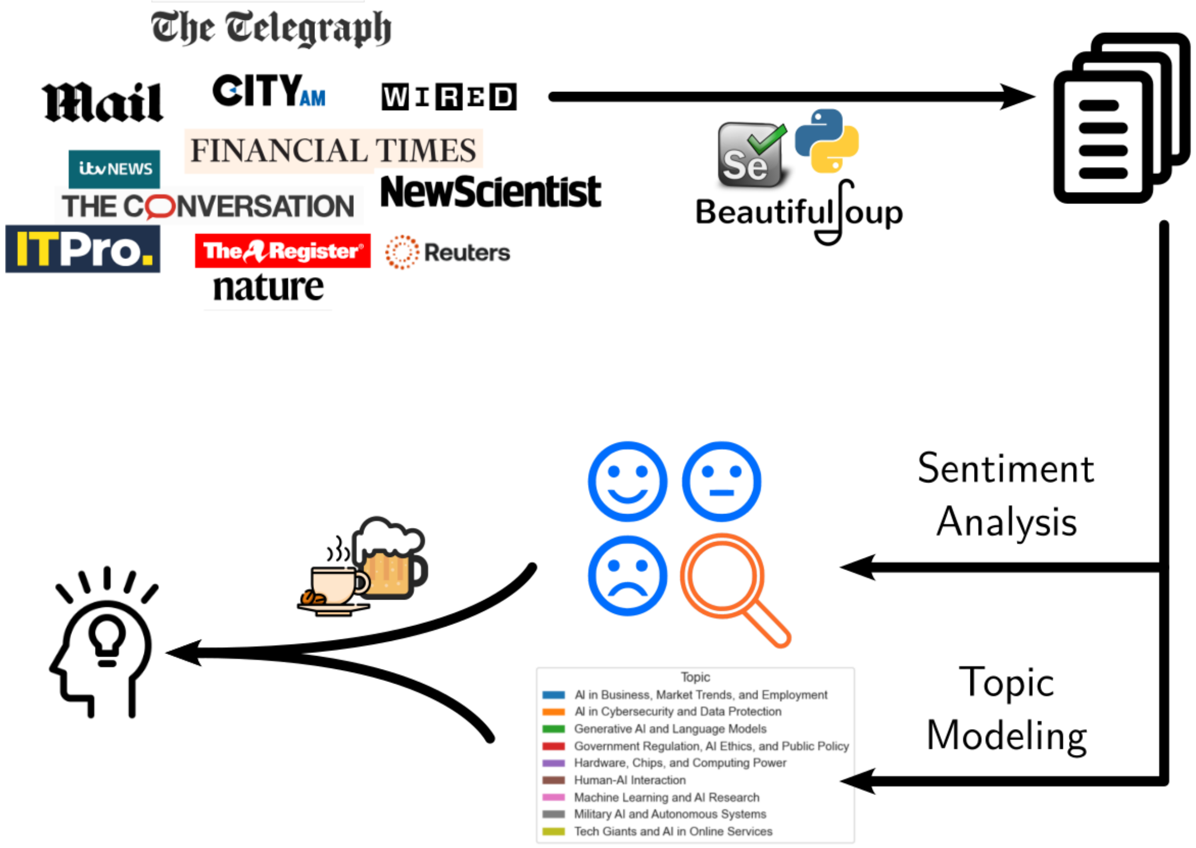
1. I can tell if I am dealing with an application based on artificial intelligence.
2. I can weigh the consequences of using AI for society.
3. I can use artificial intelligence meaningfully to achieve my goals.
4. I can assess what advantages and disadvantages the use of an artificial intelligence entails.
5. I can program new applications in the field of ”artificial intelligence
6. I can design new AI applications.
7. Although there are often new AI applications, I manage to always be ”up-to date”
8. I can also usually solve strenuous and complicated tasks when working with artificial intelligence well
9. I can handle it when interactions with AI frustrate or frighten me
10. I can prevent an AI from influencing me in my decisions

Carolus, A. et al. (2023). MAILS - Meta AI literacy scale: Development and testing of an AI literacy questionnaire based on well-founded competency models and psychological change- and meta-competencies. Computers in Human Behavior: Artificial Humans, Volume 1, Issue 2.

Public Perceptions of Trustworthy AI: Insights from a Longitudinal Study of UK News Media

Large-scale, longitudinal mapping of UK AI news discourse

- 7,691 articles from 2013 to 2024, from 12 UK news outlets
- Mainstream, business, scientific and technology themes
- Diverse range of viewpoints and readerships, aiming to capture a broad spectrum
- Descriptive statistics, sentiment analysis, and topic modeling



Topic

- 🟦 AI in Business, Market Trends, and Employment
- 🟧 AI in Cybersecurity and Data Protection
- 🟩 Generative AI and Language Models
- 🟥 Government Regulation, AI Ethics, and Public Policy
- 🟪 Hardware, Chips, and Computing Power
- 🟫 Human-AI Interaction
- ▨ Machine Learning and AI Research
- ▨ Military AI and Autonomous Systems
- ▨ Tech Giants and AI in Online Services

Murphy, T., Furnell, S.,, Heil , S. and Carpent, X. (2025). Public Perceptions of Trustworthy AI: Insights from a Longitudinal Study of UK News Media..
19th IFIP International Symposium on Human Aspects of Information Security & Assurance.

Time to spare at the end of the day?

CYBER DEFENCE DICE

Playtest for a cyber security awareness dice game

~40 minutes

6-10 players needed

No cyber expertise needed

Further Resources

- AI HLEG Report “Ethics Guidelines for Trustworthy AI”
- Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., & Zhou, B. (2023). Trustworthy AI: From Principles to Practices. ACM Computing Surveys, 55(9), 1–46. https://doi.org/10.1145/3555803
- Vereschak, O., Bailly, G., & Caramiaux, B. (2021). 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. https://doi.org/10.1145/3476068
- TAILOR Handbook of Trustworthy AI http://tailor.isti.cnr.it/handbookTAI/TAILOR.html

ENFIELD

UNIVERSITY OF TECHNOLOGY
IN THE EUROPEAN CAPITAL OF CULTURE
CHEMNITZ

Thank You

University of
Nottingham
UK | CHINA | MALAYSIA

Sebastian.Heil
@informatik.tu-chemnitz.de

Xavier.Carpent
@nottingham.ac.uk

Steven.Furnell
@nottingham.ac.uk

RG Sebastian-Heil

in sebastianheil

RG Xavier-Carpent-69827536

in xavier-carpent-2817397

RG Steven-Furnell

in stevenfurnell

Vignette 2

Domain: HEALTH Factor: Accuracy

You are consulting a doctor after experiencing symptoms that concern you. Before your appointment, you learn that the doctor uses an AI system to assist in diagnosing and deciding treatment plans. The AI is highly reliable but is known to falsely diagnose diabetes in **[2/5]** % of cases. During the consultation, the doctor explains that the AI has analyzed your medical records, lab results, and symptoms. Based on this analysis, the doctor informs you that you have been diagnosed with diabetes. The system is used by the doctor to provide faster and more accurate care.

Vignette 3

Domain: ENERGY Factor: Privacy

Your energy provider is using an AI-based energy bill prediction service. The system analyses your overall electricity usage across a range of activities to provide an estimate of your monthly bill. The prediction takes into account your past energy consumption patterns, as well as seasonal variations, and has been determined to be suitably accurate. The system collects **[overall energy consumption in your home/specific data on the types of devices you use and for how long]** as a basis for billing. You receive your latest bill with the prediction and take a look at it.

Vignette 4

Domain: SPACE Factor: Technical robustness and safety

You are about to embark on a long flight and learn that the airline uses an AI-based system to plan the optimal route, specifically designed to avoid areas of turbulence. Weather conditions are monitored in real time to ensure a smooth flight, adjusting the path as necessary to avoid any disruptions. The system has been rigorously tested and **[includes a failsafe system that can take over in case of any technical problems with the AI/can adjust the route as needed to maintain the best possible flight experience].**

Vignette 5

Domain: ENERGY Factor: Technical robustness and safety

You learn that the power plant you rely on is equipped with an AI-supported control system designed to optimise its operations and efficiency. The system is able to adjust the plant's processes in real time, ensuring it runs smoothly under varying conditions. While there have been recent cyberattacks on similar AI systems in other plants, this one is operated with ongoing human oversight to ensure its proper functioning. The system is certified to withstand **[conventional cybersecurity attacks/cybersecurity attacks specifically targeted at AI].**

Trustworthy AI

Trustworthy AI has three components:

(1) it should be lawful, ensuring compliance with all applicable laws and regulations

(2) it should be ethical, demonstrating respect for, and ensure adherence to, ethical principles and values and

(3) it should be robust, both from a technical and social perspective, since, even with good intentions, AI systems can cause unintentional harm.

Applies to AI system, and all processes & actors of AI system's life cycle

European Commission (2019). Ethics Guidelines for Trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Lawfull – EU AI Act

Unacceptable Risk

- Manipulation of human bevavior
- Real-time remote biometric identification in public spaces
- Social Scoring

High Risk

- Health
- Education
- Recruitment
- Critical Infrastructure
- Law Enforcement
- Justice

Assessment List for Trustworthy Artific Intelligence (ALTAI)

Measure
if your organisation's AI is
trustworthy
ALTAI – Assessment List for Trustworthy Artifical Intelligence

For internal use/self-assessment

Concrete Questions for the 7 HLEG Requirements

Stakeholders

- AI designers and AI developers of the AI system
- data scientists
- procurement officers or specialists
- front-end staff that will use or work with the AI system
- legal/compliance officers
- management

AI HLEG (2020). ALTAI https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment