SAND2024-14045C

Sandia National Laboratories

Exceptional service in the national interest

Trust Us: A Simple Model for Understanding Trust in AI

PRESENTED BY

Kyra Wisniewski

Christina Ting, Laura Matzen

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. SAND20XX-XXXX P

Introduction

Motivation

- Overloading the term “trust” has led to dissonance in its formal study.

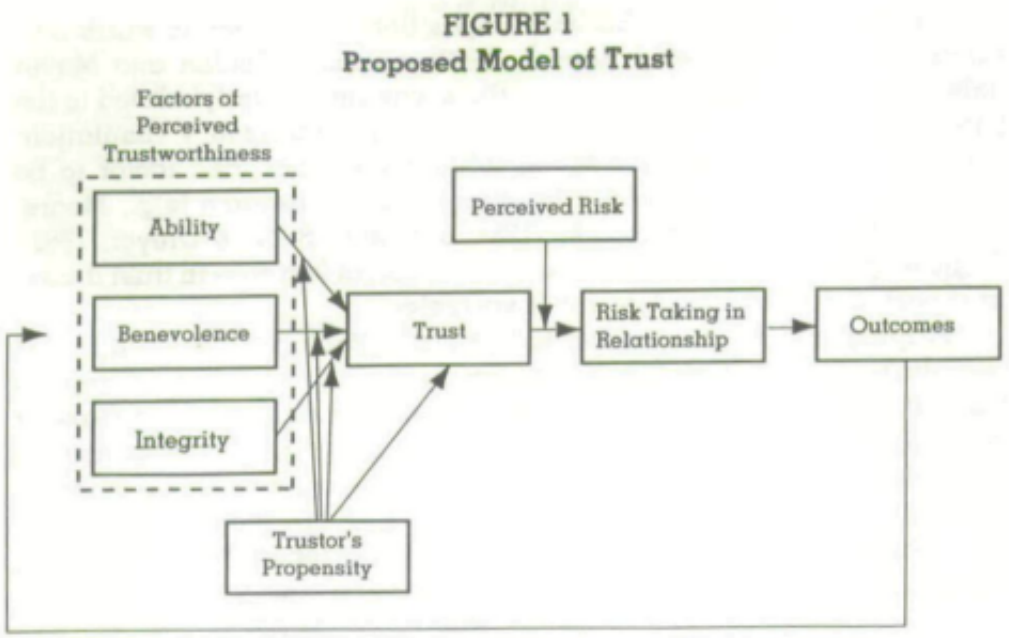
Goals

1. Connect the definitions of trust as an attitude and trust as an intention.
2. Present a trust model that clearly incorporates the two definitions of trust as separate concepts.
3. Define appropriate trust in artificial intelligence (AI) within the context of the model.

Two Definitions of Trust

"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."

[1] Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of management review*, *20*(3), 709-734.



"The **attitude** that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability."

[2] Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, *46*(1), 50-80.

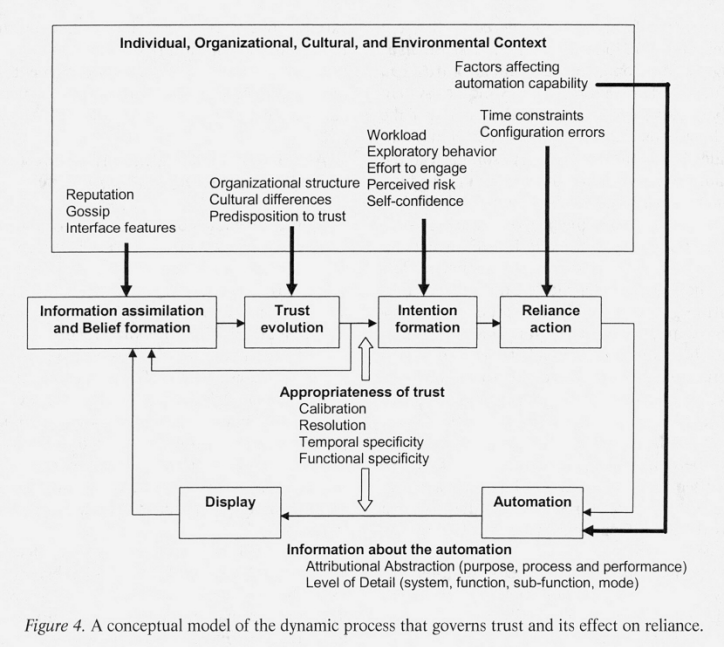
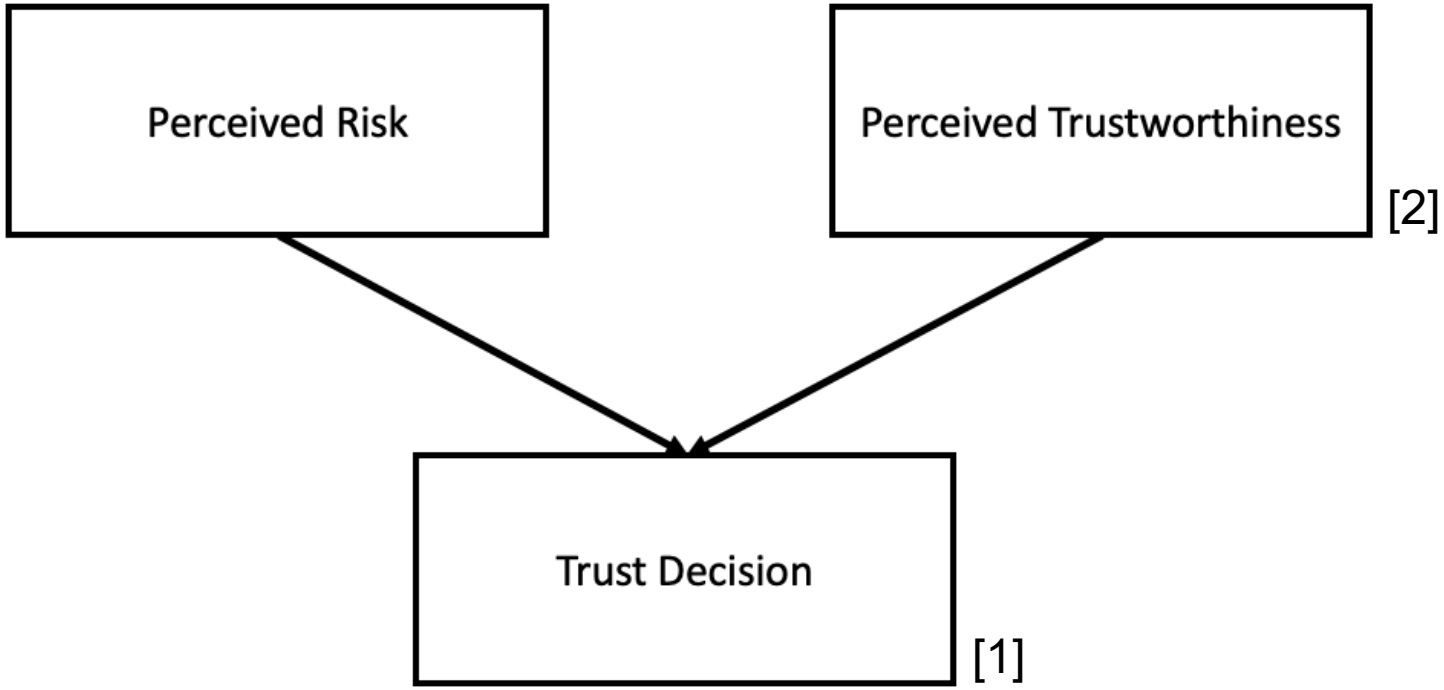


Figure 4. A conceptual model of the dynamic process that governs trust and its effect on reliance.

A Simple Trust Model

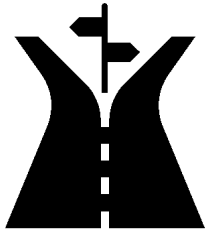
"Trust is an evaluation of attitudes about the potential for gains or losses involving the trustee against those not involving the trustee"



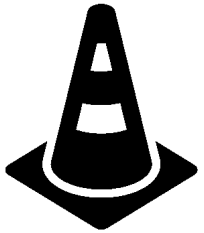
Hypothetical Trust Scenario

Imagine you are driving to work on your normal interstate route during standard rush hour traffic, and your goal is to get to work on time.

You have allowed sufficient time for your usual commute. However, your outdated navigation system suggests an alternative route using local roads.

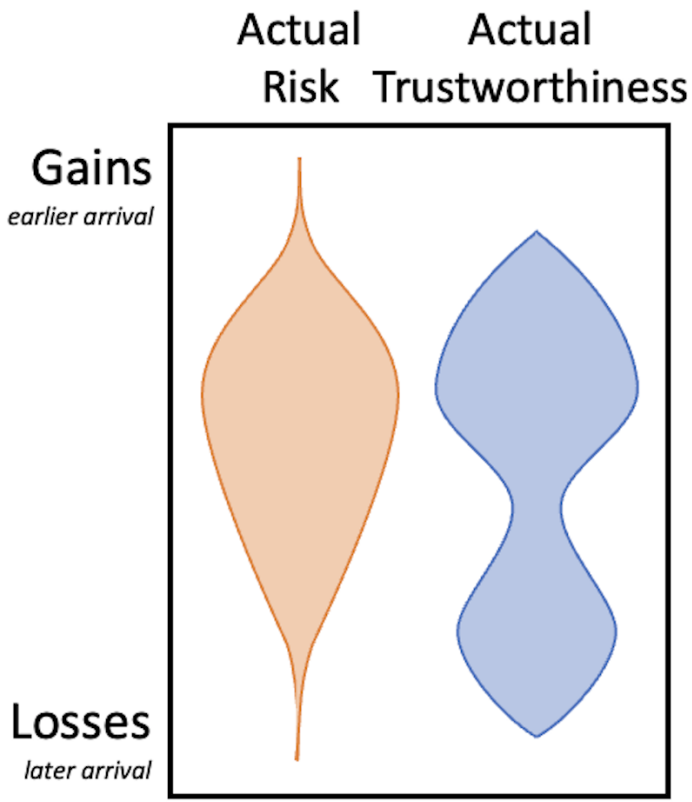


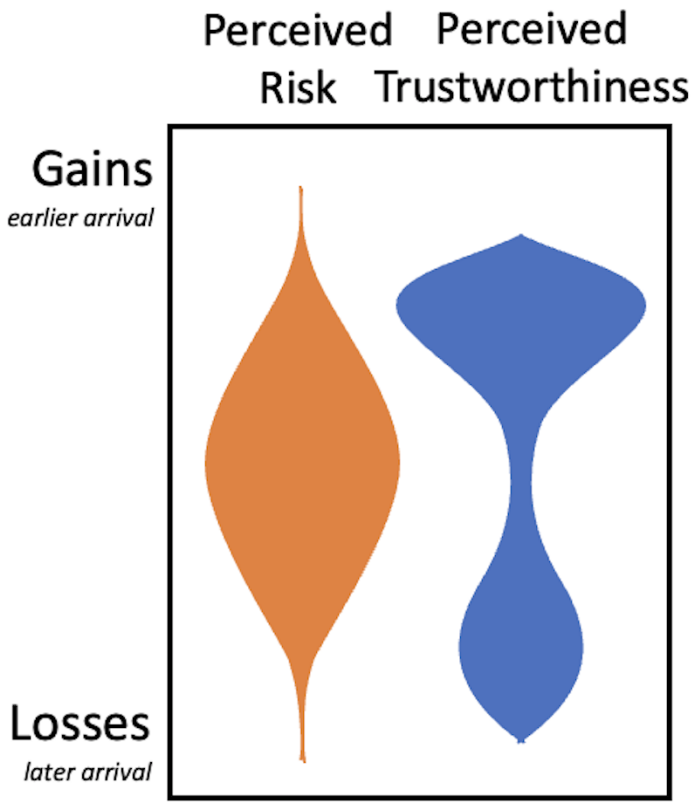
Its recommendation is based on the current traffic conditions, so by the time you reach that route it may be backed up and no longer optimal…



Risk and Trustworthiness

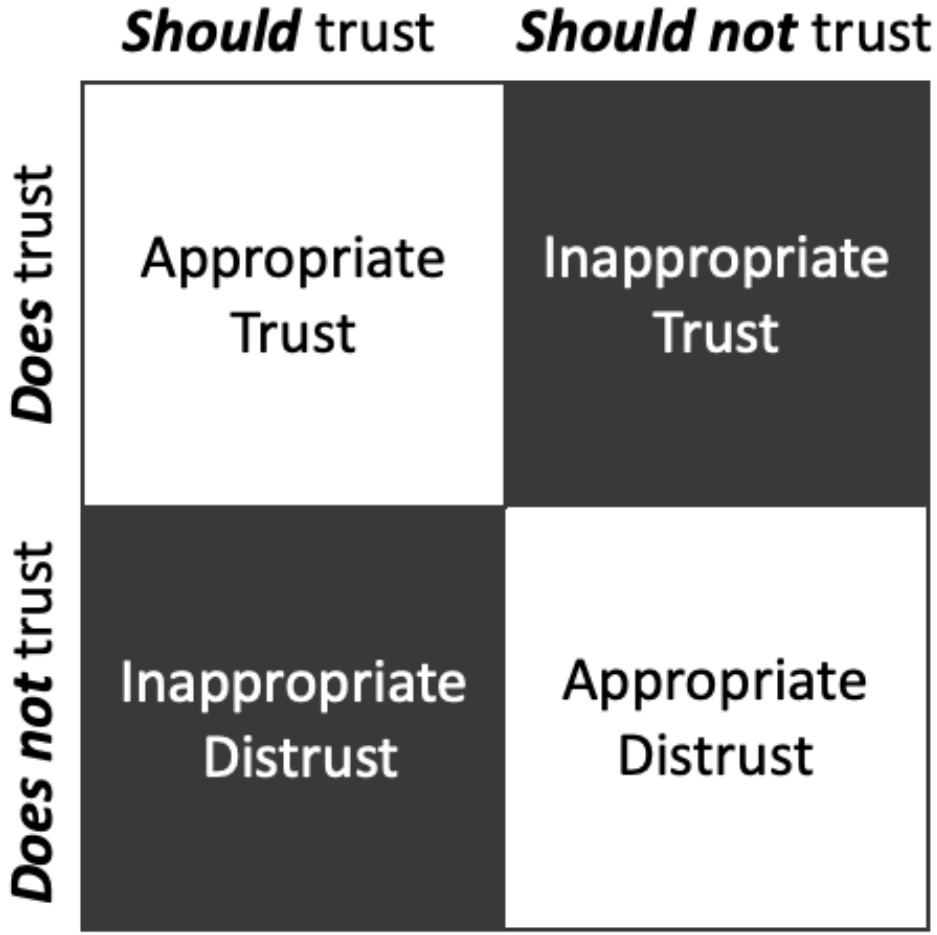
- Perceived risk and trustworthiness are an individual's attitudes about the potential for gains and losses in a situation with and without the AI's help in achieving a goal.
- Actual risk and trustworthiness can be mapped in terms of the true potential for gains and





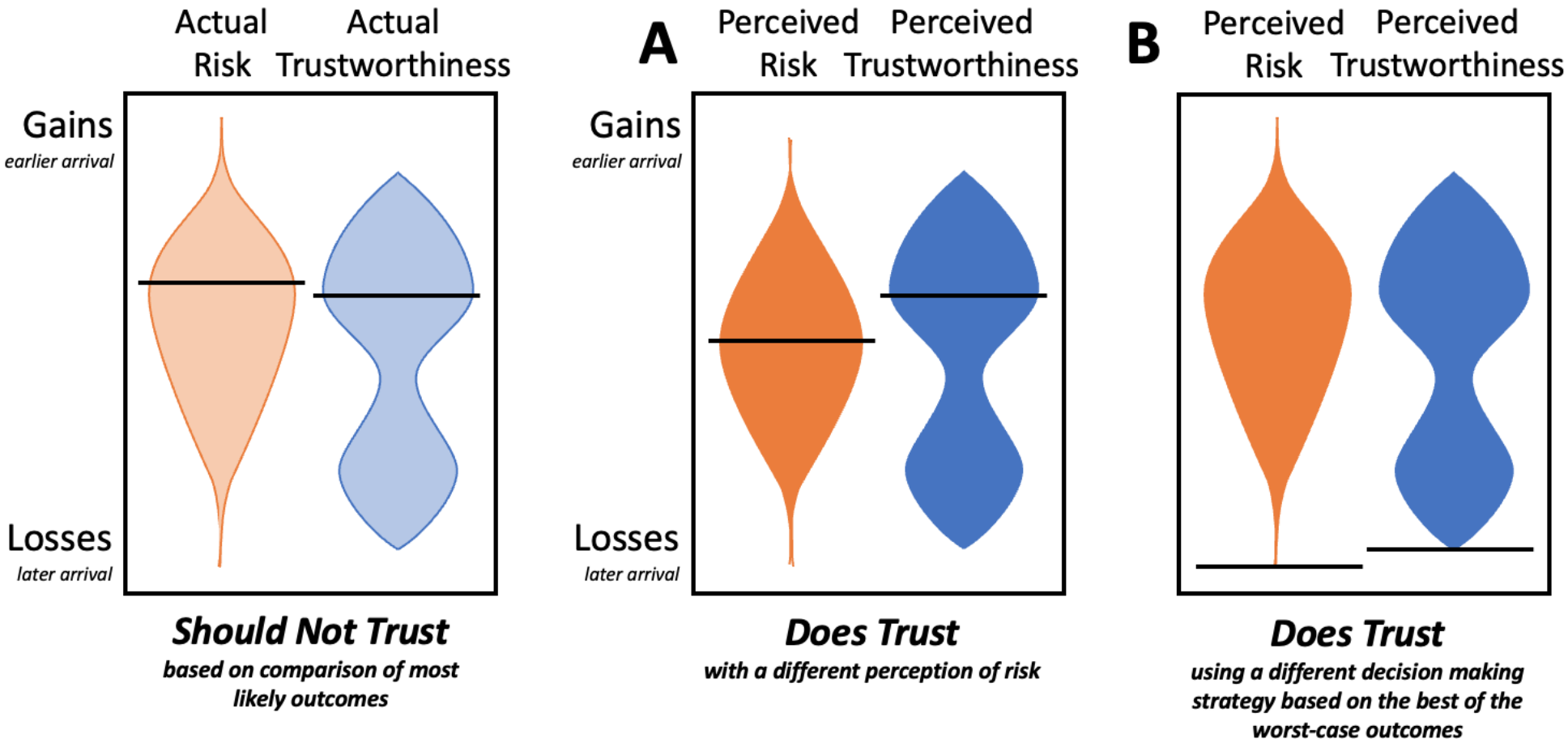
Appropriate Trust: Evaluating the Trust Decision

- Appropriateness is when the trust decision that should occur, does occur.



Alternate Definitions of Appropriateness

- Alternative definitions of appropriateness focus solely on perceived trustworthiness [2][3]



[3] De Visser, E. J., Peeters, M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020). Towards a theory of longitudinal trust calibration in human–robot teams *International journal of social robotics*, *12*(2), 459-478.

Understanding How the Trust Decision Was Made

- Understanding *how* a trust decision was made is a more challenging application of our model.
- Measuring an individual's perceptions of risk and trustworthiness and gathering information about how they are weighing these two factors in their trust decision is not straightforward.
  - Self-report data about perceptions may conflict with the trust decision.

Key Takeaways

- Trust in AI is simpler than it seems.
- Future work in this area should focus on exploring how a person perceived risk and trustworthiness to arrive at their use of an AI.
- Clearly defining actual risk and trustworthiness, perceived risk and trustworthiness, and trust as an intention to use the AI will clarify conflicting findings in the existing research and support better experimental designs and replicability in future research.