*Author(s): Naja Kathrine Kollerup, Mikael B. Skov, Niels van Berkel (2022): Using Signals to Support Trust Building in Clinical Human-AI Collaboration. In: Proceedings of the 20th European Conference on Computer-Supported Cooperative Work: The International Venue on Practice-centred Computing on the Design of Cooperation Technologies - Exploratory Papers, Reports of the European Society for Socially Embedded Technologies (ISSN XXX-XXXX), DOI: 10.18420/ecscw2022-to-be-added*

# Using Signals to Support Trust Building in Clinical Human-AI Collaboration

Author(s): Naja Kathrine Kollerup, Mikael B. Skov, Niels van Berkel
Aalborg University
*Contact Author: nkka@cs.aau.dk*

**Abstract.** Artificial Intelligence (AI) has the technological potential to transform healthcare by assisting medical personal in their everyday workflow. For successful collaboration and adoption of AI technology, end-users need to trust the AI system. In this paper we outline the use of Relational Signalling Theory, an established theory on Human-Human trust building, as a conceptual lens for designing trust-building signals in Human-AI interaction. We argue that the use of a theoretical foundation in the design and evaluation of interactions supports the development of Human-Centered AI in healthcare.

## Introduction

Artificial Intelligence (AI) systems take an increasingly larger role in assisting humans in clinical decision-making (Oh et al., 2018; Yang et al., 2020). For AI systems to be used successfully in clinical practice requires AI systems to collaborate and align with day-to-day clinical practice (Wang et al., 2020). Thus, to integrate AI systems as a collaborative partner in human workflow, especially in high-stake workflows encountered in healthcare, requires humans to put trust in the AI support system (van Berkel et al., 2022; Vereschak et al., 2021).

Current AI systems often face distrust, which can result in underestimation of the AIs capabilities, disuse, increased user workload, or deterioration of the systems performance (Okamura and Yamada, 2020). In this article we argue that in order to design AI systems that evoke trust among its end-users, we first need to understand how humans build trust. We first briefly summarise the Relational Signalling Theory (RST) (Six et al., 2010) and how it is used to describe trust-building behaviour in human-to-human relationships. Secondly, we outline how the concept and techniques of relational signals potentially can have profound implications for trust-building in Human-AI teams. Finally, we present examples within the healthcare domain of qualities AI systems should acquire to facilitate trust building.

Through our existing understanding of human-to-human trust-building, we outline concrete takeaways for Human-AI trust-building and motivate a novel research direction for the CSCW and HCI communities in relation to clinical Human-AI collaboration.

# Relational Signalling Theory

Relational signalling theory (RST) is a theory proposed by Lindenberg (2000). RST is grounded in two basic assumptions: First, human behaviour is goal-directed, and to explain the social context, one must pay attention to the goals of individuals. Second, human behaviour is context dependent (Six et al., 2010). Relational signals are signs in the behaviour of the trustee (*i.e.*, the party aiming to create trust), where the trustor (*i.e.*, the party assessing the trustee) considers two distinct aspects; Does the trustee show signs in the behaviour of interest in maintaining a relationship in the future (ability dimension of trustworthiness), and does the trustee show signs in the behaviour of having the competences to perform according to the expectations? (internal dimension of trustworthiness).

Lindenberg (2000) distinguishes between three master frames of operation: the hedonic frame, the gain frame, and the normative or solidarity frame. The first two frames are ego-oriented and serve one's own interest, whereas the third frame is alter-oriented, which means that one will show concern for the other individual (Six et al., 2010). People will look for signs in the behaviour of another individual indicating stability in the solidarity frame, and moreover, to which degree the individual is interested in maintaining a relationship in the future.

## Signalling types

A better understanding of the signal types emitted from AI systems (the trustee) supports the design of systems that are trusted by clinician and patient alike (the trustor) and helps to answer the question: 'What qualities should the AI system have in order for humans to trust it as a collaborative partner in clinical care?' To answer this question, we need to understand not only the frame of operation but consider the signals as an incorporated part of trust building. We draw upon the

Signalling theory (Donath, 2007) concerned with understanding why specific signals are reliable and others are not. Donath identifies two signals;

- *Assessment signals*, which are costly to fake and therefore considered honest and reliable signals. The quality they signal is 'wasted' in the production and therefore tend to be expensive to produce (Shami et al., 2009) and is challenging to fake. For example, lifting a heavy weight sends a reliable signal of strength – a weaker person simply cannot do it (Donath, 2007).
- *Conventional signals*, which are cheaper to produce and therefore considered less reliable and open to deception. To give an example, one may choose to use a deceptive picture of oneself on social media. This will result in loose of meaning, and these signals become unreliable (Shami et al., 2009).

Building on these established notions of trust-building in Human-Human interaction, we propose the use of RTS to frame Human-AI interactions. Through the aforementioned signal types from signalling theory, we can conceptualise, design, and study the signals that affect a trustor's belief or behaviour (Lampe et al., 2007). Especially within the healthcare domain, trust-building is essential for both clinicians and patients. We next outline how signals can be embedded in Human-AI collaboration scenarios in clinical care to enhance trust.

## Using signals in healthcare AI systems

The qualities the AI-system should acquire to facilitate trust-building for medical practitioners and patients depends on the signals being send from the AI system. These qualities can almost be anything, *e.g.*, honesty, reliability (Donath, 2007).

As an example of using signals in a healthcare context, we describe an AI-powered computer vision system designed to identify moles from melanomas. To gain the user's trust, the AI system can present a number of assessment signals. For example, the system can present alternative considerations made in its assessment and detail why these considerations did not end up being the final assessment. Alternatively, the system can highlight which specific elements of the image are deemed suspicious. These are costly signals (computational power, added explanations) that we hypothesise would increase end-user trust. Conventional signals, such as presenting the outcome of the analysis without any type of explanation, are less likely to support trust building in Human-AI interactions and come at a lower cost to the AI system.

## Conclusion and future work

In this position paper, we have provided an overview of relational signalling theory and argued for its use in trust-building in clinical Human-AI collaboration. Given the importance of trust in the healthcare domain, it is critical that we design systems that instil trust in users where appropriate, but are also able to highlight to its users when its recommendations are less trustworthy and should receive extra scrutiny. We call for future work to empirically study the degree to which assessment signals and conventional signals support trust-building in end-users of Human-AI systems.

# Acknowledgements

This work is supported by the EXPLAIN-ME project of Digital Research Centre Denmark (DIREC) under Innovation Fund Denmark.

# References

Donath, J. (2007): 'Signals in Social Supernets'. *J. Computer-Mediated Communication*, vol. 13, pp. 231–251.

Lampe, C. A., N. Ellison, and C. Steinfield (2007): 'A Familiar Face(Book): Profile Elements as Signals in an Online Social Network'. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. p. 435–444.

Lindenberg, S. (2000): 'It Takes Both Trust and Lack of Mistrust: The Workings of Cooperation and Relational Signaling in Contractual Relationships'. *Journal of Management and Governance*, vol. 4, pp. 11–33.

Oh, C., J. Song, J. Choi, S. Kim, S. Lee, and B. Suh (2018): *I Lead, You Help but Only with Enough Details: Understanding User Experience of Co-Creation with Artificial Intelligence*, p. 1–13.

Okamura, K. and S. Yamada (2020): 'Adaptive trust calibration for human-AI collaboration'. *PLOS ONE*, vol. 15, pp. e0229132.

Shami, N. S., K. Ehrlich, G. Gay, and J. T. Hancock (2009): 'Making Sense of Strangers' Expertise from Signals in Digital Artifacts'. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. p. 69–78.

Six, F., B. Nooteboom, and A. Hoogendoorn (2010): 'Actions that Build Interpersonal Trust: A Relational Signalling Perspective'. *Review of Social Economy*, vol. 68, no. 3, pp. 285–315.

van Berkel, N., J. Opie, O. F. Ahmad, L. Lovat, D. Stoyanov, and A. Blandford (2022): 'Initial Responses to False Positives in AI-Supported Continuous Interactions: A Colonoscopy Case Study'. *ACM Trans. Interact. Intell. Syst.*, vol. 12, no. 1.

Vereschak, O., G. Bailly, and B. Caramiaux (2021): 'How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies'. *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. CSCW2.

Wang, D., E. Churchill, P. Maes, X. Fan, B. Shneiderman, Y. Shi, and Q. Wang (2020): 'From Human-Human Collaboration to Human-AI Collaboration: Designing AI Systems That Can Work Together with People'. In: *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. p. 1–6.

Yang, Q., A. Steinfeld, C. Rosé, and J. Zimmerman (2020): *Re-Examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design*, p. 1–13.