Influences on User Trust in Artificial Intelligence in Healthcare: A Systematic Review Protocol

Background

Artificial Intelligence (AI) broadly refers to a *"set of advanced technologies that enable machines to carry out highly complex tasks effectively […] tasks that would require intelligence if a person were to perform them"* (Harwich & Laycock, 2018, p.11). AI has the potential to transform healthcare by altering the way in which we use data, treat patients and develop diagnostic tools. Accordingly, it is often perceived as part of a solution to tackling healthcare issues such as increasing costs and staff shortages. The promises of AI have resulted in large investments for its development. For example, the UK Government recently announced it planned to spend £250m on integrating AI into health services (Gallagher, 2019).

While funding is an important aspect of realising AI's potential, the success and potential of AI, like other technologies, does not only depend on AI itself but the user's trust in it (Lee & See, 2004). Previous research on trust in medical technology revealed that overtrust or undertrust in technology can result in suboptimal decision-making (Lyell & Coeira, 2016), potentially causing harm. Given that bodies such as the National Health Services (NHS) have not only recognized the potential of AI but are planning to increase its use (e.g. NHS Five Year Forward View in Harwich & Laycock, 2018), a better understanding of the underlying mechanisms of how users decide whether or not to trust an AI application (i.e. how they judge an AI's trustworthiness) is required.

While numerous studies have analysed trust in automation or technology in general (e.g. Lee & Moray, 1994; Lee & See, 2004; Pop, Shrewsbury & Durso, 2015; Yang, Wickens & Hölttä-Otto, 2016), the influences on trust in AI have received little attention.

What is trust?

Trust is an elusive concept and its definition and operationalization varies within and across disciplines and contexts, resulting in a somewhat fragmented understanding of what trust is (Chopra & Wallace, 2002). Trust is oftentimes described as a function of the trustor (e.g. user), trustee (e.g. machine) and situation (e.g. Hoff& Bashir, 2015; Merritt & Ilgen, 2008; Siau & Wang, 2018). Trust becomes relevant when uncertainty and risk are involved (Kini & Choobineh, 1998). The decision to trust depends, at least in part, on the trustee's *trustworthiness,* i.e. its attribute of being reliable and predictable. It reflects an evaluation of the trustee's attributes (McKnight, Carter, Thatcher & Clay, 2011). Given that *"technology lacks volition and moral agency, IT-related trust necessarily reflects beliefs about a technology's characteristics rather than its will or motives, because it has none."* (McKnight et al, 2011: 5). Accordingly, the present review adopts a working definition where trust in AI is defined as an individual's attitude towards an AI application about its ability to perform a particular action important to the individual.

Previous research on automated systems suggests that trustworthiness is fostered by characteristics such as competence, responsibility and dependability (Merritt & Ilgen, 2008) as well as certain design features (see Hoff & Bashir, 2015 for an overview). The studies reviewed in the trust in automation literature overwhelmingly focus on contexts such as monitoring tasks, autonomous driving or flight simulation tasks rather than healthcare. This prompts the question how transferable the findings are to the healthcare context. A recent scoping review provides an initial overview of personal, institutional and technological enablers and impediments of trust in digital health (Adjekum, Blasimme & Vayena, 2018). While the aspects identified in the review are insightful, digital health encompasses a wide range of technologies. Grouping together different technologies is useful for an initial investigation. However, given that AI is often singled out in the media and communicated as a superior technology with great opportunities and threats (see Ciupa, 2017), it seems essential to investigate AI and its unique characteristics separately in order to understand people's, especially users', perceptions and understanding of AI and, ultimately, their trust in AI.

Understanding which characteristics of an AI system convey trustworthiness does not only aid the development of new AI applications but also allows to better understand and meet people's expectations of such applications. The present systematic review seeks to contribute to this endeavour by analysing influences on user trust in healthcare AI.

Scope of the present review

The public's perception of and trust in healthcare AI are important matters to examine as they can influence not only the uptake but also the regulation of AI (e.g. Cave, Coughlan & Dihal, 2019). Similarly, understanding patients' trust in a decision that their physician reached with help of an AI application and their trust in the AI itself is of great importance. However, both scenarios pose a problem: There is *no direct interaction* between the human and the AI application. The patient could draw onto publicly available information and the physician but not the AI itself to reach a trusting decision. As a result, it would become very difficult for researchers to disentangle concerns regarding the specific AI application from broader concerns such as data privacy if the patient decided to distrust the application. Similarly, a measure of public trust would include individuals that have not actually interacted with an AI application. The public's perception may also reflect the image of AI as one big phenomenon (i.e. general view of AI) when in fact it is very context and application specific.

The lack of direct interaction has important implications for the way a trust(worthiness) judgement (i.e. an answer to the question 'Do I trust this AI to do x?') is reached as illustrated in the logic model in Figure 1.

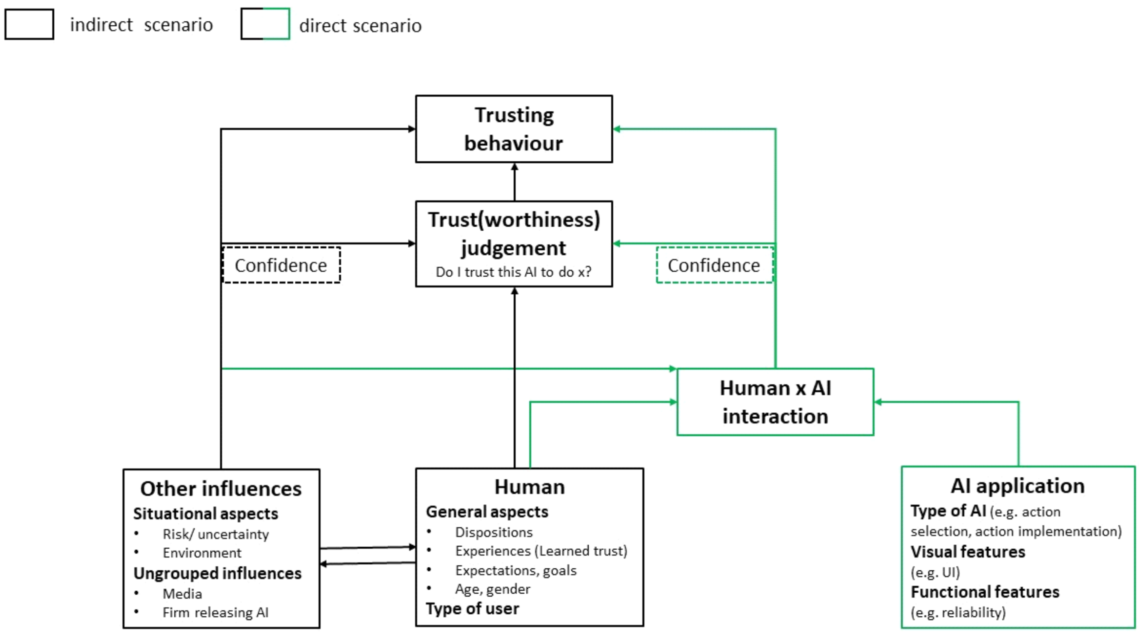

Figure 1 Logic model depicting the two different scenarios of AI usage (indirect, direct) used to determine the scope of the review. The indirect scenario (black lines) represents cases where individuals do not interact with an AI application. The green lines represent the additional paths that occur when individuals directly interact

with the AI application. I.e. the direct scenario is the full logic model whereas the indirect scenario includes only the black lines.

Figure 1 distinguishes between two scenarios: The indirect scenario (black lines) represents individuals and groups that do not interact with AI but may be affected by its existence or use (e.g. patients). In absence of an interaction with an AI application, an individual does not have the option to use experience with the application to inform his/her decision. Accordingly, the individual has to rely on factors external to the AI (e.g. media narratives or own dispositions) to reach a trust judgement. Conversely, the direct scenario (black and green lines) represents cases where individuals directly interact with an AI application (e.g. users of an AI application). Interacting with the application allows the user to experience the AI's features, enabling the user to utilize information specific to the AI on top of the other influences to reach a trust judgement. For both paths, the trust(worthiness) judgement will be made with a certain level of confidence based on the input of the different influences.

The model draws on Hoff and Bashir's (2015) review of factors influencing trust in automation. The authors grouped the reviewed evidence into three broad dimensions: Dispositional trust, situational trust and learned trust. Situational trust (ST) is highly context- and interaction-specific and variability within ST stems from the external environment as well as internal characteristics of the operator (Hoff & Bashir, 2015). External aspects include factors such as type or complexity of the system whereas internal aspects entail factors such as self-confidence or subject matter expertise. Dispositional trust (DT) defines an *"enduring tendency to trust in automation"* and thus is considered as trait-like and relatively stable over time (Hoff & Bashir, 2015, p.413). Finally, Learned trust (LT) is based on experiences *relevant to the specific automated system*. It develops from interacting with the system and informs the user's evaluation of the system (e.g. a specific AI application). While DT, ST and LT are distinct, they are interdependent, and their interactions affect the trust formation and calibration processes. Hence, all three levels influence the overall trust of a user.

Given that Learned trust emerges from an interaction between a user and an AI application, a better understanding of user trust requires a consideration of Learned trust (Human x AI interaction in Figure 1) along with the other two dimensions (i.e. user and other influences). This is only possible through considering the direct scenario in Figure 1 as Learned trust is largely absent in the indirect scenario. Accordingly, the present review focuses on the direct scenario to better understand influences on user trust in healthcare AI.

Review Questions

What influences user trust in healthcare AI? Specifically,

- What characteristics, that is functional and/ or aesthetic features of a healthcare AI, influence user trust in the AI?
- What factors external to the AI application influence user trust in the AI?

Methods

This review aims to provide an overview of influences on user trust in healthcare AI.

Literature Search

I will use a three-part search strategy to identify studies meeting the inclusion criteria: (1) I will search electronic databases for published work and grey literature, using a comprehensive search strategy for user trust in healthcare AI; (2) I will search the reference lists of primary studies included in the review and the reference list of relevant, previously published reviews (e.g. Adjekum et al., 2018); (3) I will contact authors of included papers to identify further relevant literature.

- *Searching databases*

  The following electronic databases will be searched with a pre-determined search strategy (see Appendix A for an example).

  - ACM digital library
  - IEEE Explore
  - NHS Evidence
  - Ovid ProQuest Dissertations & Thesis Global
  - Ovid PsycINFO
  - PubMed
  - Web of Science Core Collection

In addition, the online index of the most frequent journals of included papers will also be searched.

The search strategy may vary between databases as the strategy will have to be adapted to the specificities of the different database search interfaces. For example, databases such as PubMed are focused on literature from the healthcare environment and as such do not necessarily require the healthcare component in the search terms whereas others (e.g. Web of Science) do (see Appendix A for example).

The above databases were chosen in consultation with a research librarian to ensure that literature from different disciplines investigating the topic under review are represented (e.g. computer science, psychology).

- *Reference searches*

  Bibliographies of papers matching the eligibility criteria below will be searched by hand to identify any further, relevant references, which will be subject to the same screening and selection process. In addition, Microsoft Academic will be used to explore papers citing the included studies and these too will be subject to the same screening and selection process.

- *Expert consultations*

  Authors of identified and included papers will be contacted by email to identify additional (grey) literature or research that has not been found through the above process. A record of the experts contacted will be kept.

Search Strategy

The search strategy is based on formulating keywords around the main themes of the review.

1) Trust

Since trust is an ill-defined concept in the literature (Montague, Kleiner & Winchester, 2009), the search strategy will not only use the word trust but also use related terms such as trustworthiness, credibility, distrust, mistrust and confidence which are often used synonymously to trust.

2) AI

The theme of AI will be disaggregated into terms related to AI such as machine learning, intelligent agent, expert systems etc. to capture the multitude of terms used to describe AI applications. Preliminary searches revealed that more specific AI terminology such as natural language processing and artificial neural networks result in literature more focused on the implementation of a specific algorithm rather than an evaluation of an application involving a user and will thus not be used to construct the AI term to reduce the noise in the search. However, the terms will not be explicitly excluded to allow for potential overlap between specific and more general AI terminology.

3) Healthcare

As mentioned above, some databases will not require a healthcare search term. For those that do, healthcare will be described as healthcare, health care or indicated through words such as medic* or clinic*.

The searches will be restricted to publications available in English and German with no publication date restriction.

After gathering the literature, the below eligibility criteria will be applied. All references will be screened by one reviewer (EJ) whereas 10% of the sample will be screened by a second reviewer (JK) using a three-stage approach to review the title, abstract and full text.

Eligibility Criteria of Included Studies

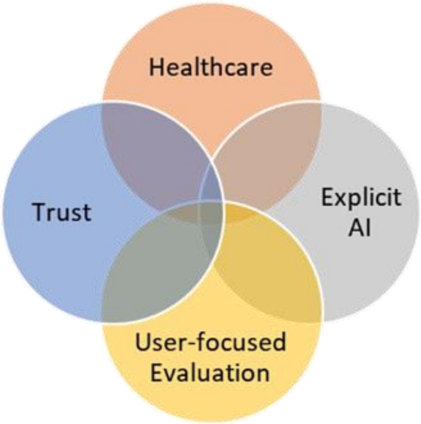

Figure 2 *Venn diagram depicting the inclusion criteria. Only studies meeting all four criteria (i.e. intersection of the four aspects) will be included in the review.*

Each study will be required to meet all of the following criteria:

(1) Be set in a healthcare context

(2) Focus on an AI application and communicate the presence of AI explicitly (e.g. through appropriate language).

(3) Be an empirical study investigating the relationship between trust and another variable with trust being the/an outcome variable.

(4) The AI application is such that it gathers information, analyses information or provides a recommendation rather than implementing an action (i.e. human remains ultimate authority).

(5) Focus on an AI application-that is **not** robotics. Studies investigating AI robotics will be excluded as the physical presence of AI adds a dimension to perceptions of AI which lays outside the scope of the current review.

(6) Evaluate the AI application with regards to users (e.g. user perceptions, experiences of AI) rather than reporting the implementation and performance-based evaluation of the AI.

There will be no limits on study participants in terms of age, gender, ethnicity of profession. There will be no limit on study setting and studies of all levels of healthcare settings (primary, secondary and tertiary) will be included.

Failure to meet any of the above eligibility criteria will result in exclusion from the review. Any disagreement between two review authors over the eligibility of a particular study will be resolved through discussion with a third review author (JT). The number of excluded studies and the reasons for exclusion will be recorded at each stage (see Figure 3).

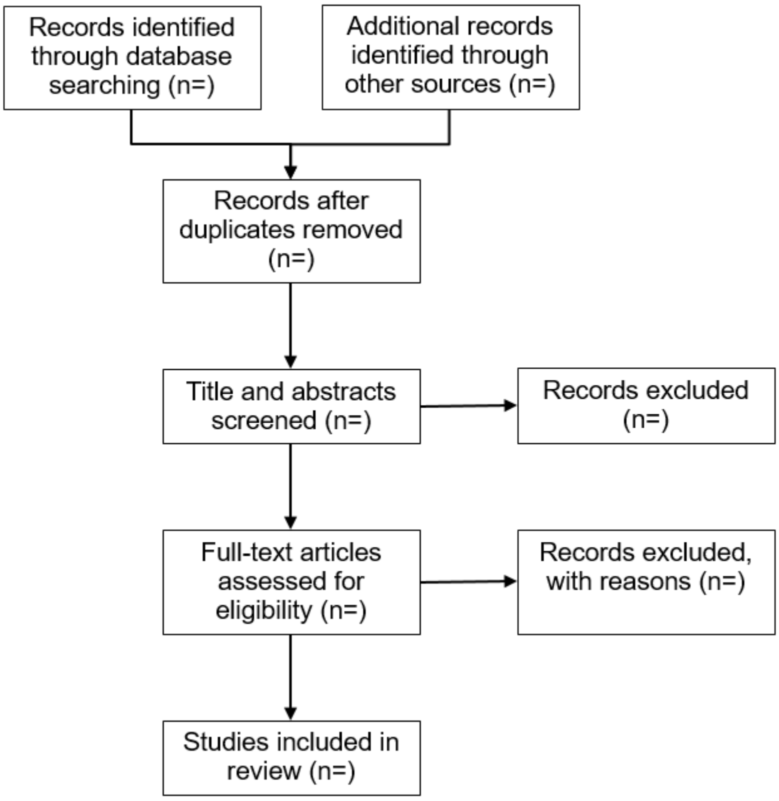
Records identified through database searching (n=)
Additional records identified through other sources (n=)
Records after duplicates removed (n=)
Title and abstracts screened (n=)
Records excluded (n=)
Full-text articles assessed for eligibility (n=)
Records excluded, with reasons (n=)
Studies included in review (n=)

Figure 3 *Study flow diagram in line with Preferred Reporting Items for Systematic Reviews and meta-Analyses (PRISMA)*

Data extraction

Two independent reviewers (EJ and DK) will extract data from the included studies. Any disagreement will be resolved by discussion between the two reviewers (EJ and DK), and further disagreements will be resolved with the help of the third reviewer (JT).

The following information will be extracted from each included study:

- General information (author, title, year of publication, country of origin, journal)

- Study characteristics (aims, design, setting, ethical approval)
- Participant characteristics (type of stakeholder, age, gender)
- AI characteristics (type of application)
- Features pertaining to participants' trustworthiness judgement of the AI application
    - Functional and visual aspects of AI application
    - Other influences (external and human-related)
- Other information (author's conclusions, references to other relevant studies)

We will approach authors for any relevant missing information from the study.

Assessing risk of bias

Two authors (EJ, DK) will assess the risk of bias and any disagreements will be resolved by a third reviewer (JT).

Based on the study design of included studies, an appropriate tool (e.g. ROBINS) will be chosen. It is likely that the literature will include not only randomized control trials (RCTs) but also focus groups, interviews, surveys etc. Given the variety in the data, several tools might be applied.

Data synthesis

The review will employ framework synthesis driven by a logic model as the method allows for data to be coded into pre-defined themes as well as new themes to emerge. Having an emergent framework allows to develop a trust in AI-specific framework from the initial framework that was borrowed from the automation literature (Figure 1). The initial model will be revised and complemented throughout the reviewing process. In particular, data from included studies will be mapped onto aspects of the initial framework presented in Figure 1. For data that cannot be accommodated, data will be aggregated to generate new themes to supplement the framework's existing themes. Given that the framework will evolve during the synthesis, an "iterative process of aggregating and reconfiguring the data" (Gough, Oliver & Thomas, 2017, p. 189) will be required. The synthesis will culminate in a revised logic model of Figure 1 to reflect the understandings gained from the review. Each key theme will be summarized (e.g. in form of a table) and all versions between the initial and final logic model will be recorded.

Glossary

The following list defines key terms of the protocol in the context of trust in AI.

Artificial Intelligence (AI): *"set of advanced technologies that enable machines to carry out highly complex tasks effectively […] tasks that would require intelligence if a person were to perform them"* (Harwich & Laycock, 2018, p.11).

Automation: Umbrella term for technologies that function automatically, i.e. without continuous input from a human.

Characteristics: Features of an AI application that a user can experience. Characteristics can be visual such as elements of the user interface as well as functional such as the reliability of an AI application.

Dispositional trust (DT): An individual's tendency (disposition) to trust a person or technology.

Healthcare AI: Umbrella term to describe different types of applications of AI to the healthcare domain (e.g. decision support tool in diagnosing a certain type of cancer).

Learned trust (LT): An individual's trust that is based on experiences relevant to a specific AI application. It develops from interacting with the system (past and present interactions).

Situational trust (ST): Depends on the specific context of an interaction. Variability in situational trust stems from the external environment as well as internal characteristics of the user (Hoff & Bashir, 2015). External aspects include factors such as type or complexity of the system whereas internal aspects entail factors such as self-confidence or subject matter expertise.

Trust: An individual's attitude towards an AI application about its ability to perform a particular action important to the individual.

Trusting behaviour: An individual's behaviour that indicates trust in the AI application e.g. acceptance of a specific technology or continued usage.

Influences: Aspects that influence an individual's trustworthiness judgement and/ or behaviour. They can be grouped into human-related, AI-related and other (external) influences.

Trustworthiness: An AI application's attribute of being worthy of trust in the eyes of the user

Trust(worthiness) judgement: An individual's judgement answering the question "Do I trust this AI application to do x?" which includes a judgement as to how trustworthy the individual considers the AI application to be.

User: Any type of person that interacts directly with an AI application.

References

Adjekum, A., Blasimme, A., & Vayena, E. (2018). Elements of Trust in Digital Health Systems: Scoping Review. *Journal of medical Internet research*, *20*(12), e11254.

Cave, S., Coughlan, K., & Dihal, K. (2019, January). Scary robots': examining public responses to AI. In *Proc. AIES http://www. aies-conference. com/wp-content/papers/main/AIES-19_paper_200. pdf*.

Chopra, K., & Wallace, A. W. (2002). Trust in electronic commerce. In *Proceedings of the 36th Hawaii International Conference on System Sciences*.

Ciupa, M. (2017). IS AI IN JEOPARDY? THE NEED TO UNDER PROMISE AND OVER DELIVER–THE CASE FOR REALLY USEFUL MACHINE LEARNING. *Computer Science {\&} Information Technology (CS {\&} IT)*.

Gallagher, J. NHS to set up national artificial intelligence lab. (2019, August 8), BBC News. Retrieved *https://www.bbc.co.uk/news/health-49270325*

Gough, D., Oliver, S., & Thomas, J. (Eds.). (2017). *An introduction to systematic reviews*. Sage.

Harwich, E., & Layock, K. (2018). Thinking on its own: AI in the NHS. *Reform). See http://www. reform. uk/wp-content/uploads/2018/01/AI-in-Healthcare-report_.pdf.*

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, 57(3), 407-434.

Kini, A., & Choobineh, J. (1998, January). Trust in electronic commerce: definition and theoretical considerations. In *Proceedings of the thirty-first Hawaii International conference on System sciences* (Vol. 4, pp. 51-61). IEEE.

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International journal of human-computer studies*, *40*(1), 153-184.

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

Lyell, D., & Coiera, E. (2016). Automation bias and verification complexity: a systematic review. *Journal of the American Medical Informatics Association*, *24*(2), 423-431.

Mcknight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on Management Information Systems (TMIS)*, *2*(2), 12.

Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human Factors*, *50*(2), 194-210.

Montague, E. N., Kleiner, B. M., & Winchester III, W. W. (2009). Empirically understanding trust in medical technology. *International Journal of Industrial Ergonomics*, *39*(4), 628-634.

Pop, V. L., Shrewsbury, A., & Durso, F. T. (2015). Individual differences in the calibration of trust in automation. *Human factors*, *57*(4), 545-556.

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

Yang, X. J., Wickens, C. D., & Hölttä-Otto, K. (2016, September). How users adjust trust in automation: Contrast effect and hindsight bias. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 60, No. 1, pp. 196-200). Sage CA: Los Angeles, CA: SAGE Publications.

Appendix

Appendix A: Example of search syntax from Web of Science Core Collection

((((TS=(trust* OR mistrust* OR distrust* OR suspic* OR "reliance" OR "credibility" OR ("confidence" NOT (confidence NEAR/3 interval*)))) AND ((TS=("artificial intelligence" OR "machine learning" OR "intelligent agent" OR "intelligent agents" OR "intelligent system" OR "intelligent systems" OR "expert system" OR "expert systems" OR "intelligent automation" OR "computational intelligence" OR "machine intelligence")) NOT (TS=(web* OR tele* OR robot*))) AND (TS=(health* OR "health care" OR medic* OR clinic*))))) *AND* **LANGUAGE:** (English OR German)