INPLASY

INPLASY2023120100

doi: 10.37766/inplasy2023.12.0100

Received: 26 December 2023

Published: 26 December 2023

Trust in Human Autonomy Teaming: A scoping review protocol for investigating role of interpretability, explainability, assurance, transparency, situational awareness and governance

Padmanabhan Poti, S[1]; Stanton, CJ[2].

Corresponding author:
Siri Padmanabhan Poti

19547451@student.westernsydney.edu.au

Author Affiliation:
The MARCS Institute of Brain, Behaviour and Development, Western Sydney University.

ADMINISTRATIVE INFORMATION

Support - Scholarship and Australian RTP.

Review Stage at time of this submission - The review has not yet started.

Conflicts of interest - None declared.

INPLASY registration number: INPLASY2023120100

Amendments - This protocol was registered with the International Platform of Registered Systematic Review and Meta-Analysis Protocols (INPLASY) on 26 December 2023 and was last updated on 26 December 2023.

INTRODUCTION

Review question / Objective To identify the ontologies, role and state of the science of interpretability, explainabililty, assurance, situational awareness, transparency and governance in the context of trust within Human Autonomy Teaming.

Background Autonomous Artificial Intelligent (AI) agents can independently perform complex tasks and exhibit decision-making abilities. However, this raises the risk of unexpected, unintended and / or poorly understood outcomes. Accordingly, human operators provide assistance or supervision to such systems through human-autonomy teaming (HAT). Due to the delegation of risks and decision-making by human operators to autonomous systems, the establishment of 'trust' in HAT is key in ensuring safety and effectiveness in mission-critical operations.
Generally, there is understood to be a role of interpretability, explainabililty, assurance and transparency in fostering trust within HAT. To understand what the effect of these are on trust in HAT, it is essential to clarify how these constructs are understood in literature, and how they are inter-related. Additionally, it is important to bring out any difference in the understanding of these constructs in current literature.

Rationale While there is considerable literature on interpretability and explainabililty in AI, this scoping review seeks to explore literature as to the established ontologies of interpretability, explainabililty, assurance and transparency in the context of trust within Human Autonomy Teaming (HAT) and related governance. The review seeks to explore the currently understood effect of interpretability, explainabililty, assurance, transparency, situational awareness and governance on trust, in the context of Human Autonomy Teaming. It also investigates the inter-relatedness of the constructs, as also any differences in the understanding of these constructs in current literature.

This scoping review is part of a larger project related to fostering trust in human-autonomy teams. This scoping review will be followed with a systematic review on specific questions.

METHODS

Strategy of data synthesis The Terms included are: (("Interpretab*" OR "Explainab*" OR "Assurance" OR "Transpare*" OR "situation* awareness" OR "Governance") AND "Trust*" AND ("Human Autonom* Team*" OR "Human Machine Interaction" OR "HAT" OR "HMI" OR "Autonom* System*")), intended to cover interpretability, Interpretable, Explainability, Explainable, Assurance, Transparency, Situational Awareness and Governance in the context of Trust, Trustability, Trustworthy Autonomous System, Human Autonomy Teaming, Human Machine Interaction.
A qualitative synthesis is undertaken of the ontologies and inter-relationships to bring about a disambiguation of key terms and their state of the science. The Databases included in the review are Scopus, Web of Science, ProQuest Central, Science Direct, ACM Digital Library.

Eligibility criteria Peer Reviewed Articles and Conference Papers published in English from 2020 to 2024 are included.
Grey Literature, Reviews and Book Chapters are excluded.

Source of evidence screening and selection An initial search strategy and list of databases is agreed upon. Search is conducted by one reviewer and the other reviewer exercises supervisory oversight. After reviewing Titles, Keywords and Abstracts of all collected candidates, duplicates are screened out. Further articles are screened out and the reason for each removal is recorded. The finalised set of candidates are uploaded onto NVivo for qualitative analysis. When the first author runs a qualitative analysis, if there are further candidates that do not fit the objectives of the search, these are identified and removed with a documented reason for removal.

Data management Data will be managed in NVivo and EndNote. The search strategy, search terms in each database and outputs from each database will be recorded in spreadsheets. The pdf of articles will be uploaded in NVivo (Local) and EndNote (Local computer with online backup). Additional Backups will be on Western Sydney University's Microsoft OneDrive.

Reporting results / Analysis of the evidence The data will be analysed in NVivo using qualitative analysis. The research questions are coded onto each of the candidates and aggregated. \The coded portions are analysed using tools in NVivo. The results are presented as a scoping review report.

Presentation of the results The presentation of the results will be as the definitions, disambiguation and inter relationships of the constructs being studied, as a table.
Each construct will be presented as a section with any differences in definition, key distinctions and any effect they have on trust within HAT.
There will be a presentation of any gaps in literature. Importantly, the scope of a follow-up systematic review will be formulated.

Language restriction The study is limited to Articles and Conference papers published in English.

Country(ies) involved The scoping review is carried out in Australia.

Keywords Interpretability; Explainability; Assurance; Transparency; Trust; Autonomy.

Dissemination plans Publish as a Peer Reviewed Paper. All or some portions of this study will be included in the doctoral thesis of Author 1, supervised by Author 2.

Contributions of each author

Author 1 - Siri Padmanabhan Poti - Author 1 drafted the manuscript, prepared search strategy, executed search, uploaded data into NVivo, analysed data, drafted presentation of results.
Email: 19547451@student.westernsydney.edu.au
Author 2 - Christopher Stanton - Author 2 provided inputs into search strategy, conducted supervisory review of activities at each step in the process and final review of output.
Email: c.stanton@westernsydney.edu.au