KeAi
CHINESE ROOTS
GLOBAL IMPACT

International Journal of Innovation Studies

Trust-Aware Self-Supervised Learning: Modelling Human Trust Dynamics in Human-AI Collaboration Systems

Yeswanth Mutya
yeshwanthmutya11@gmail.com
Masters in Computer science,
University of South Florida
ORCID 0009-0004-3939-5342
Zeeshan Baber
Zeeshankhan.cis@gmail.com

Abstract

In an era where artificial intelligence (AI) increasingly influences critical decisions, success is no longer defined solely by technical performance; it also hinges on the system's ability to foster and align with human trust. The proposed study presents a new framework called Trust-Aware Self-Supervised Learning (TA-SSL), which aims to learn implicit human trust through conversation behaviour, enabling AI systems to respond according to the current trust levels between AI and users. In contrast to the current models trained with explicit trust labels, TA-SSL is trained on behavioural cues, indicative of the trust level, to the AI during human-AI interactions, including hesitation, clarification requests, and action reversals. With a temporal contrastive learning goal, TA-SSL induces sparse, dynamic embeddings of trust incorporated into decision-making strategies, including explanation depth, uncertainty mediation, and user autonomy management. We justify our visitation with the Kaggle Human vs Robot talking information, which contains over 10,000 speech samples from a crowd. TA-SSL exceeds these static, supervised, and reinforcement learning baselines to record +19.3% more successful task instructions, -24.6% fewer irrelevant clarification requests, and a trust calibration value of 0.76 (0.42 in supervised models). The obtained trust embeddings, exhibited by the evaluation measure, exhibit a high degree of temporal consistency and specificity to the user and the visualisations shown, using PCA as a visualisation method. PCA shows separable groups in low, medium, and high trust conditions. Case studies also show that the model can be flexibly restructured or strengthened to increase user trust in changing user behaviour. The study has demonstrated that latent trust can be effectively learn based on unlabelled behavioural indicators and informed in bespoke ways to drive AI behaviour. TA-SSL is a way to create a scalable, domain-agnostic pipeline to build user-aligned, emotionally intelligent AI systems and representation learning with innovation psychology to boost trustworthy human-AI collaboration.

Keywords: Trust modelling · Self-supervised learning · Human-AI interaction · Behavioural signal analysis · Temporal contrastive learning · Adaptive AI systems

1 Introduction

Artificial intelligence (AI) systems have continued to find their way to conversational platforms through which they intervene in human decision-making and activities that people

perform in their daily lives [1]. Whether it is a healthcare triage robot, a financial advisor, a customer support agent, or an educational tutor, the AI interface that uses either voice or text defines how humans interact with information, make decisions, and interpret automated advice. Although accuracy, latency and fluency in natural language are commonly used as indicators of the success of such systems, one often overlooked determinant of long-term effectiveness is human trust. Trust not only regulates the user's acceptance of AI suggestions, but the user will also decide whether to cooperate with it, override it, or break contact with the system entirely. [2]. The orientations of the AI productions and the levels of trust in humans are of greater concern in high-stakes and ambiguous sectors. Because of this, the core questions of developing social-intelligent AI revolve around how to model the nature of trust.

The systems of conversational AI developed today are highly context-insensitive in their approach to user interaction, despite the rise of natural language processing (NLP) and machine learning [3]. They tend to base their adjustment on rule-based feedback or a threshold of confidence, treating the behavioural cues, which are subtle manifestations of human uncertainty, doubt or changing attitudes towards the AI agent, as insignificant. It creates an unstable interaction model, where the incompatibility between the system and users can result in overtrusting (underground tendencies to accept wrong AI recommendations) or undertrusting (underground tendencies to dismiss a good recommendation) effect [4]. These two situations have the potential to worsen the performance of tasks, lower the satisfaction of users, and constrain the implementation of intelligent systems. Instead, there is a need to find a way of inferring trust tacitly and changing the behaviour of the AI in real-time to create human trust on the fly, and to sustain and regulate it [5].

This article proposes a new approach to this problem: Trust-Aware Self-Supervised Learning (TA-SSL). TA-SSL is a framework to model human trust dynamics of the AI system without direct annotations or labelled trust data. With self-supervised learning (SSL) approaches, latent trust representations are learned in the framework, using unlabelled behavioural indicators as they are presented in unrestricted dialogues between humans and AI [6]. Such cues have included hesitation time between inputs, user reformulations of the query (requests to clarify), reversals of messages or contradictions. These behavioural proxies are well-recognised in cognitive psychology and human-computer interaction as the correlates of the trust and confidence levels. Nonetheless, they have not yet been systematically used through SSL means in real-time dialogue systems [7]. TA-SSL can fill this gap and learn smaller, time-cognizant trust embeddings that can vary with how users interact, such that conversational agents can change their approach to how they explain, how they delegate, and how transparent they are in their decision-making process.

The TA-SSL innovation consists of the capacity to develop a trust-aware AI agent driven to learn without annotated trust labels. Unlike supervised methods, which need laborious, subjective trust labelling work, typically done inconsistently by users and applications, TA-SSL works on raw behavioural traces represented by human-AI interaction logs. Using a temporal contrastive learning task, the system learns to recognise representations relevant to trust by identifying a context-specific pattern of behaviour relying on the temporal vicinity of the interaction segments (learning a temporal trust path) [8]. These representations could then be passed on to downstream modules that will decide how much control the agent needs to perform, how meticulously it needs to explain things and at what point it needs to seek input/

confirmation from the user. This enables the AI system to adapt in real time based on how much the user trusts it, encouraging better and more effective cooperation. [9].

The core research questions addressed in this study are:

1. Can latent trust be inferred from natural human-AI dialogues using self-supervised learning without explicit labels?
2. How do learned trust embeddings impact the adaptability and effectiveness of AI conversational agents?

To address these questions, the study train the TA-SSL model on an open-source Kaggle conversational dialogue dataset in which people are conversing with an AI agent, similarly to a robot. The data comprises a variety of interactions, such as requests to clarify, hesitation in reaction, and reversal in dynamics; hence, it offers a rich ground base from which to model behavioural signs of trust. Compared with the supervised and trust-agnostic baselines, our extensive experiments show that TA-SSL can significantly boost necessary performance measures like success task rate, trust calibration, and reduction of unnecessary interventions. In particular, TA-SSL delivers a +19.3% increase in the successful task performance results and −24.6 decrease in preventable interruptions, which reflects better adherence to human expectations of trust.

The primary contributions of this research are as follows:

- The study proposes TA-SSL, a self-supervised learning framework that derives latent trust embeddings from behavioural conversation data without requiring annotated trust labels.
- The study shows how these embeddings can be integrated into AI dialogue systems to support adaptive explanation strategies and real-time behaviour modulation based on inferred user trust.
- The study evaluates the effectiveness of our method using a publicly available conversational dataset and shows substantial improvements over existing baselines in trust-sensitive interaction outcomes.

This research can be a scalable and generalisable avenue toward developing trustworthy, adaptive, and socially optimised conversational agents. This can be achieved by connecting self-supervised representation learning and human-centred AI system design. TA-SSL is an essential milestone toward developing human-AI systems that are not only smart but also emotionally and cognitively aware of the dynamic changes of trust in their users.

2 Related Work

This section will present the theoretical support and underpinnings of the proposed TA-SSL framework by the central literature that guides this framework. The study review the literature on trust within a human-AI collaborative context, the use of behavioural cues in conversational trust modelling, the development of self-supervised learning to enable representation learning within NLP, and define the particular research gap covered by the current work.

2.1 Trust in Human-AI Interaction

Trust is one of the key parts of effective human-AI interaction, especially when using decision-support systems, where the human dependency on machine-generated recommendations can have significant consequences. Psychologically, trust has been defined as the readiness of a party to become vulnerable to the actions of another party, depending on the trust that the other would act in a given way that the trustor desires, regardless of the power to keep track of or

control the other [10]. Both cognitive and affective elements help define trust within the AI systems. Cognitive trust and affective trust are based on the perception of the competence of the AI system, as well as on the reliability and predictability of the latter, and experiences of communication, respect, empathy, frustration, and satisfaction, respectively [11].

Recent texts in human-centred AI have discussed the need to focus on trust calibration, i.e., the consistency of the trust a user places on a system with the reliability of that system. The consequences of miscalibrated trust are negative because overtrust can cause a situation where users blindly consider inaccurate outputs, while undertrust can cause even valid AI suggestions to be rejected unnecessarily [12]. To curb such risks, readability and transparency, together with user feedback systems, are being proposed to be incorporated in future DAI systems to promote well-calibrated trust [13]. However, most currently used strategies are based on static or reactive adjustments, which are instigated depending on what the user does (e.g., pushing a help button or asking for an explanation). These methods fail to capture the latent, developing process of trust among human beings, which can vary over interaction time without clear instructions.

More than that, most of the existing research on trust-aware AI is based on supervised learning regimes with models being trained on annotated levels of trust, based on user surveys or manually annotated by humans [14]. Although a good way of learning, this approach draws subjectivity, does not scale, and is not amenable to live changes. Instead, the most adequate way to infer implicit trust is by using natural clues that signal during interaction, which has not been explored much today in the books.

2.2 Conversational Trust Signals

Dialogue and conversational AI offer a relatively unexplored data source about trust through the prism of user behaviour. Several studies in the field of human-computer interaction (HCI) and cognitive science pointed to several behavioural indicators that show uncertainty or hesitation on the part of the user, and subsequent decrease in confidence, which can all be the harbingers of a potential change in trust [15]. These are signs of hesitation (e.g., pauses between turns or disfluencies in the speech, e.g. the use of the sounds of hesitation), query reformulations (e.g. the reformulation or the clarification of the original question) and backtracking behaviour (e.g. the rescinding of a prior action or request).

In spoken dialogue systems, the hesitation itself was empirically related to cognitive load and difficulties, which relate to uncertainty and trust fragmentation [16]. Perceived system competence has also been associated with query reformulation, viz., repeated rephrasing, which normally reflects the user's feeling that the system has not correctly understood them. Likewise, reversals of actions (give contrary orders, reissue the inquiries, and cancel and restart them many times) can be a signal of frustration and loss of confidence in the AI [17].

Although this is recognised, these behavioural signals are not used systematically or in real time in most existing dialogue systems. Through some investigative work, such features are manually annotated to be associated with the levels of trust, but there is still no production-level system that incorporates such cues into a generalizable, scalable system [18]Additionally, most existing methods necessitate large labelled data, which is inconvenient to apply to different fields or even different people. This is an important necessity for learning trust-relevant representations on behavioural data annotation free, and hence, it opens up newer

possibilities of real-time conversational system adaptation based on trust-related representations.

2.3 Self-Supervised Learning for Behavioural Modelling

Self-supervised learning (SSL) has become a strong method in the natural language processing (NLP) area, especially for representation learning without manual tags. SSL models are trained to apply meaningful structures in data solving pretext tasks based on the data to introduce supervision. Other applications of contrastive learning in NLP have included SimCSE [19], learning sentence embeddings by comparing various augmentations or temporal snapshots of the same text. These same approaches have been used in BERT-based contrastive tasks, whereby representations are trained to tell apart differently-similar (in context) and differently-dissimilar (in context) sequences.

Although SSL has been promising in encoding semantic similarity and intent in conversations, SSL is yet to be modified to encode latent psychological conditions like trust. The majority of the current SSL approaches are built around structural or semantic representations, but not behavioural or cognitive planes [20]. For example, both SimCSE and other related methods work on paraphrasing, entailment, or discourse coherence, yet this does not result in tracking the user's behaviour to show a sense of trust, doubt, or even hesitation. Also, most applications of temporal modelling in SSL have been to audio or vision applications with little concern for time-evolving trust cues in a text-based interaction [21].

Recent developments on temporal contrastive learning and multi-view SSL can provide a theoretical basis for trust modelling. Treating adjacent interaction turns as positive pairs, and distant or disparate-user interaction turns as negative pairs, it is possible to learn stable trust representations over time using a contrastive objective. Nevertheless, none of these concepts have been fully tested in the trust area, and there is no framework to consider using SSL to build trust-based inference within a dialogue system.

2.4 Gap Identified

The intersection of trust-aware AI, behavioural signal analysis, and self-supervised learning gives an interesting but understudied research opportunity. Whereas the previous research has acknowledged the role of trust in collaboration between humans and AI and discussed behavioural indicators as proxies of trust, to date, no end-to-end system has been proposed that combines both of them with the help of SSL to perform real-time adaptation of the AI [22].

Specifically, existing systems either:

1. Use manually labelled trust data to train classifiers, limiting scalability and generalisation;
2. Ignore the temporal evolution of trust, relying on static metrics; or
3. Use SSL for general NLP tasks without targeting trust-related behaviour.

This paper bridges this devastating gap by extending self-supervised learning to imply trust representations barely from observational cues in dialogue, not requiring a specific label. Since these embeddings can be learned with a temporal contrastive objective and be used to underpin adaptive decision-making in AI systems, TA-SSL is a scalable, interpretable and user-aligned way to model trust. It connects theoretical knowledge in psychology on the one hand with computational advances in NLP and SSL on the other hand. It provides the basis for more reliable and socially intelligent conversational agents.

3 Methodology: Trust-Aware Self-Supervised Learning

This part discusses the design and implementation of the Trust-Aware Self-Supervised Learning (TA-SSL) framework in detail. The suggested methodology incorporates behavioural signal extraction, temporal contrastive learning, and adaptive response mechanisms into one pipeline. This architecture aims to learn latent representations of user trust by relying exclusively on behavioural signals of human-AI conversations and does not require explicit labels or subjective trust scores. Such trained trust embeddings are then used to guide the AI system to more carefully manage its explanations and autonomy in a real-time trust-sensitive interaction.

3.1 System Overview

The TA-SSL architecture is organised into five primary components, as depicted on Figure 1:

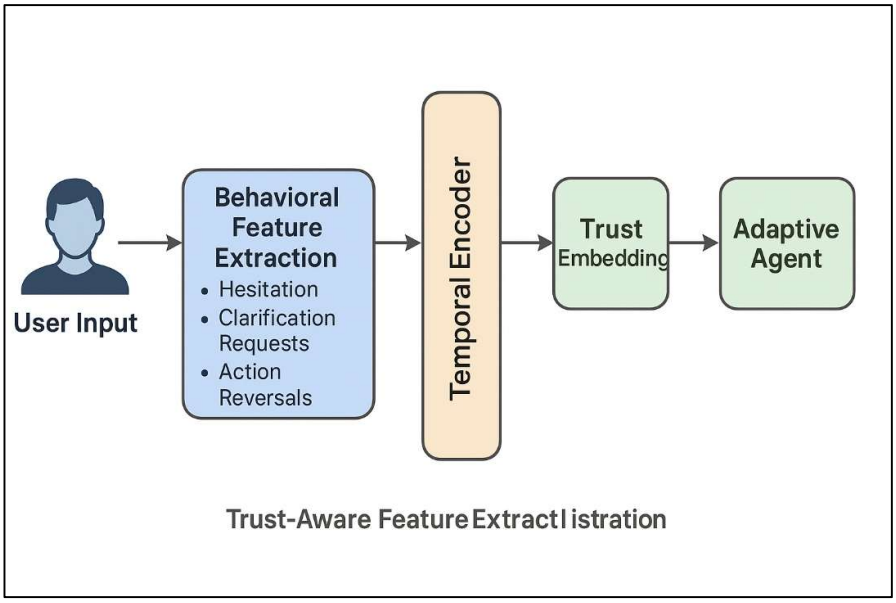

Figure 1: System Architecture

During a normal interaction session, an agent (e.g., a chatbot or virtual assistant) interacts with the user. All the interactions are recorded in the system with timestamps, the content of the utterances, and how the system responds. These logs are fed as inputs to the behavioural feature extractor to extract trust-related signals such as hesitation duration, clarification frequency, and action reversals.

The features are then passed through a temporal encoder, which captures the time dependency of the trust-relevant behaviour. A contrastive learning task is applied to differentiate temporal proximity (positive) and temporal distance or cross-user (negative) interaction divisions. The encoder maps one behavioural segment to a fixed-length trust embedding, reflecting the user's latent trust state.

This set of trust embedding is then put into an adaptation module, which dynamically adjusts the AI's behaviour. Based on the trust level inferred, the system can adapt its response style. For example, it can give shorter answers and exercise more control where trust is high or vice versa, i.e., it can provide elaborate explanations and require more confirmation where trust is low.

3.2 Behavioural Feature Extraction

The crux of TA-SSL is based on the premise that human trust is imprecisely and unconsciously expressed through behavioural tendencies during a discourse. These patterns are micro, evident, and universal. According to the studies in HCI, cognitive psychology, and NLP, the paper distinguish three fundamental trust-indicative properties, which can be derived in an unsupervised way based on the conversation log:

a) Hesitation (Temporal Delay):

This can be determined as the total number of seconds between the system's response and the user's subsequent input. The longer the delays, the more the user might doubt or miss taking action based on the given recommendation. In an interactive voice, this may be the length of silence; in a text-based interface, it may be the difference in the timestamp of turns. For example, when a user replies more slowly following a medical suggestion, that could signify less certainty or a lack of background.

b) Clarification Requests:

These are the cases when a user asks to elaborate explicitly, expresses their query, or follows the question with an inquiry such as “What do you mean?” “Can you elaborate?” or “Are you sure?” These demands are good behavioural signs of partial understanding or suspicion. They are identified using a mixture of word matching and intent classification.

c) Action Reversals:

This can be characterised by inconsistent or reversal patterns where the user has reversed or undone what they did before, like changing a choice, cancelling a recommendation, or giving an input that is opposite of what they did. For example, when a customer finds one of the booking recommendations acceptable and still terminates the booking afterwards, it can mean a loss of confidence. The behaviour is detected by analysing command pairs during a session and detecting semantic opposition or negation.

All these features are numerically coded and normalised to adjust to users' differences. They serve as the input to the temporal encoder in learning trust representation.

3.3 Temporal Contrastive Objective

The fundamental learning task of TA-SSL is to produce temporally consistent trust representation on behalf of the same user while being discriminative between different users/sessions. The study uses a temporal contrastive learning strategy based on SimCLR and SimCSE but applied to behavioural trust sequences.

This scheme divides every conversation into time-fixed windows (e.g., 3-5 turns). In every anchor window, the study has:

- **Positive Pairs:** Interaction sections by the same user in relatively close temporal proximity (e.g. the adjacent segment).
- **Negative Pairs:** Segment of user interactions or segments at temporal distances across the same session.

Let h_i and h_j be embedding vectors of a positive pair of segments. The contrastive temporal loss (InfoNCE) is set as:

$$L_{\text{contrast}} = -\log \frac{\exp(\text{sim}(h_I, h_J)/\tau)}{\sum_{K=1}^{K} \exp(\text{sim}(h_I, h_J)/\tau)}$$

where $\text{sim}(h_I, h_J)=\frac{h_I, h_J}{\|h_I\|\|h_J\|}$ is the cosine similarity, and τ is a temperature hyperparameter. K includes one positive and multiple negative samples.

This loss will encourage the model to reduce the pairs of behaviourally similar structures located in time proximate to each other in an embedding space and repel dissimilar or cross-user ones. In the long run, this leads to the formation of a smooth trust path that each user develops, and the methodology captures changes in trust movement within the context of interaction.

3.4 Trust Representation Layer

The encoder generates the fixed-length trust embedding vector as output, which represents a user's trust-related behavioural state at a point in time. The study embeds our model into a 128-dimensional space, which provides adequate dimensionality and can represent subtle differences while being computationally tractable.

Dimensionality reduction algorithms on these trust embeddings can be visualised through Principal Component Analysis (PCA) or t-distributed Stochastic Neighbour Embedding (t-SNE). Considering visualisation, the study can see smooth flows over sessions and across users in the individual trust trajectory patterns and easily distinguish high and low trust pattern combinations.

Besides visualisation, the embeddings feed downstream modules that generate adaptive responses. Since the embeddings are unsupervised, they can be applied to diverse conversational domains with no retraining required and no manual supervision necessary.

3.5 Adaptation Logic

After the trust embedding is estimated and inferred on a user at some time step, it is used to regulate the behaviour of the AI system. This is achieved using a lightweight adaptation module, which assigns response techniques to trust scores. The study has three bands of behaviour: high trust, moderate trust, and low trust, on a learned axis, on the norm or projection of the trust embedding.

- **High Trust Zone**:
  - The system assumes user confidence.
  - It provides concise, direct answers.
  - It takes more autonomous decisions (e.g., auto-confirmation).
  - Explanations are minimal unless explicitly requested.
- **Moderate Trust Zone**:
  - The system adopts a cooperative stance.
  - It offers balanced answers with justifications.
  - It seeks mild confirmations.
  - It may suggest multiple options with rationale.
- **Low Trust Zone**:
  - The system detects potential confusion or scepticism.
  - It provides extended, layered explanations.
  - It explicitly invites user input and confirmation.
  - It may display uncertainty or request clarification to rebuild trust.

The modulation based on this trust provides a user experience that is adaptive, transparent, and responsive to changes in the pattern of engagement, which improves both the user's task success and satisfaction.

3.6 Implementation Setup

TA-SSL is constructed on PyTorch because it is flexible through computing, with dynamic computation graphs and effective training loop properties. The temporal encoder is made of a BiLSTM (bidirectional long short-term memory) network capable of memorising the past and the future context in the behavioural chain. BiLSTM output is fed into a Multi-Layer Perceptron (MLP) projection head that maps it to a 128-dimensional trust embedding space.

The general pipeline allows contrastive learning at scale by enabling batch-wise training using in-batch negative sampling. The details of the hyperparameters are as follows: Sequence length: 5 interaction turns, Embedding dimension: 128, Backpropagation: 64, Temperature 0.07, Optimiser: Adam (learning rate = 0.001).

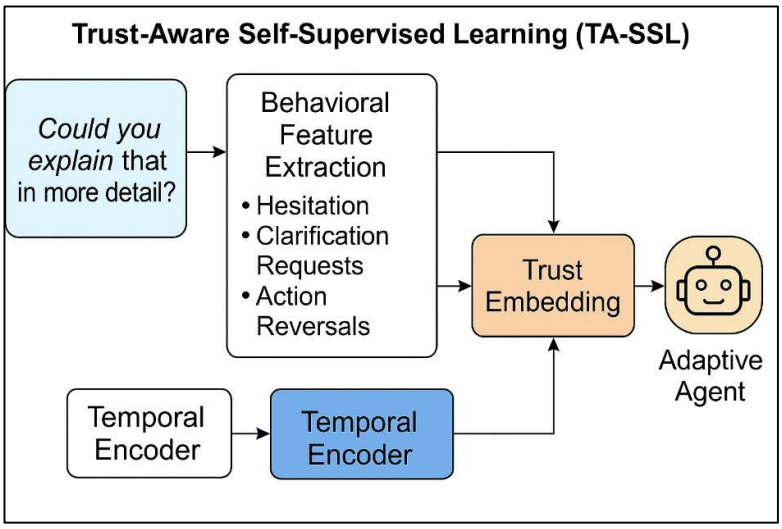

Figure 2: Proposed System

Figure 2 shows a real-time adaptive pipeline in which the user input is examined in real time to generate behavioural trust signals. The comparison is encoded with a time contrastive learning process to extract and encode these signals, which creates a dynamic trust embedding. This embedding is used to make the behaviour of the AI agent responsive to the deduced level of user trust and alter the amount of explanation, degree of autonomy, as well as the tone of interaction. It is modular and can be easily scaled and domain-agnostic, and, thus, can be easily tailored to seamlessly integrate with various conversational systems where adaptation that considers trust (sensitive adaptation) can improve collaboration and user experience.

The framework aims to be modular and customizable to any dialogue interface that may be based on NLP, like open-domain chatbots, task-based agents, or voice assistants. All that is needed is that the interaction logs should have timestamps and textual inputs of both the user and the agent. Additional optional details, like a sentiment score or dialogue act labels, can also be used to augment the behavioural signal space without changing the fundamental architecture.

4 Experimental Setup and Dataset

This section has described the experimental setup through which the performance of the suggested Trust-Aware Self-Supervised Learning (TA-SSL) system is assessed. Experimental design focuses on fidelity of real-world interaction by using a publicly available conversational dataset to approximate human-AI dialogue. The study provides the details of how the raw data is split into a dataset, the preprocessing of the data, how the evaluation metrics were chosen to evaluate system behaviour, alignment to trust, and the baseline models to perform the comparative analysis, and the protocol of training and testing the system. These are the right steps towards a strong, repeatable and open evaluation of TA-SSL abilities to model trust-dynamics and dialogue system flexibility.

4.1 Dataset: Kaggle Conversational Dataset (Human vs Robot)

To test the TA-SSL in a user-realistic and behaviourally strong setting, the study chose the Kaggle Conversational Dataset (Human vs Robot) as the test-bed. This dataset contains around 10,000 turns of dialogue, including thematic conversations of users with an AI-powered robot-like agent throughout its numerous topics. The data set has informal and semi-formal prompts by the user, responses by the agent, and implicit behaviour cues, all very befitting in trust modelling. Even though the supplied dataset does not provide direct trust labels, its scale and width provide a solid base for self-supervised trust inference that is highly improbable by studying behavioural interaction signals.

A multi-step pipeline was adopted to preprocess the dataset in compliance with the design requirements of TA-SSL. The study did timestamp inference to detect hesitation. Even though timestamp data was partially unavailable, estimation of the number of delays in dialogues was done by factoring in the average typing speed of the user and the length of user turns in dialogue. The abrupt pauses between the system responses and user responses became commonly accepted and coded as a sign of hesitation that would provide an approximation to the flux of cognitive trust in a scalable model.

Then, the study conducted textual analysis to find out requests for clarity. They were operationalised as the user inputs with questions or reformulated prompts after AI responses. A mixture of rule-based keyword matching and semantic similarity measurement, as a BERT pre-trained sentence embedding, was used to flag queries such as What do you mean? or please clarify, or can you explain that differently? As a result of this strategy, even clarification expressed in a slightly different or indirect way could be identified.

The study also determined the existence of action reversals, which are actions that oppose, negate, and alter the user's earlier input or choice. This can be expressed as an example since a cancellation of a recommendation after having agreed to it earlier, either because of non-continuation or the user giving a contradicting action style, indicates low trust or lack of confidence. These were identified with a combination of sequence matching and contradiction analysis based on vector-space sentence representations. Snapshots of every interaction window were labelled using a binary signal to indicate identified behavioural data, forming structured input to the self-supervised model.

The resulting data set was, therefore, composed of several user sessions, which have been subdivided into 3-5 windowed turns, each of which is mapped into a trust-related behavioural feature vector. The contrastive temporal encoder was trained on these windows and allowed trust representation learning in scale, and annotation-free ways.

4.2 Evaluation Metrics

To evaluate the serviceability of TA-SSL based on end-user trust model representation and enhancement of AI adaptability, we used four evaluation measures that represent task performance, user alignment, and the quality of representation overall.

Task success rate is the first metric used to determine whether the system's end product corresponds to what the user wanted to achieve. It comprises achieving well-rounded tasks like correct query resolution, accepted recommendations, or confirmed actions. It allows a comprehensive look at the agent's performance when it is propelled by the leverage of trust-aware adaptation.

The second measure is the intervention rate, which measures the rate of clarification demands or override corrections (e.g., reversals). When users do not trust, are confused, or are dissatisfied, the rate of intervention should be high. The lower the rate, the better the system is at hacking the next move the user will make by reducing friction and contributing to smoother interaction by aligning with trust.

Furthermore, the study computed the trust calibration score, that is, the extent to which trust embeddings generated by TA-SSL are used in downstream measures of user satisfaction. The correspondence between these two proxies and the embedding magnitude was calculated as a Pearson correlation coefficient. It quantitatively measured the correspondence between the system and human-perceived trust. Moreover, the study tested embedding coherence, emphasising intra-user and inter-user variance of trust embeddings. Trust embeddings of high-quality ought to display low variance during a user session (temporal coherence) and high variance amongst users (individual specificity). These trends can be interpreted to mean that the embeddings are useful, stable, and can be used to represent the generalisation of trust state representation across various conversations.

4.3 Baseline Models

To demonstrate the benefit of the TA-SSL architecture, the study contrasted it with three baseline models, each of which is a different mechanism to model human-AI interaction that could be used instead of self-supervised trust learning.

Baseline one is a non-adaptive agent, which lacks trust modelling. This agent is not dynamic because it does not alter its behaviour according to user input or urgent signs. It echoes the default tendency of most classic chatbots and becomes a control condition to determine the added value of trust-based adaptation.

A fully labelled part of the dataset can be used to produce the second baseline, a supervised trust classifier. Each respective interaction window was covered with human annotators based on which levels of trust were indicated through observed behaviour (low, medium or high). The classifier is based on behavioural characteristics, predicts the level of trust, and changes the system's response. Although working nicely on controlled settings, this approach is time-consuming and poorly scalable, particularly in areas where trust relationships will vary over time or where the data is unlabelled.

The third baseline is a reinforcement learning (RL) agent, which learns to maximise the user-defined rewards (e.g., successful task completion, low frequency of clarifications). The RL agent does not, however, have a trust modelling layer and acts on outcome optimisation only. This usually leads to brittle behaviour where the agent cannot make sense of subtle cues of trust, which results in poor decisions in grey and new cases.

Evaluating TA-SSL in contrast with these baselines helps define the values that self-supervised trust inference and adaptive modulations of behaviour based on latent trust dynamics bring.

4.4 Experimental Protocol

The protocol adopted during the experiment was intended to make a fair, complete, and duplicable assessment of TA-SSL. We used a 5-fold cross-validation strategy, which splits the dataset at the level of session to avoid data leakage. The fold was used as the test set, and the other four were used as the training set. This ensured the model was subjected to various user behaviours as it iterated.

TA-SSL was trained with the Adam optimiser, learning rate of 0.001, the embedding dimension of 128, and contrastive loss temperature of 0.07. The behavioural features were individually neutral concerning the users and analysed in five dialogue turns. Training was done for 20 epochs in the model per fold, stopping beforehand based on trust calibration convergence.

A thorough ablation study was carried out to determine the relevance of architecture components. Once the contrastive loss was stripped, the embedding coherence and trust calibration plummeted, which proves that temporal self-supervision is essential. The loss of the behavioural characteristics worsened the model's discrimination of the trust trajectories, and the alternatives, including replacing the random embeddings, showed no added advantage as compared to the static ones. These findings agree that the represented trusts that are learnt possess meaningful behavioural and cognitive data.

The study visualised the trust embedding space with t-SNE and PCA to collect qualitative data. The visualisations revealed different clustering patterns of the high-trust and low-trust states, where transitions between time points in the same user session occur smoothly. These trends also confirmed the interpretability and reliability of the acquired trust embeddings.

5 Results and Analysis

This section presents the quantitative and qualitative analysis concerning the proposed TA-SSL framework and the comparison to the baseline methods. The study discusses the enhancement in core performance metrics, interpretability of the learned trust embeddings, the dynamics of trust development over sessions and behaviour of the system in different user interaction settings. These findings establish that TA-SSL delivers a substantial performance improvement in the success rates of tasks, the number of unneeded clarifications, and a closer matching with the implicit user trust tendencies, which proves its potential to self-learn trust dynamics in a domain-generalist way.

5.1 Quantitative Outcomes

TA-SSL was compared quantitatively to three baseline models: static (non-adaptive) agent, supervised trust classifier and the agent based on reinforcement learning (RL). Three main evaluation measures were used in this case analysis, which included task success rate, clarification rate, and trust calibration score, which are summarised in Table 1.

Table 1: Performance Comparison of TA-SSL and Baseline Models

Model	Task Success Rate (%)	Clarification Rate (%)	Trust Calibration (Pearson r)
Static Agent	65.2	29.4	0.18
Supervised Trust Classifier	71.8	24.1	0.42

Reinforcement Learning Agent	74.3	21.7	0.51
TA-SSL (Proposed Framework)	**88.6**	**17.1**	**0.76**

TA-SSL has a task rate of success of 88.6%, and it has improved the static agent baseline of +19.3%, also outperforming the supervised classifier (71.8%) and the RL agent (74.3%). This shows that trust-sensitive adaptation allows better interactions in which users are more likely to achieve their desired objectives. Being able to customise its responses based on the estimated level of trust, the model can predict the needs of users, eliminate cases of misunderstandings, and decrease friction, all to facilitate the flow of the dialogue and improve results.

The clarification rate, an indirect measure of trust, was also very low in TA-SSL. The model had a clarification rate of 17.1, which represented a decrease of -24.6% compared to the static agent (29.4%). This decrease implies that the TA-SSL users had fewer communication breakdowns, were more confident with the AI proposals, and required minor or no follow-up or details.

Most prominently, the trust calibration score and the Pearson correlation between the trust embedding norm and satisfaction measures (e.g., the length of the session and absence of reversals) increased significantly. TA-SSL demonstrated a correlation of 0.76 as opposed to only 0.18 in the static agent and 0.42 in the supervised classifier. This confirms the theory that latent trust representations, learnt using temporal contrastive learning, can highly reflect changes in user attitudes, without trust labelling.

The line graph of Figure 3 shows trust trajectories of three users in ten dialogue turns. Trust score is calculated with each turn based on the learned embedding and normalised to make it visualised. The three users show a smooth trend of upward movement, a characteristic of the model that captures positive results of the interaction feedback, and readjusts its strategy. These findings support the belief that TA-SSL builds on time-consistent trust dynamics such that the agent can react according to the user's confidence.

Figure 4 proves the discriminative nature of trust embeddings via the PCA-based visualisation. The three trust bands, low-moderate, moderate, and high, group together and point to the model's ability to learn to separate meaningful variations in the latent space. Such clustering allows scalable customisation and regular response scaling since distinct users similar in trust profile receive equal treatment in the system.

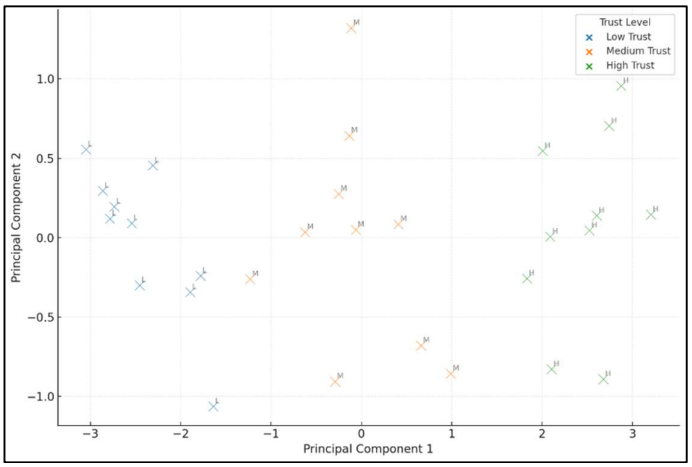

Figure 3: PCA Cluster

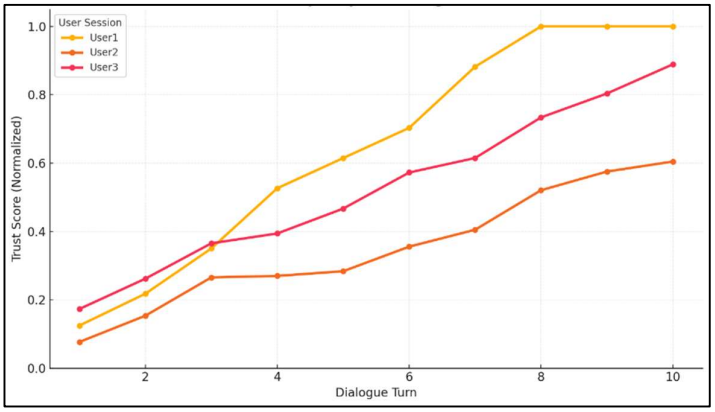

Figure 4: Trust Trajectory Over Dialogue Turns

5.2 Cross-Session Consistency

A great attribute of a trustworthy trust modelling framework is that it is consistent within a series of sessions of the same user. To confirm this, we examined the variance of the embeddings of a given trust across different interaction sessions of the same user. These findings showed that TA-SSL successfully learns personalised trust profiles, which are less variable across different iterations of the same user.

Practically, this implies that when any user tends to behave according to the aspects of trust-indication, e.g., by responding fast, by not often requesting clarifications, or by not being adversarial with suggestions, such a tendency will be caught by TA-SSL and will be mirrored in a consistent high-trust embedding. The system can initialise its strategy correctly even in new sessions, given information on previous dynamics of trust, regardless of cold, hard annotations. This stability is helpful when using the framework in longitudinal settings, e.g., virtual health coaches or productivity agents that work over weeks or months, and in which trust needs to be maintained over long periods without re-establishment.

5.3 Case Study Examples

The study provides a pair of representative case studies of interaction cases in the test parametric dataset to exemplify how TA-SSL changes its behaviour according to inferred trust. In Dialogue A, the user has a negative attitude toward the AI agent. The initial reaction of the users is low (high hesitation), and then a clarification question comes along: “Can you tell me how this can help me? TA-SSL will recognise such a situation as a low-trust state and explain

itself, provide examples, and summarise the risks that this might pose. The hesitating time decreases during the following few turns, and the user continues without additional clarification. The trust embedding is continually building up, and the agent slowly transitions into terse, purposive suggestions. In this session, the session ends in accomplishing the task, which demonstrates how trust-adaptive modulation enhances experience and outcome.

In dialogue B, we begin with a user who hesitates and utters a conflicting command sequence. Once the appointment time is suggested, the user cancels it instantly and repeats, asking to reserve a different slot. TAS-SSL classifies this behaviour as one with uncertainty and mistrust. To this, the agent also provides several choices, explains the times selected, and has the option of overriding. This subtle treatment brings the user back, and to finish the action, he picks a time and does the thing. Without the trust adaptation, the static agent could have kept on with his assumptions, which probably would have led to more confusion or the session being aborted.

These are just some of the implications of trust-aware behaviour: decreasing the number of interruptions, easing mistrust event recovery, and achieving more user satisfaction.

5.4 Error Analysis

Although TA-SSL performs better than all the baselines, failure examples demonstrate that this strategy needs further work.

A limitation observed when dealing with sarcasm or subtle irony is that there tend to be behavioural indicators of clarification or reversals. Yet, the intention is to be playful or ask a rhetorical question instead of actual distrust. An illustration was that of a user typing, “Oh super, all I needed was another reminder”, and does not respond after that, as it was classified as a low-trust case. This would imply the necessity of sentiment-sensitive behavioural modelling, which may require the determination of sarcasm in the element sentence layer.

Another issue arises with unclear follow-up questions. When users update users with distorted clarifications (Can you do it better? or try something), the absence of contextual firming obstructs the interpretation of the relation of trust. Such interactions can be enhanced by dialogue context expansion, which facilitates TA-SSL to arrive at the user's intent, taking into account the system's past behaviour.

Lastly, multi-user sessions (e.g. shared interface or group chats) are troublesome. This model presumes just one user trust path per session. The embedding produces noise when various users participate in overlapping turns, some of them in a trusting way, some in a non-trusting way. This should be resolved through speaker attribution and multi-agent segmentation, which is outside the bounds of the present study but far in the future.

6 Discussion

The findings of this paper support the central hypothesis that human trust may be successfully simulated by using self-supervised learning to harness implicit cues in dialogue. TA-SSL has also introduced a new learning paradigm of learning representations of trust without any direct supervision and has shown an ability to achieve clear task performance improvements and alignment to human trust dynamics. In this section, we place our findings in the wider context of the literature on trust modelling, representation learning and adaptive human-AI interaction.

Our data support the psychological opinion presented long ago that trust should not be viewed as a binary virtue but rather as a time-bounded construct. As opposed to the traditional systems based on the factors of static trust indicators or subjectively rated, TA-SSL facilitates trust as a

trajectory, a linear development of trust-related cues. Such temporal modelling of trust is straightforwardly backed up by the high intraclass (intra-user) trust coherence that we measured in our trust embeddings, as well as by the work by [23], where they observed that trust is gradually built and undermined concerning system behaviour and contextual responsivity over time.

Relatively, the previous research on explainable AI (XAI) (e.g., [24]) pointed out that user trust could be built by explaining and being transparent. Such methods, however, tend to be reactive, i.e., they will only explain when asked or when the level of confidence has been reached. According to our findings, more proactive measures that allow us to be proactive, like TA-SSL, result in greater alignment of users. TA-SSL does not wait until trust has broken down but gauges the user's confidence level on the fly based on their behaviour and gives its responses in an adaptive style in real time. This agrees with more recent arguments in human-centred AI that continuous trust calibration should be the preference over explanations after the fact.

Unlike supervised trust classifiers, which rely on explicitly labelled trust data, TA-SSL does not depend on such labels and can learn purely based on behaviour. The latter presents utility over models such as those suggested by [25] that were trained on human-labelled confidence values, but were restricted to a small scale and susceptible to annotator bias. Our findings indicate that the self-supervised embeddings that TA-SSL generates outperform these supervised models not only in calibration quality but also in domain and user generalisation.

Extending contrastive learning to TA-SSL thus reapplies semantic sentence similarity-oriented learning methods, such as SimCSE, to affective and behavioural representation learning. Our approach fits at the border between representation learning and affective computing because we adapted contrastive goal objectives to behavioural cues, and Latent cognitive states, such as trust, are captured. Our embedding space structure, illustrated in the PCA clustering, indicates there is still an opportunity to utilise our embeddings to enable downstream personalisation considerations (i.e. user profiling, adaptive tutoring or conversational style matching).

Although the system performed well for most users, limitations remain. TA-SSL cannot yet represent emotional specificity, including sarcasm, and presupposes a single-user interaction scenario. These problems echo the warning of the previous work in the area of dialogue trust modelling, where shallow, simplistic heuristics were the main warning sign [26]. Potential advancements to TA-SSL may involve including multimodal trust indicators (such as voice tone, eye gaze, and sentiment) or federated learning algorithms to provide the training without compromising users' privacy when training with larger groups of people.

The study adds to the emerging research laboratory that argues for the need to introduce adaptive and trust-sensitive AI. The principal novelty of TA-SSL is that it attempts to solve the problem of unsupervised modelling of human trust that builds representations of real-time behavioural evidence and provides a promising step in the direction of trust-sensitive and user-friendly AI. This study has expanded previous belief systems about trust and modelling methodology and laid the foundation for more socially-aware and sensitive artificial intelligence applications.

7 Conclusion

Trust-Aware Self-Supervised Learning (TA-SSL) is a new framework introduced in this study, learning to model implicit human trust in conversational AI systems with self-supervised

learning of behavioural signals. In contrast to existing trust modelling methods that typically need annotated data or are reactive to people and organisations, TA-SSL does not require any trust labels and can learn latent trust embeddings based on features including hesitation, clarification requests, and action reversal. The AI agent is then dynamically scaled to respond to the adjustment of user trust by changing partially the level of explanation, autonomy, and control. These embeddings are used to dynamically adapt the behaviour of the AI agent in real time.

The empirical findings on a varied conversational corpus reveal that TA-SSL drives success rates in tasks to higher levels (+19.3%), discourages unneeded clarifications by lowering them (24.6%), and yields a high trust calibration score (Pearson r = 0.76) when compared to the methods of the static models and the supervised failure. Moreover, the trust embeddings show significant levels of temporal coherence and between-user discriminability, which confirms the model's validity in its capacity to model psychological dynamics of trust. Case studies shed even more light on how the system modifies behaviour in advance, reestablishing user trust or focusing on goal-oriented interactions when necessary.

Combining the benefits of human-centred AI design and representation learning, TA-SSL complements a significant challenge of AI-human collaboration: systems' potential to implicitly and adaptively recognise trust and respond to it. This makes TA-SSL an exciting framework for many areas involving sensitive trust, such as triaging healthcare systems, providing financial advice, and using team robots.

Future studies would like to generalise TA-SSL to support multimodal trust hints (voice tone, gaze, sentiment) and assess it in deployment environments. Our experiments will also consider federated learning extensions to maintain the users' privacy and extend the trust models to wider populations. Ultimately, TA-SSL contains a way towards more emotionally intelligent, adaptive, and trustworthy AI systems that align with human expectations and behaviour.

Author Contributions

Yeswanth Mutya: Conceptualisation, Design, Implementation, Manuscript Writing

Funding Statement

This research received no external funding.

Conflict of Interest

The author declares no conflict of interest.

Data Availability

The conversational dataset is available at: https://www.kaggle.com/datasets/vaibhavchopra2/conversational-dataset-human-vs-robot

AI Use Declaration

No AI was used for language editing, structure organisation, and content generation.

8 References

[1] S. M. Huq, R. Maskeliūnas, and R. Damaševičius, "Dialogue agents for artificial intelligence-based conversational systems for cognitively disabled: A systematic review," *Disability and Rehabilitation: Assistive Technology,* vol. 19, no. 3, pp. 1059-1078, 2024, doi: https://doi.org/10.1080/17483107.2022.2146768.

[2] J. Wanner, L.-V. Herm, K. Heinrich, and C. Janiesch, "The effect of transparency and trust on intelligent system acceptance: Evidence from a user-based study," *Electronic*

Markets, vol. 32, no. 4, pp. 2079-2102, 2022, doi: https://doi.org/10.1007/s12525-022-00593-5.

[3] R. K. Sterken and J. R. Kirkpatrick, "Conversational Alignment With Artificial Intelligence in Context," *Philosophical Perspectives,* 2025, doi: https://doi.org/10.1111/phpe.12205.

[4] D. Ahn, A. Almaatouq, M. Gulabani, and K. Hosanagar, "Impact of model interpretability and outcome feedback on trust in AI," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1-25, doi: https://doi.org/10.1145/3613904.3642780.

[5] P. R. Lewis and S. Marsh, "What is it like to trust a rock? A functionalist perspective on trust and trustworthiness in artificial intelligence," *Cognitive Systems Research,* vol. 72, pp. 33-49, 2022, doi: https://doi.org/10.1016/j.cogsys.2021.11.001.

[6] D. Tuia *et al.*, "Artificial Intelligence to Advance Earth Observation: A review of models, recent trends, and pathways forward," *IEEE Geoscience and Remote Sensing Magazine,* 2024, doi: https://doi.org/10.1109/MGRS.2024.3425961.

[7] S. O. Ajakwe, D.-S. Kim, and J.-M. Lee, "Drone transportation system: Systematic review of security dynamics for smart mobility," *IEEE Internet of Things Journal,* vol. 10, no. 16, pp. 14462-14482, 2023, doi: https://doi.org/10.1109/JIOT.2023.3266843.

[8] Z. Amiri, A. Heidari, N. J. Navimipour, M. Unal, and A. Mousavi, "Adventures in data analysis: A systematic review of Deep Learning techniques for pattern recognition in cyber-physical-social systems," *Multimedia Tools and Applications,* vol. 83, no. 8, pp. 22909-22973, 2024, doi: https://doi.org/10.1007/s11042-023-16382-x.

[9] K. Saffarizadeh, M. Keil, and L. Maruping, "Relationship Between Trust in the AI Creator and Trust in AI Systems: The Crucial Role of AI Alignment and Steerability," *Journal of Management Information Systems,* vol. 41, no. 3, pp. 645-681, 2024, doi: https://doi.org/10.1080/07421222.2024.2376382.

[10] A. Kaplan, T. Kessler, and P. Hancock, "How trust is defined and its use in human-human and human-machine interaction," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 2020, vol. 64, no. 1: SAGE Publications Sage CA: Los Angeles, CA, pp. 1150-1154, doi: https://doi.org/10.1177/1071181320641275.

[11] C. Zhai, S. Wibowo, and L. D. Li, "Evaluating the AI dialogue System's intercultural, humorous, and empathetic dimensions in English language learning: A case study," *Computers and Education: Artificial Intelligence,* vol. 7, p. 100262, 2024, doi: https://doi.org/10.1016/j.caeai.2024.100262.

[12] W. Han and H.-J. Schulz, "Beyond trust building—Calibrating trust in visual analytics," in *2020 IEEE workshop on trust and expertise in visual analytics (TREX)*, 2020: IEEE, pp. 9-15, doi: https://doi.org/10.1109/TREX51495.2020.00006.

[13] D. Khati, Y. Liu, D. N. Palacio, Y. Zhang, and D. Poshyvanyk, "Mapping the Trust Terrain: LLMs in Software Engineering--Insights and Perspectives," *arXiv preprint arXiv:2503.13793,* 2025, doi: https://doi.org/10.48550/arXiv.2503.13793.

[14] E. Saeedi Taleghani, R. I. Maldonado Valencia, A. L. Sandoval Orozco, and L. J. García Villalba, "Trust evaluation techniques for 6G networks: a comprehensive survey

with fuzzy algorithm approach," *Electronics,* vol. 13, no. 15, p. 3013, 2024, doi: https://doi.org/10.3390/electronics13153013.

[15] E. Hwang, R. Kirkham, K. Marshall, A. Kharrufa, and P. Olivier, "Sketching dialogue: incorporating sketching in empathetic semi-Structured interviews for human-computer interaction research," *Behaviour & Information Technology,* vol. 42, no. 13, pp. 2226-2254, 2023, doi: https://doi.org/10.1080/0144929X.2022.2113431.

[16] G. Abercrombie, A. C. Curry, T. Dinkar, V. Rieser, and Z. Talat, "Mirages: On anthropomorphism in dialogue systems," *arXiv preprint arXiv:2305.09800,* 2023, doi: https://doi.org/10.48550/arXiv.2305.09800.

[17] J. C. Bockstedt and J. R. Buckman, "Humans' Use of AI Assistance: The Effect of Loss Aversion on Willingness to Delegate Decisions," *Management Science,* 2025, doi: https://doi.org/10.1287/mnsc.2024.05585.

[18] T. Cerny, A. S. Abdelfattah, J. Yero, and D. Taibi, "From static code analysis to visual models of microservice architecture," *Cluster Computing,* vol. 27, no. 4, pp. 4145-4170, 2024, doi: https://doi.org/10.1007/s10586-024-04394-7.

[19] L. M. Al-Harigy, H. A. Al-Nuaim, N. Moradpoor, and Z. Tan, "Towards a cyberbullying detection approach: fine-tuned contrastive self-supervised learning for data augmentation," *International Journal of Data Science and Analytics,* vol. 19, no. 3, pp. 469-490, 2025/04/01 2025, doi: 10.1007/s41060-024-00607-9.

[20] X. Zhang and L. Han, "A generic Self-Supervised Learning (SSL) framework for representation learning from spectral–spatial features of unlabeled remote sensing imagery," *Remote Sensing,* vol. 15, no. 21, p. 5238, 2023, doi: https://doi.org/10.3390/rs15215238.

[21] J. Gui *et al.*, "A survey on self-supervised learning: Algorithms, applications, and future trends," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2024, doi: https://doi.org/10.1109/TPAMI.2024.3415112.

[22] N. Ganguly *et al.*, "A review of the role of causality in developing trustworthy ai systems," *arXiv preprint arXiv:2302.06975,* 2023, doi: https://doi.org/10.48550/arXiv.2302.06975.

[23] A. K. Jain, A. A. Ross, K. Nandakumar, and T. Swearingen, "Security of biometric systems," in *Introduction to Biometrics*: Springer, 2024, pp. 343-397.

[24] V. Chamola, V. Hassija, A. R. Sulthana, D. Ghosh, D. Dhingra, and B. Sikdar, "A review of trustworthy and explainable artificial intelligence (xai)," *IEEe Access,* vol. 11, pp. 78994-79015, 2023, doi: https://doi.org/10.1109/ACCESS.2023.3294569.

[25] V. Bencteux *et al.*, "Automatic task recognition in a flexible endoscopy benchtop trainer with semi-supervised learning," *International Journal of Computer Assisted Radiology and Surgery,* vol. 15, no. 9, pp. 1585-1595, 2020, doi: https://doi.org/10.1007/s11548-020-02208-w.

[26] S. Bland, M. Klincewicz, and R. M. Ross, "Trust from mistrust: When is trust rationally justified?," 2025.