TPM Vol. 32, No. S7, 2025
ISSN: 1972-6325
https://www.tpmap.org/

TRUST IN AI SYSTEMS: A SOCIAL-PSYCHOLOGICAL INVESTIGATION OF HUMAN–AI COLLABORATION

ASHLESHA GUPTA[1], ANUPAM MUND[2], SHWETA ROY[3]. PUNEET GARG[4], DINESH KUMAR YADAV[5]

[1]ASSOCIATE PROFESSOR, JCBUST, YMCA, FARIDABAD, HARYANA, INDIA
[2]ASSISTANT PROFESSOR, ITM (SLS) BARODA UNIVERSITY, VADODARA, GUJARAT, INDIA
[3]ASSISTANT PROFESSOR, ABES ENGINEERING COLLEGE, GHAZIABAD, UTTAR PRADESH, INDIA
[4]ASSOCIATE PROFESSOR, KIET GROUP OF INSTITUTIONS, DELHI NCR, GHAZIABAD, INDIA
[5]PROFESSOR, SAITM, GURUGRAM, DELHI NCR, INDIA
EMAIL:gupta_ashlesha@yahoo.co.in[1] , anupmund90@gmail.com[2], sguddr@gmail.com[3], puneetgarg.er@gmail.com[4], dinsoft.yadav@gmail.com[5]

CORRESPONDING AUTHOR: PUNEET GARG

Abstract
Trust is a pivotal element in effective human–AI collaboration, influencing whether people adopt, use, or reject AI systems. This research work presents a comprehensive investigation of *trust in AI systems* from social-psychological and technological perspectives. We examine theoretical foundations of trust, comparing human-to-human trust paradigms with trust in artificial intelligence (AI) agents. Key factors that shape trust in AI are analyzed, including attributes of AI systems (e.g. performance, transparency), human user characteristics (e.g. personality, prior experience), and contextual dynamics (e.g. task risk, team environment). We further explore how trust develops and evolves during human–AI interactions, highlighting phenomena such as *overtrust*, *undertrust*, and *trust calibration over time*. Practical implications are discussed, focusing on strategies to build and maintain appropriate trust through explainable and reliable AI design, user education, and organizational policies. The outcomes of trust – such as user acceptance, reliance on AI recommendations, satisfaction, and collaboration effectiveness – are synthesized from recent research findings. Finally, we outline current challenges (like measuring trust and addressing cultural differences) and future research directions to foster *trustworthy AI* and optimize human–AI teamwork. Our findings underscore that calibrated trust is essential to harness the full potential of AI while safeguarding human agency and collaboration efficacy.
Keywords: Trust in AI; Human–AI Collaboration; Trustworthy AI; Trust Dynamics; Explainable AI; Human–AI Teaming

1. INTRODUCTION

The rapid integration of artificial intelligence (AI) into various facets of life and work has elevated the importance of trust between humans and AI systems. AI technologies – from machine learning algorithms and recommender systems to advanced *Large Language Models (Llms)* like ChatGPT – now assist or collaborate with humans in decision-making, creative tasks, and critical operations. However, public trust in these AI systems has not kept pace with their growing presence [6][7]. Global surveys indicate that while AI adoption is rising, only about *46% of people are willing to trust AI systems*, reflecting a significant trust gap. This gap represents a major barrier to the successful deployment of AI, as users who lack trust may refuse to use AI recommendations or underutilize AI capabilities. Conversely, excessive or misplaced trust can lead to overreliance on AI, with users following flawed AI advice uncritically, potentially causing errors or safety incidents. The development of calibrated, appropriate trust in AI is therefore critical for realizing effective human–AI collaboration [8][9].

Trust in an AI context can be defined as a user's *willingness to rely on an AI system* to perform a task, given a feeling of vulnerability and expectation of beneficial outcomes. This mirrors classic definitions of interpersonal trust which emphasize accepting vulnerability based on positive expectations of another party's competence and intentions. In human–AI relationships, the core of trust similarly involves perceptions of the AI's *ability*, *integrity*, and *predictability* [10][11]. There are important distinctions, however, between trust in human partners and trust in AI. Unlike humans, AI systems lack emotions and conscious intentions; thus qualities like benevolence or moral integrity—often central to human trust—are less directly applicable. Users primarily base trust in AI on the system's *technical performance, reliability, and helpfulness*, rather than on empathy or honesty. That said, as AI agents become more sophisticated and human-like (for example, AI assistants with conversational abilities or anthropomorphic robots), people may start attributing human characteristics to them. Researchers suggest that trust in AI then becomes a blend of traditional automation trust (focused on functionality) and interpersonal trust (involving social attributions). This convergence heightens the need for a nuanced understanding of how psychological factors (e.g. perceived intentionality or social presence) intersect with technical factors in shaping trust [12][13][14].

From a social-psychological perspective, trust in AI can be viewed through the lens of *trustor–trustee dynamics*. The human user (trustor) brings individual dispositions (such as general propensity to trust technology, personality traits, and prior experiences), while the AI system (trustee) has design attributes (such as transparency, accuracy, and reliability) that affect its trustworthiness. These interactions occur within a broader context – the task, environment, and social setting – which can modulate trust. For instance, working in a high-stakes domain (healthcare, finance, etc.) may require greater proof of reliability to earn trust than a low-stakes entertainment application [15][16]. Likewise, organizational culture and societal norms influence how comfortable people are collaborating with AI. Figure 1 illustrates the interplay between the human trustor, the AI trustee, and the surrounding context in forming trust. Effective human–AI collaboration demands alignment among all three elements so that the user's trust in the AI is well-calibrated to the AI's actual capabilities and the situation's requirements [17][18][19].

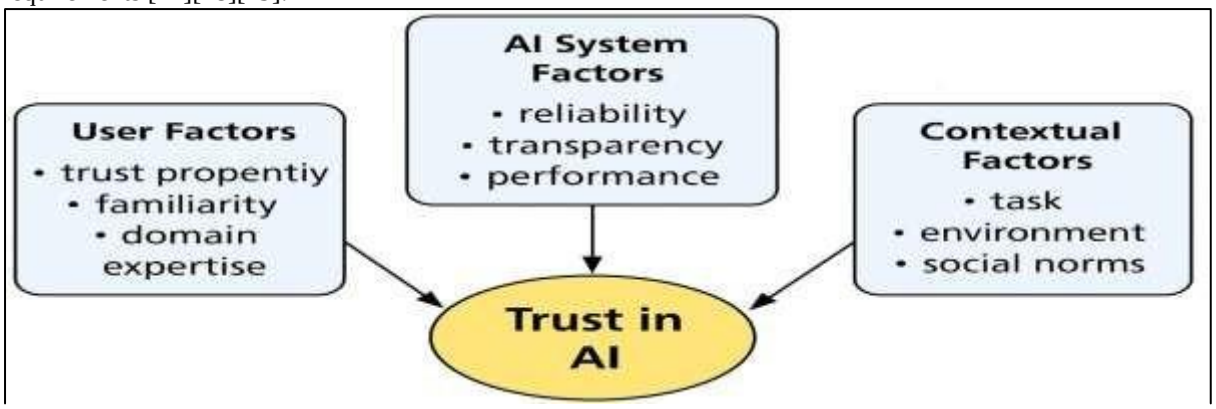

Figure 1: Trust in AI as an interplay between the human user (Trustor),

Figure 1 above highlights that trust in AI emerges from the interaction of user factors, AI system factors, and contextual factors. Misalignment in any of these can lead to distrust or mis-calibrated trust. For example, even a highly reliable AI may not be trusted by a user who has a general skepticism toward technology or who lacks understanding of the AI's workings. Conversely, a user with high trust propensity might overtrust an AI system even when it makes mistakes. The goal is to achieve appropriate trust – neither too low nor too high – corresponding to the AI's true trustworthiness [20].

2. THEORETICAL FOUNDATIONS OF TRUST IN AI

2.1 Defining Trust in Human–AI Relationships

In human–AI collaboration, *trust* can be defined as "the attitude that an AI agent will help a human achieve their goals in a situation characterized by uncertainty and vulnerability". This definition adapts concepts from interpersonal trust to the AI context, emphasizing the trustor's willingness to rely on the AI. Key to this reliance is the expectation that the AI will perform competently and predictably in the interest of the human user [21]. Trust thus involves an *assessment of risk* – the human must believe that potential benefits of using the AI outweigh the risks of errors or misuse. If the perceived risk is too high (for example, if the AI is viewed as unreliable or prone to failure), the user may withhold trust and choose not to use the AI's recommendations. On the other hand, if the AI is deemed trustworthy, the user is more willing to accept its guidance, even though doing so makes them vulnerable to the AI's decisions [22][23].

Traditional trust research in psychology (e.g. in human teams) identifies multiple dimensions of trust, such as *cognitive trust* (based on rational judgments of ability and reliability) and *affective trust* (based on emotional bonds and empathy). In AI systems, trust is predominantly cognitive: users weigh the AI's technical merits – its algorithms, accuracy, consistency – in deciding trust. AI lacks emotions and intentions, so affective and ethical dimensions (like benevolence or integrity) are not inherently present in the AI itself. However, users often anthropomorphize AI agents, perceiving them as having human-like traits especially if the AI communicates in natural language or adopts a persona. In such cases, human trustors might inadvertently apply social and affective criteria to the AI (e.g. judging an AI chatbot as "friendly" or "caring"). This can shape trust in complex ways. Some studies even find scenarios where people *trust AI more than fellow humans* – for example, if AI is seen as more impartial or less likely to act out of self-interest [24][25]. A recent survey in the UK revealed a segment of users who prefer AI's decisions, perceiving AI as unbiased and accurate compared to potentially biased human judgments. Such findings underscore that trust in AI isn't simply lower or higher than trust in humans; it operates on different grounds, with perceived *objectivity* of AI sometimes boosting trust relative to fallible human actors [26].

2.2 Trustor, Trustee, and Context: A Socio-Technical Framework

A useful way to conceptualize trust in AI is through a three-dimension framework consisting of the trustor (human user), the trustee (AI system), and the context of their interaction. Each dimension contributes specific antecedents of trust, as summarized below and in Figure 2 and Table 1. This socio-technical perspective is rooted in both

psychological theory and human–computer interaction research, aligning with frameworks from *human-automation trust* literature and interpersonal trust adapted for AI [27].

- **Trustor (User) Factors:** These are characteristics of the human who is asked to trust the AI. Different individuals bring different baselines of trust (known as *trust propensity* or disposition to trust technology). Personality traits have been linked to trust in AI; for example, people high in openness or conscientiousness tend to report higher trust in AI systems. Technological expertise and prior experience with AI also play a role – a user familiar with an AI tool's functioning may trust it more (or less, if their knowledge makes them aware of its limitations). Self-efficacy in using technology (confidence in one's ability to work with the AI) correlates with higher trust as well. On the flip side, users who are more sensitive about privacy or who fear technology might exhibit lower initial trust. Cognitive biases and cultural background can further influence how a person approaches trusting an AI. For instance, in cultures with high uncertainty avoidance, users might be generally cautious about autonomous systems [28][29].
- **Trustee (AI System) Factors:** These encompass the properties of the AI that affect its trustworthiness as perceived by users. The *performance* of the AI – including its accuracy, reliability, and predictability – is a fundamental determinant of trust. If an AI consistently produces correct and useful outputs (e.g. a recommendation system giving relevant suggestions), users gain confidence in it. *Transparency and explainability* are also crucial: when AI systems provide understandable explanations for their decisions, users' trust increases because the decision-making is less of a "black box". Studies show that explainable AI (XAI) can mitigate distrust and even make users more *resilient* in their trust, meaning they won't abandon the AI after a single error if they understood why it made that error. Other AI attributes affecting trust include *fairness* (ensuring the AI's decisions are unbiased and equitable) and *accountability* (the system's ability to log decisions and enable recourse) – these qualities reassure users that the AI will act in ethically acceptable ways. Anthropomorphic design features (giving AI human-like avatars or personalities) can sometimes increase trust by making the interaction feel more natural and socially present. However, such effects depend on user preferences and cultural context (some users find human-like AI creepy or might hold it to higher standards). *Security and privacy protections* embedded in the AI also factor into trust, as users need to trust the AI with potentially sensitive data. If an AI system is known to preserve user privacy and is robust against cyber threats, it will be perceived as more trustworthy [30][31].
- **Contextual Factors:** The environment and context in which the human–AI interaction takes place significantly influence trust. *Task characteristics* are one aspect – e.g., if the task is safety-critical (like an autonomous driving system or a medical diagnosis aid), users may be naturally more cautious and demand greater proof of reliability before trusting the AI. The *complexity or ambiguity* of the task also matters; when tasks are complex, users might rely more on AI assistance but only if they trust its competence. *Team versus individual setting* is another factor: in multi-member teams where AI is a team "member," trust can be impacted by team dynamics. Recent research indicates that in *two-person teams*, humans tend to trust human partners more than AI partners, whereas in *three-person teams* the trust levels for AI vs. human teammates become more comparable [32][33].

Figure 2 shows the Conceptual model of key antecedents and outcomes of trust in AI collaboration. User factors (e.g. trust propensity, experience), AI system factors (e.g. reliability, transparency), and contextual factors (e.g. task criticality, team environment) feed into the user's **trust in the AI**. Appropriate trust then leads to positive collaboration outcomes such as effective use of AI, better performance, and user satisfaction. (This model is conceptual; arrows indicate influence directions.)

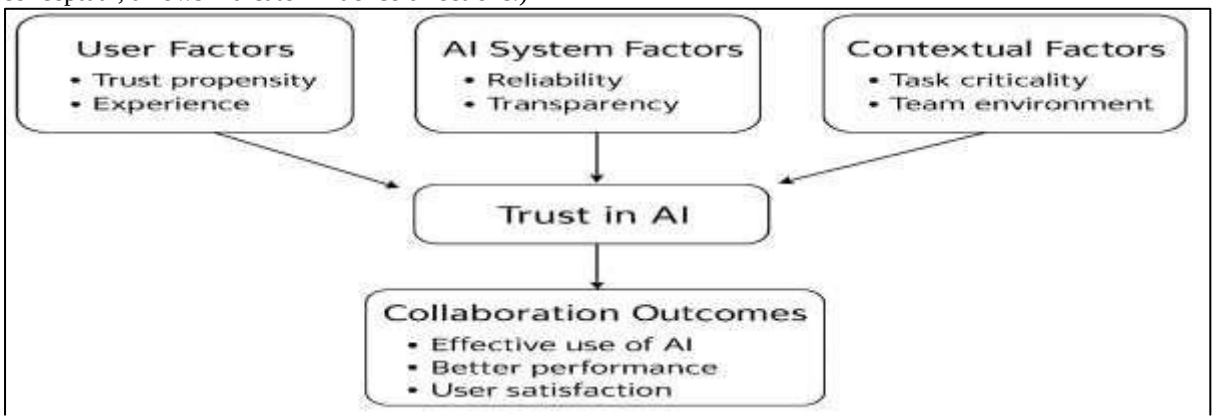

Figure 2: Conceptual model of key antecedents and outcomes of trust in AI collaboration.

Table 1 summarizes these antecedents of trust in AI, categorized by whether they stem from the user, the AI, or the context. Notably, these factors often interact. For instance, an AI's transparency (system factor) might be especially important for a novice user to build trust, whereas an expert user might focus more on the AI's performance. Likewise, contextual risk can amplify or dampen the influence of other factors (e.g., in a high-risk context, even a generally trusting person may be more skeptical of an AI until proven safe) [34][35][36].

Table 1: Key Antecedents of Trust in AI Systems.

Antecedent Category	Description and Examples
User (Trustor) Factors	Inherent characteristics of the human user. Examples: **Trust propensity** (general tendency to trust technology), relevant **personality traits** (e.g. openness, which correlates with higher AI trust), **technology experience** and expertise, **confidence** in using AI (self-efficacy), and prior exposure to or knowledge of the specific AI system. Users with more positive prior experiences or higher tech familiarity often develop trust more easily, whereas those with privacy concerns or low tech literacy may be more skeptical. Cultural background and personal values (like risk tolerance) also fall in this category.
AI System (Trustee) Factors	Attributes of the AI that affect its perceived trustworthiness. Key examples: **Performance competence** (accuracy, reliability, low error rate) – a baseline for earning trust; **Transparency & Explainability** – the AI's ability to explain its decisions or reasoning, which builds understanding and trust; **Fairness and Ethics** – assurance that the AI's decisions are unbiased and align with moral norms; **Security & Privacy** – protection of data and resistance to breaches; **Anthropomorphism** (human-like interface or communication style) – can sometimes increase user comfort and trust by making the AI seem more relatable or socially present, though this depends on user preferences and context.
Context & Interaction Factors	Aspects of the environment and human–AI interaction process. Examples: **Task criticality and risk** – high-stakes decisions demand more trustworthiness for trust to form; **Team or organizational setting** – e.g. working alongside an AI in a team, where trust might be influenced by team dynamics and whether the AI is presented as a collaborator or just a tool; **User interface and interaction quality** – a well-designed, user-friendly AI interface can enhance trust by reducing frustration and making the AI's actions clear; **Social influence** – opinions of peers, leaders, or media about the AI can raise or lower individual trust; **Cultural and regulatory context** – societal attitudes toward AI and the presence of regulations (providing oversight) shape baseline trust. For instance, users in an environment with strong AI governance may feel safer trusting AI.

2.3 Trust in AI vs. Trust in Human Teammates

Given the rising prevalence of AI "agents" working alongside, it is instructive to compare how trust operates in mixed teams versus traditional human-only teams. Research in organizational psychology has long studied **team trust** among humans, finding it to be essential for team cohesion, information sharing, and performance [37]. When an AI joins a team – for example, an AI decision-support system included in a team's deliberation – it changes the trust dynamics. Team members must not only trust each other but also develop trust in the AI's contributions. One question is whether humans trust an AI teammate differently than a human teammate. Recent experimental studies indicate that team size and composition can affect this trust relationship. In a study by Georganta and Ulfert [2], individuals in a two-member team reported higher trust in a human teammate than in an AI teammate, whereas in a three-member team, the trust gap between human and AI team members closed, with AI members being trusted nearly as much as human members [38]. The authors suggest that in a larger team, the AI's contributions might be perceived as more complementary and normalized (especially if two humans can "outvote" or validate the AI), whereas one-on-one, people have less baseline trust in an AI than in another person. Figure 3 illustrates these findings, showing average trust ratings in AI vs. human teammates in different team setups. This chart compares average trust ratings (on a 5-point scale) for human and AI teammates. In two-person teams, humans were trusted more than AI partners; in three-person teams, trust in AI members was similar to trust in human members (based on data adapted from recent research on human–AI teams). Such results highlight that team context influences trust in AI collaborators [39].

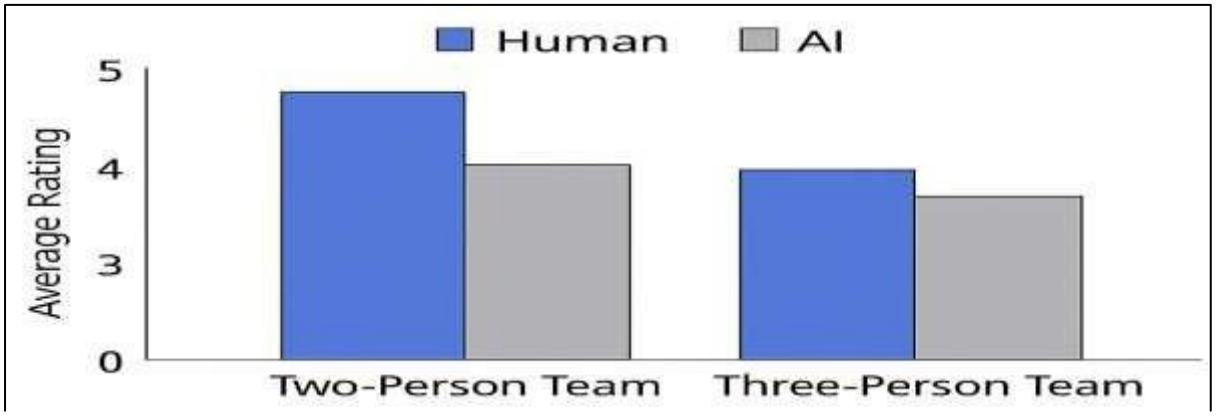

Figure 3: Trust in Human vs. AI Team Members in Different Team Sizes.

One reason human teammates often enjoy higher initial trust is the rich social signals and shared understanding humans naturally have. Humans can communicate intent, show empathy, and build rapport – factors that contribute to *affective trust* and team identity, which AI currently struggles to replicate [40]. Moreover, people may assume a human teammate is accountable to similar ethical and social norms, whereas an AI might be seen as a tool without accountability or as a black-box whose reasoning is opaque. However, as AI systems improve in natural communication and as people gain more experience working with them, these differences can lessen. For instance, if an AI consistently proves its expertise and explains its suggestions, team members may come to trust it much like a respected human expert [41][42].

An interesting phenomenon in human–AI teams is *trust asymmetry*: humans evaluate the trustworthiness of AI, but do AI agents "trust" humans? While AI systems do not possess trust emotions, designers can encode mechanisms for AI to gauge human reliability or to defer to human judgment under certain conditions (a kind of algorithmic trust in the human). In a collaborative setting, optimal performance requires *mutual trust* – humans trusting the AI's capabilities and the AI (or its governing system) trusting human inputs and decisions. For example, in semi-autonomous driving, the car's AI might monitor the human driver's alertness level (a proxy for trustworthiness) to decide when to hand over control. This idea of *reciprocal trust* in human–AI interaction is an emerging research direction. It recognizes that collaboration is a two-way street: not only do humans adjust their trust in AI, but AI systems can also be designed to adapt based on the human's behavior and reliability [43][44][45].

3. Dynamics of Trust in Human–AI Collaboration

Trust is not a static trait; it evolves over time as the human interacts with the AI system. The *dynamics of trust* involve initial trust formation, continuous updating of trust based on outcomes, and potential decay or growth of trust as conditions change. Understanding these dynamics is crucial, as it helps in designing interactions that keep trust calibrated – encouraging users to trust the AI when it is appropriate, but also to remain vigilant and not *overtrust* [46][47].

3.1 Initial Trust Formation and Calibration

When a user first encounters an AI system, they have an *initial trust level* that serves as a baseline. This might be influenced by reputation (what they've heard about the AI), analogous trust (trust in similar systems they've used), or general attitude toward technology. Some AI systems benefit from *institutional trust*: for example, if a reputable hospital deploys an AI diagnostic tool, patients may initially trust it because they trust the institution behind it [48][49]. Initial trust is also shaped by the design of the AI's introduction. An AI that provides clear documentation, demonstrations of its capabilities, or transparency about its limitations can foster a reasonable initial trust. Conversely, if an AI is introduced with hype but little explanation, users might either be skeptical (low initial trust) or have unrealistically high expectations (inflated initial trust) [50].

After initial deployment, *trust calibration* begins as the user experiences the AI's performance. Ideally, trust should increase when the AI proves reliable and decrease when the AI makes errors, tracking the true reliability of the system. However, human psychology doesn't always calibrate trust perfectly. People may exhibit *confirmation bias* – if they start with high trust, they might overlook early mistakes by the AI, staying overtrusting; if they start with distrust, they might downplay the AI's correct actions, remaining undertrusting. Designers sometimes include tutorials or controlled early tasks to help calibrate trust. For instance, a drone interface might initially show the AI's confidence level in controlling the drone, so the human pilot learns when they must intervene [51][52].

Research has introduced techniques like *dynamic trust calibration* to actively manage this process. One approach is for the AI to intentionally adjust its level of autonomy or the information it provides based on the user's trust level. If the system senses the user trusts it too much (e.g. the user is blindly accepting all AI suggestions), it might inject an occasional alert or require user confirmation on critical actions – effectively reminding the user to stay engaged. If the user trusts too little (e.g. frequently overrides a competent AI's suggestions), the AI could provide more convincing evidence or explanations to justify its recommendations [53]. The goal is to avoid *misuse* (overtrust leading to errors) and *disuse* (undertrust leading to ignoring a useful AI). Figure 4 conceptually depicts the trust development loop in human–AI interaction: the user's trust influences how they rely on the AI; this reliance leads to outcomes (successes or failures); those outcomes then feedback to adjust the user's trust moving forward. The cycle illustrates how a user's trust level affects their reliance on the AI, which in turn influences performance outcomes and feedback that update trust. For instance, higher trust leads to greater reliance on AI recommendations; if outcomes are successful, trust may further increase (positive reinforcement), but if the AI performs poorly, trust will decrease. Proper trust calibration seeks a balance where reliance on the AI is commensurate with its capability [54][55].

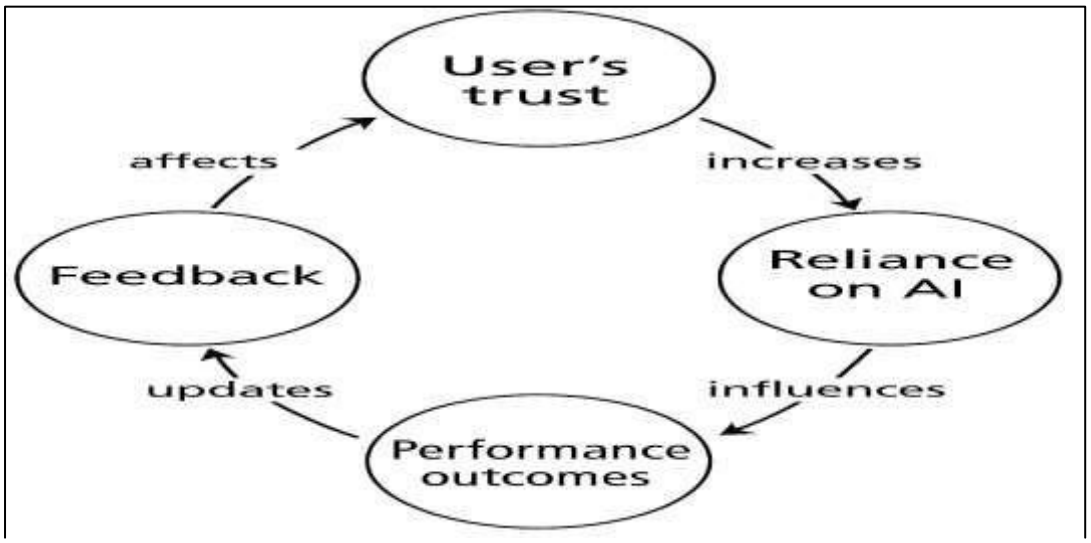

Figure 4: Trust Development Loop in Human–AI Collaboration.

One notable dynamic is that users often exhibit *initial overtrust* in novel AI systems, followed by a correction phase as they encounter imperfections. A recent overview by Schmutz et. al. [4] observed that users tend to overestimate a new AI teammate's capabilities at first, giving it *more trust than warranted*, but over time, as they see its limitations or mistakes, their trust in the AI declines to a more realistic level [56][57]. This pattern was seen in contexts like creative collaboration and decision teams – the novelty and advanced appearance of AI can induce a brief "halo effect." However, without sustained performance, trust drops, and teams sometimes even become less effective than human-only teams due to coordination issues and loss of trust. This underscores that maintaining trust requires consistent AI performance and good user understanding of the AI. It also highlights that the *trajectory* of trust is important: a small early failure by the AI can significantly dampen users' future trust (sometimes disproportionately so – one mistake might outweigh many correct actions in the user's mind). Conversely, if an AI handles an early critical test successfully, it can earn a strong trust credit moving forward [58].

3.2 Overtrust, Undertrust, and Trust Repair

Overtrust occurs when users trust an AI more than its actual performance merits. Overtrust is dangerous because it can lead to complacency. For example, drivers of cars with advanced driver-assistance may become too trusting and stop paying attention, assuming the AI will handle every situation – sometimes with fatal results if the system fails unexpectedly [59]. Overtrust often stems from automation bias, where users assume the AI is always correct. It can be exacerbated by *opaque AI*: if users do not understand how the AI arrives at decisions, they might simply defer to it, especially if it's right most of the time. To combat overtrust, researchers emphasize the importance of keeping the human "in the loop" and designing for *appropriate reliance*. One strategy is *providing continuous feedback* about the AI's confidence and status, so that users have cues when the AI is unsure or encountering novel conditions. Another strategy is training users with scenarios of AI failure, so they appreciate the AI's limits. Some recent work on *explainable AI* suggests that good explanations not only build trust but also prevent *unwarranted* trust by revealing uncertainties: if the AI explains its reasoning and also highlights what it doesn't know, users are less likely to trust it blindly [60][61].

Undertrust is the opposite problem – the user fails to trust a capable AI, thus forgoing its benefits. Undertrust can manifest as the human frequently overriding or ignoring the AI's suggestions, even when the AI is correct. This leads to suboptimal outcomes since the AI's valuable inputs are wasted. Undertrust often arises from initial skepticism, negative first impressions, or previous experiences where an AI errored and lost the user's confidence [62]. It can also occur due to lack of transparency: if users can't tell *why* the AI's recommendation is good, they may err on the side of caution and stick to their own judgment. *Trust repair* techniques are important when trust has been broken (e.g., after an AI malfunction). In human teams, trust repair might involve apologies or compensatory actions. For AI, researchers have explored analogous concepts: an AI could acknowledge an error and provide a detailed analysis of what went wrong and how it's been corrected, to attempt to regain user trust. Another approach is *performance improvement* – the AI needs to perform flawlessly for a period to rebuild confidence, possibly combined with assurances (for instance, updated software that fixes the bug that caused the failure). Empirical studies have found that explaining the cause of a failure can partially restore trust, especially if the cause is understood and addressed, whereas a mysterious failure with no explanation can permanently damage trust in the system [63][64].

A related dynamic concept is *trust resilience* – how robust a user's trust is in the face of errors or conflicting information. Ideally, users should not swing from complete trust to total distrust from a single incident; rather, their trust adjustments should be proportional. Designing AI for *graceful failure* can help here: if an AI can detect its own likely failure and warn or ask for human confirmation, the user's trust might be only mildly reduced instead of completely shattered. For example, a medical AI that flags "I'm not very confident about this case"

allows the human doctor to double-check, preventing a blind error and preserving trust in future cases where the AI is confident.

The *CHAI-T framework* proposed by McGrath et. al. [1] explicitly includes *active trust management* as part of human–AI teaming. In this process framework, the idea is that teams (and system designers) should monitor trust levels throughout the collaboration and apply interventions if trust is mis-calibrated. This could involve interface changes, additional communication between human and AI, or training interventions during the team's lifecycle. The framework also stresses that trust is not an end in itself, but a means to achieve better team performance. Thus, the goal is not to maximize trust blindly, but to find the *optimal* trust level where the human appropriately relies on the AI to maximize outcomes. Figure 5 provides a hypothetical illustration of trust calibration over a series of interactions, highlighting how user trust might rise or fall in response to the AI's performance over time [65].

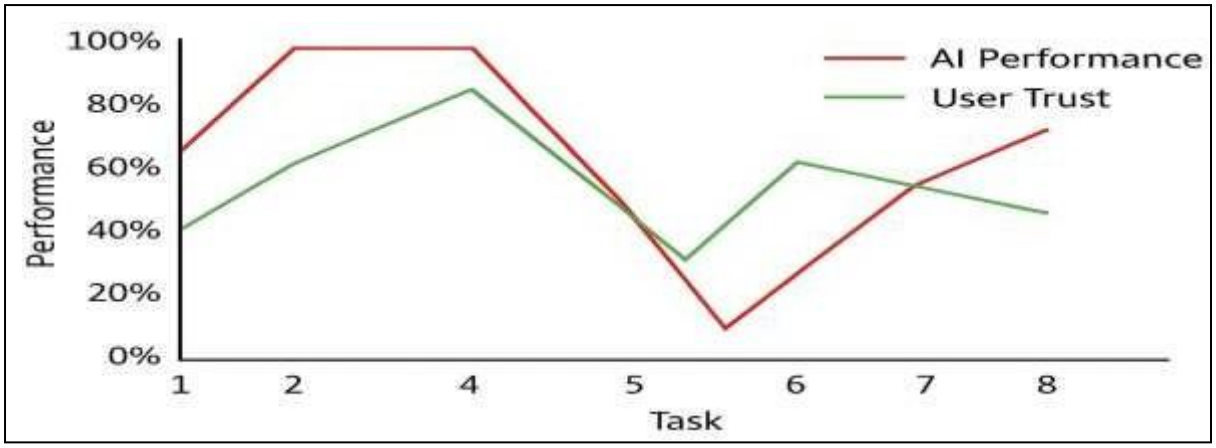

Figure 5: Example of Trust Calibration Over Time.

The plot in Figure 5 shows a hypothetical scenario of an AI system's performance accuracy (red line) across repeated tasks and a user's self-reported trust level in the AI (green line). Trust is updated based on the AI's successes and failures. In this example, the user's trust initially grows as the AI performs well, dips when the AI's performance drops (at tasks 4 and 8), and partially recovers after the AI's performance improves again. Ideally, the trust trajectory (green) should align with the AI's true reliability (red), indicating calibrated trust. Misaligned patterns (e.g. trust remaining high despite poor performance, or vice versa) would suggest overtrust or undertrust [66][67].

3.3 Social and Cognitive Factors in Trust Dynamics

Social-psychological processes heavily influence how trust dynamics play out in human–AI interaction. One such factor is *social identity and team dynamics*. If a human views an AI as part of "the team," they may extend trust more readily. Some organizations personify AI agents (giving them names, avatars, or human-like communication styles) precisely to encourage users to treat them as teammates rather than mere tools. However, studies have found that making an AI too human-like can sometimes backfire if the AI then violates social expectations. For example, an anthropomorphic AI that errs might be judged more harshly ("it misled me") than a plain tool-like AI that errs ("it malfunctioned") [68]. There is evidence that teams with strong *psychological safety* – an environment where human members feel safe to take risks and admit mistakes – can also incorporate AI more effectively, since users are less afraid to critique or question the AI, leading to healthier calibration. A recent systematic review noted that team-level trust in AI is linked to both individual trust perceptions and organizational culture, suggesting a *multi-level trust* phenomenon. In other words, a company that fosters trust in technology and provides a supportive climate can raise the baseline trust its employees have in AI tools, and vice versa [69].

Cognitive factors also play a role. *Human cognitive biases* can distort trust. Automation bias (as mentioned) leads to overtrusting automated systems. The opposite bias is algorithm aversion – some people inherently distrust algorithms especially in subjective domains (e.g. creative judgment), preferring human judgment even if it's statistically worse. *Confirmation bias* might cause users to interpret the AI's behavior in a way that confirms their pre-existing trust level. There's also the **halo** *effect* for AI [70]: if the AI is very good at one task, users might assume it's good at unrelated tasks too. Managing these biases is challenging. Some interface designs attempt to debias users, for instance by periodically highlighting when the human's decision differed from the AI's and what the outcomes were, to force reflection on whether distrust or trust was justified [71].

4. Impacts of Trust on Collaboration and Performance

Why is trust so frequently cited as the "make-or-break" factor in human–AI teamwork? The reason is that trust directly affects *user behavior and decisions* regarding the AI, which in turn determine whether the collaboration fulfills its potential. In this section, we discuss the practical implications of trust: how it influences adoption of AI systems, the usage patterns (reliance vs. override), the overall performance of human–AI teams, and user-related outcomes like satisfaction and acceptance. Table 2 at the end of this section provides a summary of major outcomes associated with trust in AI, based on recent research findings [72][73].

4.1 Adoption, Usage, and Reliance

At the most fundamental level, trust in AI is a key driver of whether people *choose to use* AI systems at all. If an individual or organization doesn't trust an AI application, they are unlikely to adopt it. Numerous studies across domains (from manufacturing to medicine) have identified *lack of trust* as a primary barrier to AI implementation. For instance, medical practitioners often refrain from using AI diagnostic tools unless they trust the system's accuracy and believe it will behave safely and transparently. When trust is present, however, it *significantly increases the intention to adopt and continue using AI*. In technology acceptance terms, trust boosts the perceived usefulness and willingness to depend on the system in the long run [74].

Beyond initial adoption, trust levels govern how users interact with an AI on a day-to-day basis. High trust generally leads to *greater reliance* – users follow the AI's recommendations or delegate tasks to the AI more frequently. This can improve efficiency and outcomes if the AI is competent. For example, a trusted AI assistant in customer support might be allowed to handle many queries autonomously, saving human agents' time and providing faster responses. On the other hand, low trust results in users *double-checking* or overriding the AI often, essentially negating the benefits the AI could provide. One study on AI in hiring showed that when trust was low, hiring managers would ignore the AI's candidate rankings and stick to manual review, whereas those with higher trust integrated the AI's insights into their decisions. Proper trust thus ensures that AI is used where it can add value, and not used when it's not appropriate – a concept sometimes referred to as *appropriate reliance or calibrated reliance* [75][76].

Wen et. al. [3] found that managers with greater trust in an AI decision support system were willing to give the AI more weight in decisions, *but* they also recognized which decision aspects required human intuition, achieving better synergy. Thus, trust enabled an effective division of labor – routine, data-driven evaluations were left to the AI, and unusual or value-sensitive judgments were handled by humans, reflecting the manager's calibrated approach.

From a broader perspective, widespread trust in AI can accelerate the *diffusion of AI innovations*. Societies with higher public trust in AI might see faster uptake of AI-driven services (like autonomous public transport or AI-based medical screening), reaping benefits in productivity and convenience. Conversely, low public trust can stall such initiatives. For example, if only 30% of citizens trust autonomous vehicles, policymakers will face resistance in deploying them widely. A global study in 2025 highlighted this tension: although 66% of people reported using some form of AI regularly, barely half actually trust AI systems, indicating that many use AI with caution and perhaps under duress (e.g. because it's imposed at workplaces). Building trust is therefore seen as essential to unlock AI's full value for society [77][78].

4.2 Human–AI Team Performance and Decision Quality

In collaborative settings, trust markedly influences *team performance* and the quality of decisions made by human–AI pairs or teams. Ideally, a well-calibrated trust leads to what some call *augmented intelligence* – the human and AI together outperform what either could do alone. This has been demonstrated in certain domains: for instance, in radiology, an AI plus a human radiologist (with appropriate trust) can catch more abnormalities than either one separately. However, achieving this synergy is not guaranteed. If trust is lacking, the collaboration may underperform even a single human or AI [79][80]. Indeed, a thought-provoking finding by some researchers is that *human–AI combinations sometimes perform worse than the best human or the best AI alone*, particularly when team processes (like trust and communication) break down. Schmutz et. al. [4] observed that many human–AI teams did not meet expectations because of coordination failures and declining trust – humans either ignored good AI advice or overruled it at wrong times, while at other times they followed AI into errors due to misplaced trust. Shared *situational awareness* and mutual trust (to the extent applicable) were missing, which are elements well-known to be required in high-performing human teams [81].

When trust is optimal, several positive effects on performance are noted in research: faster decision-making (because the human is not second-guessing the AI constantly), higher accuracy (as the human can effectively catch the AI's mistakes and vice versa), and improved *team learning* (the human and AI adapt to each other's patterns). For example, in a study of AI-assisted drone operations, operators who trusted the AI's flight control were able to focus on strategic mission elements, leading to better outcomes than operators who mistrusted the AI and micromanaged the controls (distracting them from bigger issues). On the contrary, when operators overtrusted the drone AI, some responded too slowly to system failures, leading to crashes that a moderately skeptical operator might have averted [82].

4.3 User Attitudes, Satisfaction, and Acceptance

Trust not only affects task performance but also shapes users' *attitudes and overall acceptance* of AI systems. When users trust an AI, they tend to have more positive attitudes toward the technology – viewing it as useful, reliable, and worth integrating into daily processes. This often translates into higher *user satisfaction*. For example, customers interacting with a trusted AI chatbot report greater satisfaction with the service, because trust reduces anxiety and friction in the interaction. They feel the system is on their side and competent, which makes the experience smoother. Empirical evidence confirms a strong link between trust and satisfaction: in a variety of settings (from e-commerce recommenders to virtual assistants), users who reported higher trust also reported greater satisfaction and willingness to continue using the AI. Trust contributes to satisfaction by instilling a sense of *security* and *confidence* – the user is confident the AI will work well, which removes stress. Moreover, trust

can enhance the *perceived reliability* of the system; even if occasional issues occur, a trusting user views them as exceptions against a backdrop of generally reliable service [83][84].

Importantly, trust influences the threshold for *acceptance of AI decisions*. In critical applications, a user often has final say whether to accept or reject the AI's output (e.g., a doctor deciding whether to follow an AI diagnostic suggestion). High trust raises the likelihood of acceptance of the AI's input in the decision process. In domains like judicial risk assessments or hiring, studies have found trust to be a *robust predictor* of whether decision-makers accept AI recommendations, sometimes even more so than the AI's actual accuracy. That is, if the decision-maker doesn't trust the AI, they might reject its correct recommendation; if they trust it, they'll accept its recommendation, sometimes even when it might be wrong (hence the need for calibration). Factors such as *empathy perception* can mediate this: research suggests that if users feel the AI understands nuanced human considerations (like fairness or empathy in hiring), they are more inclined to accept its outputs. Conversely, if the AI is seen as too cold or "unfeeling," a human decision-maker might override it on emotional or ethical grounds even if the AI's logic is sound [85][86].

Trust also fosters *user engagement*. A trusted AI is more likely to be used frequently and explored deeply. Users might try more of the AI's features and engage in a more interactive manner (e.g. asking a digital assistant more follow-up questions, or exploring creative options suggested by an AI). This engagement can lead to better outcomes because the AI can assist more fully [87][88]. For instance, in educational technology, students who trusted an AI tutor engaged with it more and ended up learning more, whereas those who distrusted it only sporadically used it and gained less. In customer service, when customers trust an AI agent, they're more willing to share relevant information about their needs (since they trust the AI will use it properly), enabling the AI to provide better assistance. Trust thus can facilitate *information disclosure* and cooperation. A 2024 study by Amin et. al.[5] demonstrated that users were significantly more willing to disclose personal information to an AI chatbot (for mental health advice) when they had higher trust in the system's benevolence and competence. This points to a virtuous cycle: trust encourages users to input more data and context to the AI, which can improve the AI's performance and personalization, further reinforcing trust [89][90].

Table 2 compiles key outcomes associated with trust in AI, supported by findings from recent research. Broadly, trust is a facilitator: it enables greater usage, better integration of AI into tasks, and more positive user perceptions. However, it must be the *right amount* of trust to truly yield benefits – too little and the AI's power is untapped, too much and users risk errors. The succeeding section will focus on how we can achieve that right balance by design – discussing practical methodologies and design principles to build trustworthy AI and foster well-calibrated trust. Trust (or lack thereof) in an AI system has significant consequences on how users behave and the overall success of human–AI collaboration. This table outlines several key outcomes influenced by user trust, along with brief descriptions [91].

Table 2: Major Outcomes of Trust in AI Systems.

Outcome / Metric	Influence of Trust
Adoption & Acceptance	Strong trust increases the *likelihood of adopting* AI technology and accepting its recommendations. Users are more willing to deploy and rely on an AI tool they trust, whereas distrust can prevent deployment or lead to rejection of AI outputs. High trust has been linked to greater *intention to use* AI across domains – e.g. physicians adopting diagnostic AI, or customers opting for AI services.
Reliance on AI & Usage Level	Users with higher trust tend to *rely on the AI's suggestions or decisions* more often in practice. This can manifest as allowing the AI to automate tasks or frequently following AI recommendations. Appropriate trust leads to **optimal reliance**, where users leverage the AI when it is beneficial. With low trust, users either use the AI minimally or constantly override its recommendations, reducing potential benefits.
Decision Quality & Performance	Properly calibrated trust generally *improves decision outcomes* in human–AI teams. When users trust a competent AI, they combine human judgment with AI input effectively, often achieving higher accuracy or productivity than either alone. If trust is too low or too high, team performance can suffer – undertrust may ignore correct AI solutions, and overtrust may accept incorrect AI outputs, both leading to poorer decisions. Studies have shown trust to be critical for realizing performance gains from AI assistance.
User Satisfaction & Confidence	Users are more *satisfied* with AI-assisted processes when they have trust in the system. Trust reduces anxiety and creates a sense of partnership, improving the user experience. A trusted AI instills confidence – users feel more secure and positive about the outcomes. Conversely, interacting with an untrusted AI often yields frustration, stress, or dissatisfaction. High trust thus correlates with favorable user evaluations and comfort with the AI.

Outcome / Metric	Influence of Trust
Teamwork & Collaboration Quality	In settings where humans and AI work together (e.g. decision-making teams, creative collaboration), trust in AI leads to smoother **collaboration**. The human treats the AI as a reliable team member, facilitating open information exchange and efficient coordination. Low trust results in frictions – the human might constantly double-check the AI or exclude it from critical parts of the process, undermining teamwork. Trust also contributes to a positive "team climate" where the human is receptive to the AI's input.
User Engagement & Information Sharing	When users trust an AI system, they tend to *engage more deeply* with it – exploring its features, providing it with more input, and interacting with it more frequently. For example, a person will converse longer with a trusted AI assistant and volunteer more relevant information (like personal preferences or concerns) to it. This improved information sharing (enabled by trust) can further enhance the AI's effectiveness, creating a virtuous cycle. If trust is low, users keep interactions with the AI to a minimum and withhold information (e.g. a patient not telling a health chatbot certain symptoms due to distrust), which in turn limits the AI's ability to help.

5. Building and Maintaining Trust in AI Systems

Given the importance of appropriate trust for successful human–AI collaboration, a key question is: *How can we design and deploy AI systems in ways that foster well-calibrated user trust?* This section discusses approaches and best practices – both technical and organizational – to build initial trust, maintain it over time, and avoid trust breaches [92]. The focus is on creating AI systems that are not only **trustworthy in design** but also effectively communicate their trustworthiness to users. We divide the discussion into several interrelated strategies: enhancing AI transparency and explainability, ensuring reliability and safety, using human-centered design (including anthropomorphic or interactive elements), and educating or training users for better trust calibration [93].

5.1 Transparency and Explainability

"*Transparency***"** in AI refers to making the system's workings visible or understandable to users, and **"***explainability***"** refers to the AI's ability to provide understandable reasons for its outputs. Decades of research in automation have shown that people trust systems more when they can comprehend how they operate. This remains true for modern AI: one of the most consistently recommended trust-building measures is to integrate explainable AI techniques [94]. For example, an AI image classifier might highlight the sections of an image that led to its decision, giving the user insight into the AI's reasoning. Such explanations demystify the "black box" and allow users to judge the AI's competence. When users see that an AI's reasoning aligns with their own logical analysis, their trust in its conclusions increases. Moreover, if the AI makes a mistake, a good explanation can make it a learning moment rather than a trust-breaking event (the user sees *why* the error happened and can maintain trust that errors will be identifiable and rare) [95].

Recent research supports the impact of explainability on trust. Ha and Kim (2024) found that providing explanations for an AI's recommendations significantly boosted users' trust and their resilience to occasional AI errors. Another study showed that *contextual explanations* (tailored to the user's situation) in an AI loan decision system improved both trust and perceived fairness of the system [96]. On the other hand, *poorly implemented* explanations (e.g. too technical, or obviously irrelevant boilerplate) can backfire, as they might be seen as confusing or even misleading. Thus, explainability must be done in a user-centered way. Guidelines for XAI often suggest using simple language, visual aids (charts or highlights), and providing explanation at the right level of detail for the target user (e.g. a doctor might want a different explanation than a patient) [97].

Transparency goes beyond just output explanations. It also involves being open about the AI's *capabilities and limitations*. For instance, an AI assistant might clearly state: "I have knowledge up to 2022 and may not know recent events," or a medical AI tool might indicate it is not designed to handle pediatric cases if that's a limitation. This up-front transparency sets correct expectations, which is crucial for trust [98][99]. If users know what the AI can and cannot do, they are more likely to trust it within its scope and not to push it beyond, preventing misuses that lead to disappointment. Transparency can also include the AI's confidence levels or uncertainty in its outputs. By exposing uncertainty (e.g. "I am 60% confident in this prediction"), the AI actually *increases* user trust in the long run. It sounds counterintuitive – admitting uncertainty – but it makes the AI appear more honest and allows the user to apply appropriate caution when confidence is low. Many modern AI interfaces now present confidence scores or ranges for this reason [100].

5.2 Reliability, Robustness and Safety

No amount of explanation can make up for an AI that frequently fails. Thus, the foundation of building trust is to ensure the AI system is technically reliable and robust. Trust is closely tied to the AI's *trustworthiness* – if the system behaves consistently well, users naturally become more trusting. Several measures can enhance reliability and safety:

- **Rigorous Testing and Validation:** AI systems, especially those deployed in critical applications, should undergo extensive testing across diverse scenarios to ensure they perform as expected. By catching and fixing bugs or biases before deployment, we reduce the chance of trust-eroding failures in the field. For instance, stress-

testing an AI medical diagnostic on edge cases can improve its reliability and prevent misdiagnoses that would destroy physician trust [101].

- **Fail-safes and Redundancies:** Incorporating safety nets builds trust that even if something goes wrong, the system will handle it gracefully. For example, a robotic assistant might have emergency stop mechanisms or an AI might be programmed to recognize when it's out of its depth and request human intervention. Users trust systems that *know their limits.* As Lee and See (2004) famously noted in automation trust, design should facilitate appropriate trust by making systems *predictable, self-revealing,* and *fail-safe*. In practical terms, an AI could switch to a safe mode or alert the user when anomalies occur [102].
- **Continuous Performance Monitoring:** For long-lived AI services, having monitors that track performance and detect drifts or degradations helps maintain trust. If an AI model's accuracy starts to drop (perhaps due to changing data patterns), the system can either retrain or inform users about decreased confidence until fixed. Users of AI are more likely to trust a system that demonstrates *self-monitoring* and improvement over time, as it shows the AI's maintainers are ensuring ongoing quality [103].
- **Security and Privacy Safeguards:** Trust is not only about accuracy; it's also about whether the AI will do something harmful with user data or be manipulated. Building secure AI systems that protect data and resist adversarial attacks is crucial for user trust, especially in an era of frequent data breaches. For example, if users know that an AI assistant keeps their data encrypted and doesn't share it without consent, they will trust it more with personal information. Clear privacy policies and compliance with regulations (like GDPR) can be communicated to users to bolster trust in how the AI handles their data [104].
- **Consistency and Predictability:** Humans tend to trust systems that behave consistently. If an AI's output varies wildly or seems erratic, trust falters. Ensuring the AI's behavior is predictable (or when it changes, the change is explainable) helps maintain user confidence. Consistency can be improved by smoothing outputs, avoiding random-like behavior, or at least explaining variability. For instance, a finance AI advisor should not give contradictory advice to a user in a short span; if market changes cause different advice, it should explain that context, maintaining an image of logical consistency [105].

5.3 Human-Centered Interface and Communication

The way an AI system communicates with the user is a critical factor in trust. A *Human-Centered Design* that takes into account user needs, mental models, and comfort can significantly enhance trust. Several interface and interaction design considerations have proven effective:

- **Clarity and Consistency in UI:** The user interface should present information in a clear, organized manner. Inconsistent or confusing interfaces can cause user errors and erode trust ("if the interface is sloppy, what does that say about the underlying AI?"). Using familiar design patterns, clear language (avoiding technical jargon), and logical workflows helps users feel in control and understand the AI's actions. For instance, labeling AI-generated content clearly vs. user-provided content can avoid misunderstandings that might otherwise reduce trust if users are unsure who (AI or human) produced what [106.
- **Feedback Mechanisms:** A trust-enhancing interface encourages two-way communication. Users should be able to give feedback to the AI (like correcting it or indicating preferences) and see that the AI adapts or responds to that feedback. When users feel they have *agency* in the interaction, their trust increases because it becomes a collaboration rather than a one-sided automation. For example, a recommender system might allow users to thumbs-up or thumbs-down recommendations; seeing future recommendations change accordingly shows the user that the AI is listening, which builds trust that the system is respectful of their input [107].
- **Anthropomorphic and Social Cues:** Introducing human-like elements in AI interactions can sometimes foster a social form of trust. Polite language, conversational tone, or even a simple avatar can make the AI feel more approachable and trustworthy on an interpersonal level. Research has shown that moderate anthropomorphism (like giving a chatbot a name and a bit of personality) can increase user engagement and trust, as long as the AI's competence also meets expectations. The *CASA (Computers Are Social Actors***)** paradigm suggests people subconsciously apply social rules to computers; hence, an AI that follows social etiquette (e.g. saying thank you, apologizing for errors) may engender more trust. Caution is needed: if the AI is too human-like and then fails, users may feel betrayed in a personal way. The design should align the AI's persona with its actual capabilities [108].
- **Trust Signals and Reassurance:** Sometimes small design elements can reassure users. For instance, displaying certifications or verifications ("This AI's algorithm is approved by FDA" in a medical app, or "Model last updated on ___with X% validated accuracy") can serve as trust signals. Providing access to credentials of data sources ("Trained on 1 million verified cases from XYZ database") also helps. In collaborative scenarios, if the AI can **expose its reasoning process** step by step (perhaps in a side panel), the user can follow along, which makes the process feel more like working with a human partner and builds trust in each step rather than only the final answer.
- **Preventing Emotional Misinterpretation:** A subtle issue is that users sometimes anthropomorphize AI in negative ways too, for example interpreting a factual, terse response from a chatbot as "rude" or "unhelpful," which can hurt trust. Designing the AI's communication style to avoid such impressions is important. If a chatbot injects a bit of empathy ("I'm sorry to hear you're facing this issue, let's see how I can assist"), users often trust it more than if it gives a cold response, even if the end solution is the same. Emotional intelligence in AI responses, to the extent possible, contributes to trust especially in domains like healthcare or counseling [109].

User training and onboarding also belong in human-centered approaches. A well-crafted *Onboarding Tutorial* that shows users how the AI works, including its benefits and limitations, can set a foundation for trust. For example, introducing a fraud-detection AI to analysts might involve showing a few example cases side-by-side where the AI was right and where it was wrong, teaching the analysts what to look out for. This transparency from the get-go creates informed trust: users know when to lean on the AI and when to be cautious, which ironically makes them *more* likely to trust it when appropriate (because they don't fear unknown failure modes).

5.4 User Education and Organizational Practices

Beyond the AI system itself, there are external measures to cultivate a trust-friendly environment. *User education* is one such measure. This can range from basic training on how to use the AI system to more general education on AI's strengths and weaknesses. When users understand how AI algorithms learn and make decisions, they are more likely to trust the system rationally. For instance, if doctors learn that an AI diagnostic tool uses established medical imaging patterns and has been validated in clinical trials with a certain accuracy, their trust grows because they understand the underpinning (versus viewing it as magic). On the public front, improving AI literacy is key – initiatives to educate consumers about AI (how recommendations are generated, what biases can occur) help set realistic expectations and build trust through understanding, addressing the fear of the unknown [110].

Within organizations, *leadership and culture* play a role. If company leadership openly supports the AI, explains why it's adopted, and sets an example of trusting it (while also having appropriate oversight), employees will be more willing to give it a chance. Conversely, if the introduction of AI comes without clear communication, it may breed suspicion (e.g. employees might think the AI is there to monitor or replace them, leading to distrust and even sabotage). Therefore, change management around AI deployment is crucial – involving users in the deployment process, gathering their feedback, and iteratively improving the system helps in *co-development*, making users feel a sense of ownership and trust.

Another practice is *algorithmic accountability*: having clear processes for when the AI makes an error – how it will be addressed, who takes responsibility. When users know there's accountability (e.g. a human supervisor reviews AI decisions periodically, or there's a contact to report issues), they trust the system more because it's not a wild unchecked entity. Some organizations set up AI ethics panels or monitoring teams, which indirectly boosts trust among users who are aware that someone is ensuring the AI remains fair and reliable [111].

Finally, incorporating user feedback loops into ongoing development can maintain trust. As users gain experience, they might identify blind spots or suggest improvements. Organizations that actively listen and update the AI system engender trust that the AI is evolving to meet their needs. It tells users "we care that you trust this system, and we're working to keep it worthy of your trust." For example, a software update addressing a known issue and communicating that to users ("We heard your concern about X, the new version has fixed it…") can recover or reinforce trust.

6. Practical Applications and Broader Implications

Trust considerations in AI systems apply broadly across domains and types of applications. In this section, we discuss how the theoretical principles of trust manifest in various practical applications of AI, and we generalize lessons that cut across specific domains. The aim is to illustrate that while the context may differ – from healthcare to finance to everyday consumer gadgets – the fundamental trust dynamics and the need for calibrated trust remain universal. We also examine how focusing on trust can lead to better AI adoption outcomes and what it means for the future of human–AI collaboration at a societal level.

6.1 Collaborative Decision Support Systems

A common class of AI applications is decision support systems, where AI provides recommendations or insights to a human decision-maker. Examples include AI-powered diagnostic systems in healthcare, financial advisory tools, or even AI assistants for military or emergency response decision-making. In all these cases, trust is the linchpin of utilization. If a doctor doesn't trust the AI's diagnosis suggestion, they will ignore it, nullifying its value. If they trust it appropriately, it can augment their decision – say by catching a rare condition the doctor hadn't considered, thereby improving patient outcomes. But if they overtrust it, they might accept a flawed suggestion without cross-checking, potentially harming the patient [112].

Human resources (HR) is another emerging area – AI is used for screening resumes or even advising on employee retention. Trust issues here involve fairness and biases. HR professionals might distrust AI if they fear it's biased against certain groups. Building trust thus requires transparency in the AI's criteria (to show it's fair) and possibly keeping a human review stage to ensure nothing egregious happens. When trust is established, AI can speed up hiring significantly, but companies often still keep a "human in the loop" to maintain a level of oversight that reassures both the HR staff and the candidates that the process is accountable.

Across these decision support contexts, a few general lessons appear:

- **Start with AI in an advisory role**, not an authoritative role, until trust is earned. People prefer to have AI as a recommendation agent initially, with themselves having final say. As trust grows, they may start deferring more to the AI's decisions in routine cases.
- **Domain knowledge and trust**: People with more domain expertise might be initially skeptical of AI (feeling it encroaches on their expertise), but if shown that the AI is a tool that follows domain rules, they warm up. Meanwhile, novices might overtrust AI because they assume it's infallible. Training both groups differently

is crucial – experts might need to see the AI's alignment with their knowledge (to trust it), whereas novices might need cautionary training to not overtrust just because "the computer said so."

- **Audit trails**: In sensitive decisions (like finance or law), trust is enhanced if the AI's decisions are auditable. Knowing that an AI's recommendation can be reviewed and explained later makes human decision-makers more comfortable using it, because it provides accountability. For example, a bank might trust an AI's credit decision if there's a clear record of factors that led to that decision, which can be shown to regulators if needed.

6.2 AI Assistants and Human-Centered AI Tools

A very visible category of AI in everyday life is AI assistants – think of Siri, Alexa, Google Assistant, or AI customer service chatbots. For these, user trust is crucial for continued use. If an assistant repeatedly gives wrong answers or has unclear sources, users lose trust and abandon it. Conversely, if it reliably helps with tasks (like scheduling, information retrieval) and maintains a friendly, helpful demeanor, users integrate it into daily routines, showing **user loyalty** born from trust [113].

One interesting observation is that the *threshold for trust* may be different depending on the application's stakes. Users might readily trust a virtual assistant to set an alarm or play music (low stakes), but not to give medical advice (high stakes) unless it has proven credibility. So, trust calibration often involves context: a single AI system might be trusted for some things and not others. Some voice assistants explicitly state when they can't handle a request ("I'm sorry, I'm not sure about that") – which, while a limitation, actually builds trust because it prevents the AI from confidently giving a wrong answer. Users learn that if the assistant does answer, it's likely correct within its domain, and if it doesn't know, it will admit it. This honesty is critical. A study of user interactions with chatbots found that *admitting uncertainty* and providing sources increased users' trust in the information the chatbot did provide, versus a chatbot that always gave a answer even if it was a guess.

6.3 Organizational and Societal Implications

On a broader level, fostering trust in AI has implications for how organizations structure work and how society at large views AI integration. Organizations that successfully implement AI often create new *roles and processes* around the AI. For instance, some companies have "AI translators" or analysts whose job is to interpret AI outputs for others, bridging the gap and thereby building trust among the broader team. These are roles that recognize not everyone will automatically trust or understand the AI, so a human facilitator can help. Over time, as trust increases, these roles might evolve or be less hands-on, but they serve as scaffolding in the interim.

Companies also need to consider *ethical use* because public trust can quickly erode if AI is misused. A company could have an incredibly accurate AI, but if users find it invades privacy or is used in a way they find uncomfortable, trust is lost (for example, the backlash to some social media algorithms that were seen as manipulative). Thus, building trust is also about aligning AI use with human values and transparently communicating those values. As one example, when an AI is used in hiring, companies often publicly share how they mitigate bias in that AI. This transparency aims to build trust not just with immediate users (HR staff) but also with stakeholders like job applicants and regulators.

At the societal level, surveys like the KPMG 2025 global study highlight that while people see the *benefits* of AI, they remain wary – demanding regulation and oversight as a condition for their trust. This has led to calls for *policy frameworks* that ensure AI systems are audited and certified for aspects like fairness, safety, and privacy. Government and industry standards (e.g., the EU's proposed AI Act) might, in the future, serve a similar role to FDA approvals in medicine – giving the public a baseline assurance that an AI system meets certain trustworthy criteria. Such measures could raise general public trust in AI technologies, facilitating their acceptance (just as people trust that airplanes are safe largely due to stringent aviation regulations and oversight) [114].

7. Challenges and Future Research Directions

While significant progress has been made in understanding and improving trust in AI systems, many challenges remain. Trust is a nuanced, context-dependent phenomenon, and achieving the "right" level of trust in practice can be difficult. In this section, we outline some key challenges that researchers and practitioners face in fostering trust in AI, and we highlight promising future research directions to address these issues. These include developing better trust metrics, adapting to cultural differences, dealing with increasingly autonomous AI, and ensuring ethical alignment.

7.1 Measuring and Evaluating Trust

One fundamental challenge is how to measure trust in human–AI interaction reliably. In research studies, trust is often measured via user surveys (asking users to rate their trust or perceived trustworthiness of the AI) or by proxy behaviors (like degree of reliance on the AI's suggestions). Each method has limitations. Self-reported trust can be subjective and influenced by users' interpretation of what "trust" means. Behavioral measures (e.g. how quickly a user follows an AI recommendation) might not capture the full picture—someone might trust the AI but still double-check due to protocol, or conversely not trust fully but follow due to lack of alternatives. Developing *robust trust metrics* that can be applied in real-world settings is an ongoing area of research. For instance, using physiological signals (heart rate, eye tracking) to infer user stress or ease during AI interaction is one experimental approach, but linking that definitively to trust is complex [102].

7.2 Cultural and Individual Differences

As identified, trust in AI is not one-size-fits-all. *Cultural differences* can lead to different default trust levels in technology. For example, surveys have found that in some East Asian contexts, people might be more accepting of AI in roles of authority (perhaps due to cultural attitudes towards authority and high technology adoption rates), whereas in some Western contexts, individuals emphasize personal autonomy and might be more skeptical of AI making decisions for them. Designing AI systems that accommodate these differences is important as AI gets deployed globally. Future research should delve deeper into *how cultural values (like individualism vs. collectivism, uncertainty avoidance, power distance) impact trust in AI*. With such knowledge, AI interfaces could potentially adapt – for instance, an AI might present itself differently or offer different explanation styles depending on the user's cultural background or personal preference (some users might want a very detailed explanation to trust it, others may find that tedious and prefer to just see outcomes and develop trust through usage) [108].

7.3 AI Evolution: From Tools to Teammates to Agents

As AI systems become more advanced, possibly attaining greater autonomy or even forms of general intelligence, the trust paradigm will also evolve. One challenge is what some call the **opacity paradox** of advanced AI: as AI (like deep learning networks or large language models) become more powerful, they also often become more complex and harder to interpret, potentially undermining transparency-based trust approaches.

7.4 Ethical and Policy Challenges

Ensuring trust is not misused is also an important future focus. There is a potential dark side: techniques to increase user trust could be exploited to get users to accept AI decisions beyond what they should (like persuasive AI that encourages overtrust to push certain outcomes). It's crucial that trust-building is tied to genuine trustworthiness, not manipulation. Ethical guidelines need to emphasize that any trust calibration must aim for appropriate trust aligned with the user's interests, not simply maximal trust for the AI's or provider's benefit. Future policies might require transparency not just of AI systems, but of their *trust calibration strategies* – for instance, if an AI is using a certain tone or explanation to influence user trust, users might have the right to know that [82].

Another policy issue is *accountability in trust failures*. When a user trusts an AI and that leads to harm, who is responsible? Was it the user's "fault" for trusting too much, or the designer's fault for making the AI appear more capable than it is? These questions will shape regulations. Clear standards might emerge about how AI should communicate uncertainty or limitations, and failing to do so could be seen as negligence on the developer's part, not a user error.

Finally, building public trust at large might involve *certification of AI systems*. We might see independent bodies evaluating AI for criteria that matter to trust – such as fairness, security, and transparency – and giving them trust "scores" or labels (similar to nutrition labels on food or safety ratings on cars). Research can contribute here by identifying which factors most strongly correlate with end-user trust and should thus be part of such evaluations. For example, a label might convey: accuracy level in domain, whether the AI explanation method is human-auditable, bias audit results, etc. A well-informed public could then trust certified AI much like people trust FDA-approved medicine [74].

7.5 Dynamic and Reciprocal Trust Considerations

Future human–AI relationships might involve more *reciprocal trust* elements. For instance, we might have AI systems that selectively trust human input – e.g., a smart home AI might learn to trust one family member's commands over another in certain domains if one has given better feedback historically. There could be interesting emergent behaviors: imagine two AIs trying to gauge each other's reliability when collaborating (like self-driving cars negotiating at an intersection). Developing frameworks for "trust" between AI agents, and how humans fit into those loops, could become relevant (though this stretches the traditional definition of trust) [104].

Furthermore, *long-term trust* in continuous use scenarios will need more attention. Many current studies are short-term; what happens when someone works with the same AI assistant for years? Does trust plateau, or could there be cycles of complacency and shock if a rare error occurs after a long time of perfection? Maintaining vigilance without losing trust in long-term human–AI partnerships (like a person and their AI caregiver over decades) will be an intriguing area. Perhaps periodic re-validation or "refreshers" might be needed – analogous to renewing a certification, an AI might need to re-prove itself or update the user on how it has improved (or changed) over time to sustain trust.

8. CONCLUSION

Trust is central to successful human–AI collaboration, dictating system *adoption* and *utilization*. This review comprehensively investigates trust in AI from social-psychological and technological angles, contrasting human-to-human trust with trust in artificial agents. We analyze key determinants of trust, including AI system attributes like *performance* and *transparency*, user characteristics (e.g., personality and experience), and contextual factors (e.g., task risk). The paper explores the dynamic evolution of trust during human–AI interaction, focusing on the critical process of *trust calibration*—avoiding pitfalls like overtrust and undertrust. Practical implications are discussed, highlighting design strategies for building appropriate trust through explainable and reliable AI, user education, and effective organizational policy. Ultimately, *calibrated trust* is underscored as the vital component needed to harness AI's full potential while preserving human agency and collaboration efficacy, setting the stage for future research challenges.

REFERENCES

[1] McGrath, M. J., Duenser, A., Lacey, J., & Paris, C. (2025). *Collaborative human–AI trust (CHAI-T): A process framework for active management of trust in human–AI collaboration*. Computers in Human Behavior: Artificial Humans, **6**, 100084. (Preprint arXiv:2404.01615)

[2] Georganta, E., & Ulfert, A. S. (2024). *Would you trust an AI team member? Team trust in human–AI teams*. Journal of Occupational and Organizational Psychology, **97**(3), 1212–1241. DOI: 10.1111/joop.12504

[3] Wen, Y., Wang, J., & Chen, X. (2025). *Trust and AI weight: Human–AI collaboration in organizational management decision-making*. Frontiers in Organizational Psychology, **3**, 1419403. DOI: 10.3389/forgp.2025.1419403

[4] Schmutz, J. B., Outland, N., Kerstan, S., Georganta, E., & Ulfert, A.-S. (2024). *AI-teaming: Redefining collaboration in the digital era*. Current Opinion in Psychology, **58**, 101837. DOI: 10.1016/j.copsyc.2024.101837

[5] Amin, A., Shi, W., & Abrams, J. (2024). *Trust and self-disclosure to AI: The case of chatbots in healthcare and caregiving*. ACM Transactions on Computer-Human Interaction, **31**(1), 1–26. DOI: 10.1145/3571783

[6] P. Garg, A. Dixit and P. Sethi, "Ml-fresh: novel routing protocol in opportunistic networks using machine learning," *Computer Systems Science and Engineering*, vol. 40, no.2, pp. 703–717, 2022.

[7] Yadav, P. S., Khan, S., Singh, Y. V., Garg, P., & Singh, R. S. (2022). A Lightweight Deep Learning-Based Approach for Jazz Music Generation in MIDI Format. *Computational Intelligence and Neuroscience*, *2022*.

[8] Soni, E., Nagpal, A., Garg, P., & Pinheiro, P. R. (2022). Assessment of Compressed and Decompressed ECG Databases for Telecardiology Applying a Convolution Neural Network. *Electronics*, *11*(17), 2708.

[9] Pustokhina, I. V., Pustokhin, D. A., Lydia, E. L., Garg, P., Kadian, A., & Shankar, K. (2021). Hyperparameter search based convolution neural network with Bi-LSTM model for intrusion detection system in multimedia big data environment. *Multimedia Tools and Applications*, 1-18.

[10] Khanna, A., Rani, P., Garg, P., Singh, P. K., & Khamparia, A. (2021). An Enhanced Crow Search Inspired Feature Selection Technique for Intrusion Detection Based Wireless Network System. *Wireless Personal Communications*, 1-18.

[11] Garg, P., Dixit, A., Sethi, P., & Pinheiro, P. R. (2020). Impact of node density on the qos parameters of routing protocols in opportunistic networks for smart spaces. *Mobile Information Systems*, *2020*.

[12] Upadhyay, D., Garg, P., Aldossary, S. M., Shafi, J., & Kumar, S. (2023). A Linear Quadratic Regression-Based Synchronised Health Monitoring System (SHMS) for IoT Applications. Electronics, 12(2), 309.

[13] Saini, P., Nagpal, B., Garg, P., & Kumar, S. (2023). CNN-BI-LSTM-CYP: A deep learning approach for sugarcane yield prediction. *Sustainable Energy Technologies and Assessments*, *57*, 103263.

[14] Saini, P., Nagpal, B., Garg, P., & Kumar, S. (2023). Evaluation of Remote Sensing and Meteorological parameters for Yield Prediction of Sugarcane (Saccharumofficinarum L.) Crop. Brazilian Archives of Biology and Technology, 66, e23220781.

[15] Beniwal, S., Saini, U., Garg, P., & Joon, R. K. (2021). Improving performance during camera surveillance by integration of edge detection in IoT system. *International Journal of E-Health and Medical Communications (IJEHMC)*, *12*(5), 84-96.

[16] Garg, P., Dixit, A., & Sethi, P. (2019). Wireless sensor networks: an insight review. *International Journal of Advanced Science and Technology*, *28*(15), 612-627.

[17] Sharma, N., & Garg, P. (2022). Ant colony based optimization model for QoS-Based task scheduling in cloud computing environment. *Measurement: Sensors*, 100531.

[18] Kumar, P., Kumar, R., & Garg, P. (2020). Hybrid Crowd Cloud Routing Protocol For Wireless Sensor Networks.

[19] Raj, G., Verma, A., Dalal, P., Shukla, A. K., & Garg, P. (2023). Performance Comparison of Several LPWAN Technologies for Energy Constrained IOT Network. *International Journal of Intelligent Systems and Applications in Engineering*, *11*(1s), 150-158.

[20] Garg, P., Sharma, N., & Shukla, B. (2023). Predicting the Risk of Cardiovascular Diseases using Machine Learning Techniques. *International Journal of Intelligent Systems and Applications in Engineering*, *11*(2s), 165-173.

[21] Patil, S. C., Mane, D. A., Singh, M., Garg, P., Desai, A. B., & Rawat, D. (2024). Parkinson's Disease Progression Prediction Using Longitudinal Imaging Data and Grey Wolf Optimizer-Based Feature Selection. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(3s), 441-451.

[22] Gudur, A., Pati, P., Garg, P., & Sharma, N. (2024). Radiomics Feature Selection for Lung Cancer Subtyping and Prognosis Prediction: A Comparative Study of Ant Colony Optimization and Simulated Annealing. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(3s), 553-565.

[23] Khan, A. (2024). Optimisation Methods Based on Soft Computing for Improving Power System Stability. *J. Electrical Systems*, *20*(6s), 1051-1058.

[24] Sharma, K. K., Verma, P. K., & Garg, P. (2024). IoT-Enabled Energy Management Systems For Sustainable Energy Storage: Design, Optimization, And Future Directions. *Frontiers in Health Informatics*, *13*(8).

[25] Gupta, S., Yadav, N., Singh, K., & Garg, P. (2025). APPLICATIONS OF SIMULATIONS AND QUEUING THEORY IN SUPERMARKET. *Reliability: Theory & Applications*, *20*(1 (82)), 135-140.
[26] Beniwal, S., Garg, P., Rajpal, R., Sharma, N., & Mittal, H. K. (2025). Fusion of Opportunistic Networks with Machine Learning: Present and Future. *Metallurgical and Materials Engineering*, *31*(1), 311-324.
[27] Garg, P. (2025). Explainable AI & Model Interpretability in Healthcare: Challenges & Future Directions. *EKSPLORIUM-BULETIN PUSAT TEKNOLOGI BAHAN GALIAN NUKLIR*, *46*(1), 104-133.
[28] Rani, P. (2025). From Data to Diagnosis: Unleashing AI and 6G in Modern Medicine. *EKSPLORIUM-BULETIN PUSAT TEKNOLOGI BAHAN GALIAN NUKLIR*, *46*(1), 69-103.
[29] Dixit, A., Garg, P., Sethi, P., & Singh, Y. (2020, April). TVCCCS: Television Viewer's Channel Cost Calculation System On Per Second Usage. In *IOP Conference Series: Materials Science and Engineering* (Vol. 804, No. 1, p. 012046). IOP Publishing.
[30] Sethi, P., Garg, P., Dixit, A., & Singh, Y. (2020, April). Smart number cruncher–a voice based calculator. In *IOP Conference Series: Materials Science and Engineering* (Vol. 804, No. 1, p. 012041). IOP Publishing.
[31] S. Rai, V. Choubey, Suryansh and P. Garg, "A Systematic Review of Encryption and Keylogging for Computer System Security," *2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT)*, 2022, pp. 157-163, doi: 10.1109/CCiCT56684.2022.00039.
[32] L. Saraswat, L. Mohanty, P. Garg and S. Lamba, "Plant Disease Identification Using Plant Images," 2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT), 2022, pp. 79-82, doi: 10.1109/CCiCT56684.2022.00026.
[33] L. Mohanty, L. Saraswat, P. Garg and S. Lamba, "Recommender Systems in E-Commerce," 2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT), 2022, pp. 114-119, doi: 10.1109/CCiCT56684.2022.00032.
[34] C. Maggo and P. Garg, "From linguistic features to their extractions: Understanding the semantics of a concept," 2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT), 2022, pp. 427-431, doi: 10.1109/CCiCT56684.2022.00082.
[35] N. Puri, P. Saggar, A. Kaur and P. Garg, "Application of ensemble Machine Learning models for phishing detection on web networks," 2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT), 2022, pp. 296-303, doi: 10.1109/CCiCT56684.2022.00062.
[36] R. Sharma, S. Gupta and P. Garg, "Model for Predicting Cardiac Health using Deep Learning Classifier," 2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT), 2022, pp. 25-30, doi: 10.1109/CCiCT56684.2022.00017.
[37] Varshney, S. Lamba and P. Garg, "A Comprehensive Survey on Event Analysis Using Deep Learning," 2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT), 2022, pp. 146-150, doi: 10.1109/CCiCT56684.2022.00037.
[38] Dixit, A., Sethi, P., Garg, P., & Pruthi, J. (2022, December). Speech Difficulties and Clarification: A Systematic Review. In *2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART)* (pp. 52-56). IEEE.
[39] Garg, P., Dixit, A., Sethi, P., & Pruthi, J. (2023, December). Strengthening Smart City with Opportunistic Networks: An Insight. In *2023 International Conference on Advanced Computing & Communication Technologies (ICACCTech)* (pp. 700-707). IEEE.
[40] Rana, S., Chaudhary, R., Gupta, M., & Garg, P. (2023, December). Exploring Different Techniques for Emotion Detection Through Face Recognition. In *2023 International Conference on Advanced Computing & Communication Technologies (ICACCTech)* (pp. 779-786). IEEE.
[41] Mittal, K., Srivastava, K., Gupta, M., & Garg, P. (2023, December). Exploration of Different Techniques on Heart Disease Prediction. In *2023 International Conference on Advanced Computing & Communication Technologies (ICACCTech)* (pp. 758-764). IEEE.
[42] Gautam, V. K., Gupta, S., & Garg, P. (2024, March). Automatic Irrigation System using IoT. In *2024 International Conference on Automation and Computation (AUTOCOM)* (pp. 100-103). IEEE.
[43] Ramasamy, L. K., Khan, F., Joghee, S., Dempere, J., & Garg, P. (2024, March). Forecast of Students' Mental Health Combining an Artificial Intelligence Technique and Fuzzy Inference System. In *2024 International Conference on Automation and Computation (AUTOCOM)* (pp. 85-90). IEEE.
[44] Rajput, R., Sukumar, V., Patnaik, P., Garg, P., & Ranjan, M. (2024, March). The Cognitive Analysis for an Approach to Neuroscience. In *2024 International Conference on Automation and Computation (AUTOCOM)* (pp. 524-528). IEEE.
[45] Dixit, A., Sethi, P., Garg, P., Pruthi, J., & Chauhan, R. (2024, July). CNN based lip-reading system for visual input: A review. In *AIP Conference Proceedings* (Vol. 3121, No. 1). AIP Publishing.
[46] Bose, D., Arora, B., Srivastava, A. K., & Garg, P. (2024, May). A Computer Vision Based Framework for Posture Analysis and Performance Prediction in Athletes. In *2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE)* (pp. 942-947). IEEE.
[47] Singh, M., Garg, P., Srivastava, S., & Saggu, A. K. (2024, April). Revolutionizing Arrhythmia Classification: Unleashing the Power of Machine Learning and Data Amplification for Precision Healthcare. In *2024 Sixth International Conference on Computational Intelligence and Communication Technologies (CCICT)* (pp. 516-522). IEEE.

[48] Kumar, R., Das, R., Garg, P., & Pandita, N. (2024, April). Duplicate Node Detection Method for Wireless Sensors. In *2024 Sixth International Conference on Computational Intelligence and Communication Technologies (CCICT)* (pp. 512-515). IEEE.
[49] Bhardwaj, H., Das, R., Garg, P., & Kumar, R. (2024, April). Handwritten Text Recognition Using Deep Learning. In *2024 Sixth International Conference on Computational Intelligence and Communication Technologies (CCICT)* (pp. 506-511). IEEE.
[50] Gill, A., Jain, D., Sharma, J., Kumar, A., & Garg, P. (2024, May). Deep learning approach for facial identification for online transactions. In *2024 International Conference on Emerging Innovations and Advanced Computing (INNOCOMP)* (pp. 715-722). IEEE.
[51] Mittal, H. K., Dalal, P., Garg, P., & Joon, R. (2024, May). Forecasting Pollution Trends: Comparing Linear, Logistic Regression, and Neural Networks. In *2024 International Conference on Emerging Innovations and Advanced Computing (INNOCOMP)* (pp. 411-419). IEEE.
[52] Malik, T., Nandal, V., & Garg, P. (2024, May). Deep Learning-Based Classification of Diabetic Retinopathy: Leveraging the Power of VGG-19. In *2024 International Conference on Emerging Innovations and Advanced Computing (INNOCOMP)* (pp. 645-651). IEEE.
[53] Srivastava, A. K., Verma, I., & Garg, P. (2024, May). Improvements in Recommendation Systems Using Graph Neural Networks. In *2024 International Conference on Emerging Innovations and Advanced Computing (INNOCOMP)* (pp. 668-672). IEEE.
[54] Aggarwal, A., Jain, D., Gupta, A., & Garg, P. (2024, May). Analysis and Prediction of Churn and Retention Rate of Customers in Telecom Industry Using Logistic Regression. In *2024 International Conference on Emerging Innovations and Advanced Computing (INNOCOMP)* (pp. 723-727). IEEE.
[55] Mittal, H. K., Arsalan, M., & Garg, P. (2024, May). A Novel Deep Learning Model for Effective Story Point Estimation in Agile Software Development. In *2024 International Conference on Emerging Innovations and Advanced Computing (INNOCOMP)* (pp. 404-410). IEEE.
[56] Shukla, S. M., Magoo, C., & Garg, P. (2024, November). Comparing Fine Tuned-LMs for Detecting LLM-Generated Text. In *2024 3rd Edition of IEEE Delhi Section Flagship Conference (DELCON)* (pp. 1-8). IEEE.
[57] Kumar, B., IQBAL, M., Parmer, R., Garg, P., Rani, S., & Agrawal, A. (2025, March). The Role of AI in Optimizing Healthcare Appointment Scheduling. In *2025 3rd International Conference on Disruptive Technologies (ICDT)* (pp. 881-887). IEEE.
[58] Kumar, B., Garg, V., Ahmed, K., Garg, P., Choudhary, S., & Baniya, P. (2025, March). Enhancing Healthcare with Blockchain: Innovations in Data Privacy, Security, and Interoperability. In *2025 3rd International Conference on Disruptive Technologies (ICDT)* (pp. 932-938). IEEE.
[59] Raj, V., Prakash, B. K., Kumar, A., & Garg, P. (2024, December). Optimize the Time a Mercedes-Benz Spends on the Test Bench Using Stacking Ensemble Learning. In *2024 International Conference on Progressive Innovations in Intelligent Systems and Data Science (ICPIDS)* (pp. 445-450). IEEE.
[60] Kaushik, N., Kumar, H., Raj, V., & Garg, P. (2024, December). Proactive Fault Prediction in Microservices Applications Using Trace Logs and Monitoring Metrics. In *2024 International Conference on Progressive Innovations in Intelligent Systems and Data Science (ICPIDS)* (pp. 410-415). IEEE.
[61] Kumar, A. A., Sri, C. V., Bohara, K. S. K., Setia, S., & Garg, P. (2024, December). Capnivesh: Financing Platform for Startups. In *2024 International Conference on Progressive Innovations in Intelligent Systems and Data Science (ICPIDS)* (pp. 261-265). IEEE.
[62] Bhandari, P., Setia, S., Kumar, K., & Garg, P. (2024, December). Optimizing Cross-Platform Development with CI/CD and Containerization: A Review. In *2024 International Conference on Progressive Innovations in Intelligent Systems and Data Science (ICPIDS)* (pp. 175-180). IEEE.
[63] Chaudhary, A., & Garg, P. (2014). Detecting and diagnosing a disease by patient monitoring system. *International Journal of Mechanical Engineering And Information Technology*, *2*(6), 493-499.
[64] Malik, K., Raheja, N., & Garg, P. (2011). Enhanced FP-growth algorithm. *International Journal of Computational Engineering and Management*, *12*, 54-56.
[65] Garg, P., Dixit, A., & Sethi, P. (2021, May). Link Prediction Techniques for Opportunistic Networks using Machine Learning. In *Proceedings of the International Conference on Innovative Computing & Communication (ICICC)*.
[66] Garg, P., Dixit, A., & Sethi, P. (2021, April). Opportunistic networks: Protocols, applications & simulation trends. In *Proceedings of the International Conference on Innovative Computing & Communication (ICICC)*.
[67] Garg, P., Dixit, A., & Sethi, P. (2021). Performance comparison of fresh and spray & wait protocol through one simulator. *IT in Industry*, *9*(2).
[68] Malik, M., Singh, Y., Garg, P., & Gupta, S. (2020). Deep Learning in Healthcare system. *International Journal of Grid and Distributed Computing*, *13*(2), 469-468.
[69] Gupta, M., Garg, P., Gupta, S., & Joon, R. (2020). A Novel Approach for Malicious Node Detection in Cluster-Head Gateway Switching Routing in Mobile Ad Hoc Networks. *International Journal of Future Generation Communication and Networking*, *13*(4), 99-111.

[70] Gupta, A., Garg, P., & Sonal, Y. S. (2020). Edge Detection Based 3D Biometric System for Security of Web-Based Payment and Task Management Application. *International Journal of Grid and Distributed Computing*, *13*(1), 2064-2076.
[71] Kumar, P., Kumar, R., & Garg, P. (2020). Hybrid Crowd Cloud Routing Protocol For Wireless Sensor Networks.
[72] Garg, P., & Raman, P. K. Broadcasting Protocol & Routing Characteristics With Wireless ad-hoc networks.
[73] Garg, P., Arora, N., & Malik, T. Capacity Improvement of WI-MAX In presence of Different Codes WI-MAX: Speed & Scope of future.
[74] Garg, P., Saroha, K., & Lochab, R. (2011). Review of wireless sensor networks-architecture and applications. IJCSMS International Journal of Computer Science & Management Studies, 11(01), 2231-5268.
[75] Yadav, S., &Garg, P. Development of a New Secure Algorithm for Encryption and Decryption of Images.
[76] Dixit, A., Sethi, P., & Garg, P. (2022). Rakshak: A Child Identification Software for Recognizing Missing Children Using Machine Learning-Based Speech Clarification. International Journal of Knowledge-Based Organizations (IJKBO), 12(3), 1-15.
[77] Shukla, N., Garg, P., & Singh, M. (2022). MANET Proactive and Reactive Routing Protocols: A Comparison Study. International Journal of Knowledge-Based Organizations (IJKBO), 12(3), 1-14.
[78] Arya, A., Garg, P., Vellanki, S., Latha, M., Khan, M. A., & Chhbra, G. (2024). Optimisation Methods Based on Soft Computing for Improving Power System Stability. *Journal of Electrical Systems*, *20*(6s), 1051-1058.
[79] Garg, P. (2025). Cloud security posture management: Tools and techniques. Technix International Journal for Engineering Research, 12(3).
[80] Tyagi, P., Sharma, S., Srivastava, A., Rajput, N. K., Garg, P., & Kumari, M. (2025). AI in Healthcare:Transforming Medicine with Intelligence. In *First Global Conference on AI Research and Emerging Developments (G-CARED 2025)*, New Delhi, India. https://doi.org/10.63169/GCARED2025.p4
[81] Bhatt, M., Parmar, R., Arsalan, M., & Garg, P. (2025). Generative AI: Evolution, Applications, Challenges And Future Prospects. In *First Global Conference on AI Research and Emerging Developments (G-CARED 2025)*, New Delhi, India. https://doi.org/10.63169/GCARED2025.p6
[82] Saraswat, P., Garg, P., & Siddiqui, Z. (2025). AI & the Indian Stock Market: A Review of Applications in Investment Decision. In *First Global Conference on AI Research and Emerging Developments (G-CARED 2025)*, New Delhi, India. https://doi.org/10.63169/GCARED2025.p10
[83] Sharma, S., Mittal, S., Tevatia, R., Tyagi, V. K., Garg, P., & Kapoor, S. (2025). Unlocking Workforce Potential: AI-Powered Predictive Models for Employee Performance Evaluation. Ind Emerging Developments (G-CARED 2025), New Delhi, India. https://doi.org/10.63169/GCARED2025.p21
[84] Shrivas, N., Kalia, A., Roy, R., Sharma, S., Garg, P., & Agarwal, G. (2025). OSINT: A Double-edged Sword. In *First Global Conference on AI Research and Emerging Developments (G-CARED 2025)*, New Delhi, India. https://doi.org/10.63169/GCARED2025.p22
[85] Aditi, Garg, P., & Roy, B. (2025). A System of Computer Network: Based On Artificial Intelligence. In *First Global Conference on AI Research and Emerging Developments (G-CARED 2025)*, New Delhi, India. https://doi.org/10.63169/GCARED2025.p24
[86] Parmar, R., Kapoor, S., Saifi, S., & Garg, P. (2025). Case Study on Intelligent Factory Systems for Improving Productivity and Capability in Industry 4.0 with Generative AI. In *First Global Conference on AI Research and Emerging Developments (G-CARED 2025)*, New Delhi, India. https://doi.org/10.63169/GCARED2025.p28
[87] Singh, R., Sharma, R., Kumar, R., Nafis, A., Siddiqui, M. A. M., & Garg, P. (2025). Detection of Unauthorize Construction using Machine Learning: A Review. In *First Global Conference on AI Research and Emerging Developments (G-CARED 2025)*, New Delhi, India. https://doi.org/10.63169/GCARED2025.p30
[88] Kapoor, S., Singh, V., Sharma, S., Garg, P., & Ankita (2025). A Bridge between Blockchain and Decentralized Applications Web3 and Non-Web3 Crypto Wallets. In *First Global Conference on AI Research and Emerging Developments (G-CARED 2025)*, New Delhi, India. https://doi.org/10.63169/GCARED2025.p35
[89] Verma, M., Sharma, S., Garg, P., & Singh, A. (2025). The Hidden Dangers of Prototype Pollution: A Comprehensive Detection Framework. In *First Global Conference on AI Research and Emerging Developments (G-CARED 2025)*, New Delhi, India. https://doi.org/10.63169/GCARED2025.p36
[90] Sharma, A., Sharma, S., Garg, P., & Bhardwaj, P. (2025). LockTalk: A Basic Secure Chat Application. In *First Global Conference on AI Research and Emerging Developments (G-CARED 2025)*, New Delhi, India.
[91] Arora, K., Bawane, R., Gupta, C., Ahmed, K., & Garg, P. (2025). Detection and Prevention of Cyber Attack and Threat using AI. In *First Global Conference on AI Research and Emerging Developments (G-CARED 2025)*, New Delhi, India. https://doi.org/10.63169/GCARED2025.p38
[92] Dhruv, Rahman, A. A., Rai, A., Siddiqui, M. A. M., Garg, P., & Yadav, D. (2025). Easeviewer: An Esports Production Tool. In *First Global Conference on AI Research and Emerging Developments (G-CARED 2025)*, New Delhi, India. https://doi.org/10.63169/GCARED2025.p46

[93] Lakshita, Mehwish, Nazia, Ahmed, K., & Garg, P. (2025). Emerging Trend in Computational Technology: Innovations, Applications, and Challenges. In *First Global Conference on AI Research and Emerging Developments (G-CARED 2025)*, New Delhi, India. https://doi.org/10.63169/GCARED2025.p51
[94] Chauhan, S., Singh, M., & Garg, P. (2021). Rapid Forecasting of Pandemic Outbreak Using Machine Learning. *Enabling Healthcare 4.0 for Pandemics: A Roadmap Using AI, Machine Learning, IoT and Cognitive Technologies*, 59-73.
[95] Gupta, S., & Garg, P. (2021). An insight review on multimedia forensics technology. *Cyber Crime and Forensic Computing: Modern Principles, Practices, and Algorithms*, *11*, 27.
[96] Shrivastava, P., Agarwal, P., Sharma, K., & Garg, P. (2021). Data leakage detection in Wi-Fi networks. *Cyber Crime and Forensic Computing: Modern Principles, Practices, and Algorithms*, *11*, 215.
[97] Meenakshi, P. G., & Shrivastava, P. (2021). Machine learning for mobile malware analysis. *Cyber Crime and Forensic Computing: Modern Principles, Practices, and Algorithms*, *11*, 151.
[98] Garg, P., Pranav, S., & Prerna, A. (2021). Green Internet of Things (G-IoT): A Solution for Sustainable Technological Development. In *Green Internet of Things for Smart Cities* (pp. 23-46). CRC Press.
[99] Nanwal, J., Garg, P., Sethi, P., & Dixit, A. (2021). Green IoT and Big Data: Succeeding towards Building Smart Cities. In *Green Internet of Things for Smart Cities* (pp. 83-98). CRC Press.
[100] Gupta, M., Garg, P., & Agarwal, P. (2021). Ant Colony Optimization Technique in Soft Computational Data Research for NP-Hard Problems. In *Artificial Intelligence for a Sustainable Industry 4.0* (pp. 197-211). Springer, Cham.
[101] Magoo, C., & Garg, P. (2021). Machine Learning Adversarial Attacks: A Survey Beyond. *Machine Learning Techniques and Analytics for Cloud Security*, 271-291.
[102] Garg, P., Srivastava, A. K., Anas, A., Gupta, B., & Mishra, C. (2023). Pneumonia Detection Through X-Ray Images Using Convolution Neural Network. In *Advancements in Bio-Medical Image Processing and Authentication in Telemedicine* (pp. 201-218). IGI Global.
[103] Gupta, S., & Garg, P. (2023). 14 Code-based post-quantum cryptographic technique: digital signature. *Quantum-Safe Cryptography Algorithms and Approaches: Impacts of Quantum Computing on Cybersecurity*, 193.
[104] Prakash, A., Avasthi, S., Kumari, P., & Rawat, M. (2023). PuneetGarg 18 Modern healthcare system: unveiling the possibility of quantum computing in medical and biomedical zones. Quantum-Safe Cryptography Algorithms and Approaches: Impacts of Quantum Computing on Cybersecurity, 249.
[105] Gupta, S., & Garg, P. (2024). Mobile Edge Computing for Decentralized Systems. *Decentralized Systems and Distributed Computing*, 75-88.
[106] Gupta, M., Garg, P., & Malik, C. (2024). Ensemble learning-based analysis of perinatal disorders in women. In Artificial Intelligence and Machine Learning for Women's Health Issues (pp. 91-105). Academic Press.
[107] Malik, M., Garg, P., & Malik, C. (2024). Artificial intelligence-based prediction of health risks among women during menopause. *Artificial Intelligence and Machine Learning for Women's Health Issues*, 137-150.
[108] Garg, P. (2024). Prediction of female pregnancy complication using artificial intelligence. In *Artificial Intelligence and Machine Learning for Women's Health Issues* (pp. 17-35). Academic Press.
[109] Pokhrel, L., Arsalan, M., Rani, P., Garg, P., & Pinheiro, P. R. (2026). AI-Powered Healthcare Solutions: Bridging the Medical Gap in Underserved Communities Worldwide. In *Applied AI and Computational Intelligence in Diagnostics and Decision-Making* (pp. 57-86). IGI Global Scientific Publishing.
[110] Kapoor, S., Parmar, R., Sharma, N., Garg, P., & Singh, N. J. (2026). AI and Computational Intelligence in Healthcare: An Introductory Guide. In *Applied AI and Computational Intelligence in Diagnostics and Decision-Making* (pp. 1-26). IGI Global Scientific Publishing.
[111] Pokhrel, L., Kumar, A., Garg, P., Anand, N., & Singh, N. (2026). AI and IoT in Global Health: Ethical Lessons From Pandemic Response. In *Development and Management of Eco-Conscious IoT Medical Devices* (pp. 367-394). IGI Global Scientific Publishing.
[112] Parmar, R., Singh, A., Garg, P., Sharma, T., & Pinheiro, P. R. (2026). Blockchain for Ethical Supply Chains: Transparency in Medical IoT Manufacturing. In *Development and Management of Eco-Conscious IoT Medical Devices* (pp. 337-366). IGI Global Scientific Publishing.
[113] Gupta, S., Garg, P., Agarwal, J., Thakur, H. K., & Yadav, S. P. (2024). Federated learning based intelligent systems to handle issues and challenges in IoVs (Part 1). Bentham Science Publishers. https://doi.org/10.2174/9789815313031124030l
[114] Gupta, S., Garg, P., Agarwal, J., Thakur, H. K., & Yadav, S. P. (2025). Federated learning based intelligent systems to handle issues and challenges in IoVs (Part 2). Bentham Science Publishers. https://doi.org/10.2174/9789815322224125030l