# Schema-Based Understandability in Human-Robot Interactions: A Cognitive Framework

Victoria Williams
RAIL Lab, MIND Institute
School of Computer Science and Applied Mathematics
University of the Witwatersrand
Johannesburg, South Africa
victoria.williams@wits.ac.za

Benjamin Rosman
RAIL Lab, MIND Institute
School of Computer Science and Applied Mathematics
University of the Witwatersrand
Johannesburg, South Africa
Benjamin.Rosman1@wits.ac.za

## ABSTRACT

Robots are entering hospitals, airports, classrooms and homes, yet a persistent barrier to effective human–robot interaction is often not capability itself, but whether people can make sense of robot behaviour quickly enough to coordinate with it. Prevailing work in HRI and explainable AI has largely treated this as an information-disclosure problem: the assumption is that revealing more of a system's internal reasoning will improve understanding. We argue that this framing overlooks how people interpret behaviour under real-time constraints. Drawing on cognitive schema theory, we define robot understandability as the user's ability to form a timely, workable interpretation of what the robot is doing, sufficient to anticipate what it will do next and respond appropriately, without access to its internal decision process. We propose a schema-alignment framework organised around four interdependent schemas that structure social sensemaking: context, role, procedure and strategy. Across case studies in embodied social robotics and large language model (LLM) interaction, we show that coordination breakdowns - pauses, errors, and interactional repairs - arise when system cues fail to support the schemas users rely on. We use LLM interaction as a useful comparison because it removes the demands of embodiment and helps isolate breakdowns that may reflect more general processes of sensemaking. At the same time, the comparison has clear limits: embodied robots introduce additional demands, including coordination in shared physical space and the interpretation of nonverbal cues, such as gaze and gesture. Together, these arguments reframe understandability as a problem of interactional cueing rather than information access, and provide a psychologically grounded basis for robot design.

## CCS CONCEPTS

• Human-centered computing • Human-computer interaction

## KEYWORDS

Schema theory; robot understandability; human-robot interaction; explainability

## 1 Introduction

As robots move from controlled laboratories into shared human environments - triaging patients in emergency departments, guiding travellers through airport terminals, and supporting rehabilitation in clinics - their success increasingly depends not only on what they can do, but on whether the people around them can make sense of what they are doing [1, 2]. We refer to this capacity as "understandability": the user's ability to form a timely, workable interpretation of the robot's current behaviour and to predict what it will do next. At its core, this is a psychological requirement. Interaction with any autonomous agent demands ongoing perception, inference and prediction under conditions of divided attention and time pressure [3]. When a robot's behaviour does not support a clear interpretation, users hesitate and become uncertain about when to intervene. Across repeated encounters, these small judgements - when to rely on the robot and when to step in - shape trust over time. Trust, in this sense, reflects a willingness to rely on the robot, and that willingness is continually updated as users encounter uncertainty and learn from the robot's behaviour [4, 5].

Evidence from motion-legibility research and studies of "automation surprises" points to the same conclusion: successful coordination - that is, choosing the right action at the right time in response to the robot, whether to wait, move around it, follow it, assist it, or intervene - depends on how readily people can interpret behaviour in the moment. Research on "motion legibility" shows that when a robot's movements make its goal easier to infer, observers predict its next action more quickly and with greater confidence [6]. By contrast, in complex automated systems, behaviour that violates expectations ("automation surprises") diverts attention away from the primary task towards figuring out what the system is doing and why, resulting in increased cognitive load and degraded performance [7].

In HRI and explainable AI, understandability has largely been framed as an information problem: if users are given the right information about the system, understanding should follow. The Situation Awareness-based Agent Transparency (SAT)

model [8, 9] reflects this view by organising transparency into three levels: (1) what the system is currently doing, (2) the reasoning behind its actions, and (3) projected future outcomes. Inspired by Endsley's model of situation awareness [10], and applied across military, automotive and healthcare settings, SAT has become one of the most influential frameworks in transparency research, shaping nearly half the experimental studies reviewed in the HRI transparency literature [11].

Many explainable HRI approaches take the same basic approach: they translate the robot's decision-making into cues that are legible to users - through visualisations, verbal justifications, confidence indicators or expressive behaviours - and assume that understanding will follow [12]. Yet, in practice, more transparency does not reliably produce better understanding, and can sometimes make matters worse. A systematic review of controlled experiments across military, automotive, aviation and nuclear domains found that although transparency often helped, its effects were highly context-dependent: outcomes varied by task and agent characteristics, several studies reported no improvement in situation awareness, and some reported increased workload [11]. An earlier review similarly concluded that validation of dominant transparency models remained "largely incomplete" or inconclusive [13]. More recently, a meta-analysis in medical and service robotics found that excessive disclosure can reduce trust and increase cognitive load [14]. Together, these findings point to a limitation of the disclosure paradigm: it treats understanding as a problem of information access, even though users construct understanding through real-time sensemaking - using cues to update expectations about the situation and what the robot will do next.

Cognitive psychology helps explain this mismatch. In real-world settings, attention is distributed across competing demands and working memory is constrained, especially when information must be actively maintained and updated while action is ongoing [15, 16, 17, 18]. Under such conditions, people rarely infer an agent's internal decision-making processes in any explicit or sustained way. Instead, they rely on "schemas": learned, culturally embedded knowledge structures that help people notice what matters, make sense of unclear or missing information, and predict what will happen next [19, 20, 21, 22]. Bartlett's classic experiments showed that when people retell unfamiliar stories, they do not reproduce them verbatim; they reshape them to fit familiar cultural expectations [19]. Understanding and remembering are therefore active, schema-guided processes. Bransford and Johnson [23] showed that contextual framing provided before information dramatically improved comprehension, whereas identical context provided afterwards had little effect - direct evidence that schemas function as cognitive prerequisites for understanding [23]. In social interaction, schemas allow people to infer rapidly: what kind of situation is this, who are the actors to one another, and what normally happens next? When schemas are stable, interaction is smooth and cognitively undemanding. When they fail - because competing schemas are activated or none is adequately established - people experience what Goffman described as "frame breaks": moments in which expectations no longer support prediction, making users stop, reassess the sisutation and work out how to respond [24, 25].

From this perspective, robot understandability is better described as a "schema-alignment problem" than an "information-disclosure problem". A robot is experienced as understandable when its observable cues - movement, timing and communication - make it easy for users to interpret what it is doing and anticipate what it will do next. We organise this framework around four interdependent schemas, drawn from Nishida's Cultural Schema Theory [26, 27]: context (what situation is this?), role (what kind of agent is this?), procedure (what happens next?), and strategy (how should deviations be interpreted?).

From this perspective, understandability improves when robots communicate in ways that (a) stabilise these four schemas early in the interaction, (b) maintain them through consistent behavioural cues as the interaction unfolds, and (c) explicitly repair them when breakdowns occur. In the sections that follow, we elaborate on the theoretical foundations of this framework, ground it in case studies from large language model (LLM) interaction and embodied social robotics, and discuss how it can guide robot design. We draw on LLM interaction as a comparative case because, even in the absence of embodiment, users must still work out what kind of exchange this is, what the system is to them, what should happen next, and what an unexpected response means. LLMs therefore provide a text-only testbed for the same schema-alignment problem. At the same time, the comparison has clear limits. Embodied robots share physical space with users, bringing demands that text-based systems do not: navigating around people, ensuring physical safety, and communicating through gaze, gesture and body positioning. We return to these embodiment-specific issues when we discuss the design implications in Section 4.

## 2 Theoretical Background

Which schemas matter most for understanding a robot? Nishida [26] distinguishes eight "Primary Social Interaction" (PSI) schema types - fact-and-concept, person, self, role, context, procedure, strategy and emotion - learned through repeated cultural experience and organised as an interdependent network, such that activation of one schema reshapes expectations across the others. When people lack the relevant schemas, as in an unfamiliar cultural setting, they often experience uncertainty and social anxiety because they cannot easily interpret what is happening or anticipate what should happen next [26]. A similar problem arises when a robot's behaviour does not fit any familiar interaction script: users lose their footing and cannot readily infer what the robot is doing. In real time, there is little opportunity to reason

through mechanisms. What users need instead is a quick, usable model that supports prediction and action. We therefore focus on four schemas - context, role, procedure and strategy - because they correspond to the minimal questions users must answer in order to coordinate with a robot: What kind of situation is this? What is this agent to me? What happens next? Why is it doing that?

## 2.1 Context Schema

Sensemaking begins with context: a working sense of what kind of situation this is, which goals are relevant, what behaviour is appropriate, and what others are likely to do [19, 20, 23]. People do not usually start with this frame already in place. Rather, they derive it from environmental cues and unfolding interaction [28, 29, 30]. In Turner's [31] computational model, the context schema is activated early and recruits compatible strategies and procedures, thereby narrowing the range of plausible interpretations and supporting rapid prediction. The more specific the context, the more precise the expectations it generates [20, 21].

This has direct implications for HRI. Thompson et al. [32] argue that "social context" is often underspecified in robot interaction, even though environmental cues - location, layout, objects and constraints - strongly shape what users expect and how they interpret a robot's behaviour. When such cues are weak or ambiguous, a stable context schema may fail to form, leaving users to interpret behaviour action by action. Establishing context early - through the physical setting, task framing or the robot's opening cues - is therefore central to understandability [10, 33, 34].

## 2.2 Role Schema

Once the situation is recognised, the next question concerns the social position of the other agent. Role schemas encode expectations associated with particular social positions - what a teacher, nurse or waiter does [26, 35, 36]. They are among the most powerful determinants of social expectation, shaping not only what people anticipate from others, but also what they expect of themselves within the interaction.

Role assignment occurs rapidly, and its effects are far-reaching. The same principle applies to robots. Thompson et al. [32] note that whether a person is a resident, staff member or visitor in a care facility shapes how the robot should interact with them, while the robot's own designated role - whether active collaborator or passive tool - shapes human behaviour in return. Huang and Mutlu [37] similarly identify participant roles and relative status as foundational components of social context. In the present framework, the role schema determines what the robot is to the user - assistant, authority, companion or instrument - and with that come expectations about competence, initiative, deference and the boundaries of appropriate action. When role assignment is ambiguous, users cannot readily predict what the robot will do or what they themselves ought to do, and the interaction loses coherence.

## 2.3 Procedure Schema

A procedure schema captures the expected sequence of an interaction: what normally happens next, who takes which turn, and at what pace. It answers perhaps the most immediate predictive question in social exchange: What is the next step, and what should I do now? In cognitive terms, procedure schemas align with Schank and Abelson's [21] notion of scripts, as well as with the broader distinction between procedural knowledge (knowing how) and declarative knowledge (knowing that) [38, 39]. They are acquired through repeated experience, as successful action sequences become internalised and social or institutional conventions stabilise what "the right next step" looks like in a given setting [26, 31].

When a procedure schema is in place, interaction feels effortless because each participant can prepare their next move in advance - pulling out a boarding pass while approaching a gate agent, for example. When that schema is violated - because a robot skips a step, acts out of sequence or shifts tempo unexpectedly - users can no longer rely on automatic, script-based processing. Instead, they must switch to more effortful, case-by-case interpretation, evaluating each new action on its own terms [40]. The result is not merely momentary confusion; it increases cognitive effort at precisely the point where timely turn-taking and action sequencing matter most.

## 2.4 Strategy Schema

Context, role and procedure schemas support prediction when interaction unfolds as expected. Strategy schemas become especially important when it does not. They help users interpret deviations by connecting an unexpected action to a goal and a constraint - what the robot is trying to achieve, what has changed in the situation, and why an alternative route or response now makes sense. In this way, strategy schemas let users see a departure from the expected script as the robot adapting its plan to the circumstances, rather than a robot making an error [26, 41].

## 2.5 The Schema Network

Nishida's [26] central claim is that PSI schemas form an interdependent network: when one schema shifts, the others recalibrate, altering how behaviour is perceived, predicted and evaluated [26, 42]. Turner's [31] computational account makes this dependency even clearer. Interaction typically begins with context, which frames the situation and narrows the range of plausible goals [20, 23]. That frame recruits a strategy - an interpretive stance for pursuing the goal under current constraints - which in turn organises a procedure by specifying the expected sequence of actions [21, 31]. Role shapes the process throughout by defining what each participant can and should do, and by influencing how deviations are interpreted [26, 35].

For design purposes, the implication is not that robots must reveal their internal decision processes, but that they must

provide cues that keep this schema network stable for the user. When users can quickly establish what kind of interaction this is (context), what the robot is to them (role), what happens next (procedure), and what a deviation means (strategy), coordination remains coherent - even when the robot behaves unexpectedly [26, 28]. When any one of these schemas is weak, ambiguous or undermined by the robot's cues, understanding becomes unstable: users can no longer reliably interpret what the robot is doing, predict what it will do next, or decide how to respond. In Goffman's terms, the frame "breaks": the interaction ceases to feel like a coherent episode, users shift into effortful repair, and disengagement becomes more likely [24, 33].

## 3 Case Studies

In the previous section, we introduced four schemas - context, role, procedure and strategy - and showed how they support real-time interpretation of an agent's behaviour. Here, we use this framework as an analytical tool for diagnosing "coordination breakdowns": moments when user and system actions no longer align, such that users can no longer reliably anticipate what the system will do next or time their own responses appropriately, resulting in mistrust or disengagement. We examine breakdowns in two kinds of systems - LLMs and embodied social robots - and show that, in each case, the failure can be traced to a misalignment between the schemas users rely on to interpret the interaction and the cues the system provides.

The two systems differ in modality, but the comparison serves a specific purpose. It allows us to ask whether coordination breakdowns can be explained by the same underlying mechanism across interaction types. When the same schema–cue mismatch appears in both language-based and embodied systems, this suggests that the problem is not solely modality-specific, but reflects a more general issue of schema alignment - how users form and maintain expectations during interaction. At the same time, embodied robots impose coordination demands that LLMs do not: users must position themselves in space, interpret bodily movement, and read nonverbal signals, such as gaze and gesture in real time. The schema-alignment mechanism may therefore be shared across these interaction contexts, even though design solutions for embodied robots must additionally account for the practical demands of sharing space and coordinating movement in real time.

### 3.1 Context Schema Misalignment

When context is unclear, users cannot reliably activate a context schema - the learned "what situation am I in?" frame that tells them what this encounter is, what role the system is playing, and which norms apply (e.g., do I queue, approach, speak, avoid, or ignore?). In public spaces, people rapidly classify entities using environmental cues like placement, signage, orientation, and approach behaviour. Field studies of public robot deployments show how fragile this is. Tonkin et al. [43] placed a Pepper robot in an airport to provide travel information. Although the system was functionally capable, many passers-by hesitated or disengaged because the scene did not clearly support any single context schema: was this an information kiosk to walk up to, an obstacle to navigate around, or a social agent inviting interaction? When multiple frames compete, users default to caution - stalling, avoiding, or disengaging - because they cannot tell what "the right next move" is in that context.

LLM interfaces reveal a similar context-schema problem. Most chat systems use a single input field for activities as different as search, tutoring, brainstorming, and editing, without reliably signalling which interactional frame is currently in play. Zamfirescu-Pereira et al. [44] found that non-expert users did not primarily struggle with prompt wording; they struggled with a more basic question: what kind of conversation is this? Many approached the system with an instruction-following (Q&A) context schema - ask a question, get an answer - yet the model could shift, without warning, into coaching, planning, or drafting. Kim et al. [45] similarly show that mismatched expectations about what the system is doing and how to use it can drive user dissatisfaction and erode trust. The issue is therefore not only response quality, but the absence of stable framing cues that would let users evaluate whether a response is appropriate to the situation. Context schemas set the interpretive terms for everything that follows; when the interaction type remains ambiguous, users cannot anticipate what should happen next or judge outputs against a clear set of norms.

### 3.2 Role Schema Misalignment

People judge how much to trust a system by reading its cues - what it looks like, how it communicates and how it behaves. These cues lead users to place the system into a role, such as expert adviser, helpful assistant or social companion. Role schema misalignment arises when the system signals a role that carries strong expectations, but cannot reliably meet them. For LLMs, the strongest role cue is often how the robot sounds. Fluent, confident answers readily invite an interpretation of the model as an expert authority, even when its underlying knowledge is uncertain. When the model then produces an obvious failure, such as a hallucinated fact or fabricated citation, users often recalibrate their trust rapidly, rely on it less, and withdraw from the interaction [45, 46].

For embodied robots, role expectations are strongly shaped by anthropomorphic design. Even minimal social cues – a head that turns to "look" at people, or human-proportional body features - can lead users to expect social understanding and responsive interaction [47]. When a highly human-like robot then fails to understand the user, behaves inappropriately, or cannot repair a breakdown, it is typically judged more harshly than a less human-like robot [48]. This difference matters for schema alignment. In LLMs, role expectations are constructed largely from language during the interaction; in embodied

robots, they are set immediately by physical design and presence, before any dialogue begins. As a result, role calibration in embodied settings happens earlier and carries higher stakes for how breakdowns are interpreted.

## 3.3 Procedure Schema Misalignment

Smooth coordination, i.e., keeping the interaction moving without awkward pauses, interruptions, or breakdowns, depends on a procedure schema: a shared script for how the interaction works - how to begin, how turns are taken, what input is expected and what signals that the exchange is complete. In everyday encounters with robots, people often are not sure how the interaction is supposed to work, where to stand, when to speak, and whether the robot is waiting for them to do something [49]. In museum deployments, for example, visitors interrupt, speak over the robot, or walk away when the system fails to it obvious whose turn it is and what should happen next [50]. In contrast, The SPENCER airport guide robot provides a useful counterexample. Coordination improved when SPENCER provided clear procedural cues - explicit "follow me" prompts, a steady and predictable walking pace, and signalling early when it was about to speak or change activity - because these cues allowed users to recognise a familiar "leader–follower" routine [51]. The robot's underlying capabilities did not fundamentally change; what changed was that users could anticipate the next step and align their own behaviour accordingly.

LLMs show the same problem in text-based form. Web search comes with a well-learned procedure - enter a query, scan results, choose a source, refine the query [52] - whereas chat prompting offers no equally shared script for what to do after an imperfect answer. When a response is partial or off-target, users must improvise: rephrase the question, add constraints, request a rewrite or restart entirely, often without clear feedback about which move is most likely to help. Studies of novice prompting describe the same trial-and-error cycle: users iterate repeatedly, investing more effort without reliably moving closer to a solution [44, 53]. The failure, then, is not primarily one of insufficient explanation. It is the absence of procedural guidance - cues that tell the user what the next step should be.

## 3.4 Strategy Schema Misalignment

Deviations are inevitable in real-world interaction. The difficulty arises when a system deviates without giving the user an interpretable reason. In such moments, users are not seeking a technical account of the system's internals; they are trying to answer three practical questions: *what changed, why did it change, and what should I do now?* When the system does not provide cues that support those inferences, the deviation feels arbitrary, and the interaction becomes difficult to repair.

In LLMs, this often appears as refusals or abrupt shifts in tone that users experience as inconsistent. Recent work on refusal behaviour and "over-refusal" shows that LLMs can be highly sensitive to framing, and can refuse in ways that are difficult to predict from the user's perspective - especially when the system does not make it clear whether the issue is a safety-related policy constraint, missing information or a capability limit [54, 55, 56]. When users cannot tell which of these cases they are facing, they cannot choose an effective recovery strategy - for example, adding context, reframing the request or ceasing to rely on the system - and reliance declines [46].

Embodied robots show the same coordination problem when they pause, reroute or refuse a request without indicating why the change occurred or what the user should do next. Recent HRI work on repair and trust recovery suggests that explanations are most useful when they are next-step oriented: they connect the robot's change in behaviour to an intelligible goal and constraint, and indicate what will happen next or what the user should do in response [57, 58, 59].

# 4. Design Implications

Transparency in HRI and explainable AI is often framed as an information-disclosure problem: if users can see more of the system's reasoning, coordination should improve [60, 61, 62]. The breakdowns discussed in Section 3 point to a different problem. In each case, the failure was one of interpretation: users could not sustain a stable understanding of what kind of encounter this was, what role the system occupied, what would happen next, or how to interpret deviations. The issue was not simply that information had been "withheld", but that the system's cues did not support the rapid sensemaking required in real-time interaction [33, 63].

This view is consistent with cognitive and social psychology. Under time pressure, people do not typically infer hidden mechanisms; they rely on schemas - learned structures that guide attention, shape expectations, and support rapid prediction [17, 20, 64]. The design problem, therefore, is less about making computation legible than about stabilising the schemas that allow users to anticipate, evaluate and repair interaction as it unfolds [26]. On this basis, we propose five design principles. Each targets a schema implicated in the case studies and specifies the kinds of cues that help users predict what the system is doing and what will happen next. To make these links explicit, each principle identifies the schema it addresses and connects back to the relevant breakdowns in Section 3.

**Principle 1: Communicate context before content**

**Target schema:** context ("what kind of interaction is this?")

As Section 3.1 showed - a Pepper robot that passers-by could not readily identify as an information assistant, and LLM chat interfaces that silently shifted between interaction modes - understanding depends on a frame, and that frame must come first. Bransford and Johnson [23] showed that people understand and remember the same information very differently when a contextual cue is provided in advance; without it, material is harder to organise and interpret.

Goffman [24] makes the parallel point for social encounters: people rely on "primary frameworks" to determine what kind of situation they are in before they can make sense of behaviour within it. Cognitive accounts help explain why this ordering matters. The brain uses top-down expectations to constrain interpretation early, shaping what people notice and how they assign meaning to incoming cues [65, 66]. If the interaction type is not established at the outset, users are forced to infer it from behaviour already underway - an effortful and error-prone process in real time.

Design implication: The system should establish context before asking anything of the user or providing detailed information. This can be done through a short, user-facing cue that names the interaction type, states the goal, and sets the immediate expectation for what happens next [21]. In public deployments, such early framing increases engagement because it gives people a clear category within which to act [43, 67]. For example: "I'm a wayfinding assistant - tell me your destination." "I help with check-in - scan your boarding pass." "I can triage questions - describe your issue in one sentence." Crucially, context-setting does not require the user to speak first. Embodied robots can activate context schemas before any verbal exchange through physical placement (for example, beside an information desk), visible signage, screen-based labels, approach behaviours such as orienting towards an approaching person, or simple ambient signals such as a status light. This is especially important in public settings, where many users will not initiate contact unless they already understand what the robot is for [43].

**Principle 2: Calibrate role expectations early**

**Target schema:** role ("what is this system to me?")

Section 3.2 showed that role misalignment plays out in much the same way in both systems: LLMs whose fluent language encourages expectations of expert-level accuracy, and embodied robots whose human-like appearance encourages expectations of social competence. In both cases, failure arises when the system cannot live up to the role its own cues have implied. Once context is established, users rapidly infer the system's role - assistant, adviser, companion or tool - and that role inference shapes what they expect the system to do and how much they are willing to rely on it [68, 69]. Early cues shape how users interpret what happens next [70], even small mismatches between implied role and actual capability can destabilise trust.

**Design implication:** At the outset, the system should communicate the capabilities and limitations most relevant to the current interaction. This gives users a basis for judging what to rely on the system for, and how to proceed when it cannot help. For example: "I can help you find your gate, but I cannot access bookings or rebook flights. For changes, please go to the service desk." This supports calibrated reliance by setting an appropriate expectation from the start. A practical question follows: how much should the robot say about what it can and cannot do? Exhaustively listing capabilities is neither realistic nor desirable – the range of possible constraints is effectively open-ended, and experienced users may find lengthy disclosure unnecessary or irritating. Consistent with how schemas operate, the better approach is to be selective and adaptive. The robot should communicate the capabilities and limits most relevant to the immediate task and then reveal additional boundaries only when they become interactionally relevant. A hospital triage robot, for example, might begin with its core function ("I can help assess your symptoms") and only surface a further limit when the user reaches it ("I'm not able to prescribe medication - I'll put you in touch with a doctor"). In this sense, role calibration should be treated as an ongoing process rather than delivered once at the start, because user expectations are shaped and updated over the course of interaction.

**Principle 3: Make the procedural sequence explicit**

**Target schema:** procedure ("what happens next?")

Section 3.3 showed that when no shared script is available, users struggled to proceed: museum visitors interrupted the robot or walked away, and LLM users repeatedly revised prompts without knowing which repair strategy was most likely to work. Even when context and role are clear, interaction will still break down if the user cannot answer a simple question: what happens next? When the sequence is not made explicit, users must guess how to proceed - when to speak, what to provide, and whether they are making progress. This additional burden of "process management" competes with the task itself and increases cognitive load [71]. Procedure schemas, or scripts, reduce that burden by supplying a familiar sequence of steps and making the next expected action explicit [21, 72].

Procedural cues also support grounding by maintaining a shared sense of where the exchange stands and what counts as progress [73]. In robot encounters, missing procedural signals leave users uncertain about where to stand, when to speak, or what the robot is waiting for, thereby producing hesitation and disengagement [49].

Design implication. The system should state the procedure early and succinctly: what will happen first, what it needs from the user, and what will happen next. This may be as simple as a recognisable routine ("follow me") or a short interaction plan ("I'll ask three questions, then I'll recommend an option"). The point is to give users a script they can follow, rather than forcing them to infer the interaction structure from behaviour alone [51, 57, 74].

**Principle 4: Mark deviations and explain them within the user's frame**

**Target schema:** strategy ("why did that just happen?")

Section 3.4 showed that unexplained deviations - an LLM refusing a request without clarifying whether the issue is safety, capability or missing context, or a robot rerouting

without saying why - leave users unable to tell whether the system is adapting or failing. Coordination depends on predictability. People anticipate what an agent will do next and revise their understanding when that expectation is violated [75, 76, 77]. Small deviations are often manageable when the reason is clear. The problem is the unexplained deviation: it interrupts action, diverts attention away from the primary task, and forces the user to work out what is happening [78]. Without a usable reason, users cannot tell whether the system is adapting to a constraint, making an error, or pursuing a different goal. The result is a broader loss of predictability - users no longer know what to expect next - which undermines coordination and reduces trust [41].

Design implication. When the system departs from the expected course, it should do two things: first, signal that a deviation is occurring; second, provide a brief explanation that supports action by identifying the goal, the constraint, and what will happen next. For example: "I'm taking a different route because this area is restricted; please follow me." This restores predictability without requiring users to reason about internal optimisation, and explanations that do not support the user's next action are unlikely to repair the interaction effectively [79]. What, then, counts as a "deviation" in practice? We define it simply: any action the user would not have predicted given the interaction script currently in play. Designers need not label every possible robot action in advance. Deviations are relative to the expectations the system itself has established. If a wayfinding robot initiates a "follow me" routine and then unexpectedly stops or reverses, it has broken the active script and should explain why when the robot behaves as expected, no special cue is needed.

The reverse problem also matters: what happens when the user deviates - walking away mid-interaction, providing unexpected input, or skipping a step? Our framework focuses on how users interpret the robot, but the same coordination logic applies from the robot's side too. To respond effectively, the robot needs an explicit expectation of the interaction sequence - what should happen next - so it can detect when the user has gone off-script and choose an appropriate recovery response (e.g., prompt, clarify, or restart the step). We see this as an important direction for future work as robots develop richer behavioural models.

**Principle 5: Repair when confusion is detected**

**Target schema:** schema network as a whole (restoring the interpretive framework)

Even when cues are well designed, users will sometimes become confused. When this happens, they stop coordinating and start trying to work out what the system is doing [78]. If the system does not help, the confusion can spread: users lose confidence not only in the last action, but in the broader interaction - what this encounter is, what the system can do, and what should happen next [41, 75, 76, 77].

Design implication: Confusion should be treated as a prompt for repair. The system should briefly re-establish the shared frame so that the user can continue: remind them what this interaction is for (context), what the system can and cannot do (role), what happens next (procedure), and, where relevant, why the system changed course (strategy). A key limitation is that detecting confusion in real time remains technically difficult. Current approaches rely on cues, such as facial expressions, hesitation, gaze aversion and speech disfluencies e.g., [80, 81], but these signals are noisy and often unreliable across contexts. This principle therefore specifies what the robot should do once confusion has been detected, while acknowledging that detection itself remains an open engineering challenge. As a practical starting point, designers can use simpler triggers - response timeouts, repeated failed inputs, or the user explicitly asking for help - to initiate repair sequences, and move towards richer detection methods as the technology matures.

## 5 Conclusion

We argue that robot understandability is better conceived as a schema-alignment problem than as an information-disclosure problem: in real-world settings, such as clinics, airports, classrooms and homes, people have limited attention and cannot pause to work out in detail what the robot is doing or why, so they rely on schemas - learned expectations that let them rapidly infer what kind of interaction this is, what the robot is to them, what is likely to happen next, and how to interpret unexpected actions. Across our case studies, a consistent pattern emerged: when a robot's cues support these expectations, coordination remains smooth; when they do not, users hesitate, lose confidence and disengage. This pattern appears in both embodied social robots and text-based language models, suggesting a broader challenge in how people make sense of autonomous systems during interaction, while also reinforcing that embodied robots introduce demands the LLM comparison cannot capture (shared physical space, spatial coordination and nonverbal cueing), so robot design frameworks must reflect these interactional realities rather than assume that findings from text-only systems carry over to embodied robots. The framework we propose now requires empirical testing to examine (i) how schema-aligned cues shape users' ability to predict the system, coordinate with it, and trust it across tasks and settings; (ii) how prior experience, cultural background, and situational context shape the schemas users carry into an interaction; and (iii) how the relative weight of the four schemas varies by context (for example, whether procedural clarity matters more in time-critical medical settings than in open-ended social encounters). This agenda also highlights two pressing technical challenges: reliable real-time detection of user confusion, and adaptive role calibration that updates how the system communicates its capabilities and limits as the interaction progresses." By reframing understandability as expectation alignment rather

than information disclosure, this work offers both a diagnostic lens for explaining why interactions break down and a psychologically grounded basis for designing systems that people can readily interpret and coordinate with.

## ACKNOWLEDGMENTS

VW received funding from the Oppenheimer Memorial Trust Award (OMT Ref: 2150701).

## REFERENCES

[1] Seraj, E., Lee, K.-M., Zaidi, Z., et al. 2024. Interactive and explainable robot learning: A comprehensive review. Found. Trends Robot. 12, 2–3, 75–349.

[2] Kumar, S., et al. 2026. Enhancing robot understandability – A model to estimate varying levels of discrepancy. Int. J. Soc. Robot. DOI: https://doi.org/10.1007/s12369-025-01356-w

[3] Peter A. Hancock, Deborah R. Billings, Kristin E. Schaefer, Jessie Y. C. Chen, Ewart J. de Visser, and Raja Parasuraman. 2011. A Meta-Analysis of Factors Affecting Trust in Human–Robot Interaction. Human Factors 53, 5 (Oct. 2011), 517–527. DOI: https://doi.org/10.1177/0018720811417254

[4] Alhaji, B., Büttner, S. T., Kumar, S. S., and Prilla, M. 2025. Trust dynamics in human interaction with an industrial robot. Behav. Inf. Technol. 44, 2, 266–288. DOI: https://doi.org/10.1080/0144929X.2024.2316284

[5] Chi, V. B. and Malle, B. F. 2023. People dynamically update trust when interactively teaching robots. In Proceedings of the 2023 ACM/IEEE International Conference on Human–Robot Interaction (Stockholm, Sweden, March 13–16, 2023). HRI '23. ACM, New York, NY, 554–564. DOI: https://doi.org/10.1145/3568162.3576962

[6] Dragan, A. D., Lee, K. C. T., and Srinivasa, S. S. 2013. Legibility and predictability of robot motion. In Proceedings of the 8th ACM/IEEE International Conference on Human–Robot Interaction (Tokyo, Japan, March 03–06, 2013). HRI '13. IEEE Press, 301–308.

[7] Sarter, N. B., Woods, D. D., and Billings, C. E. 1997. Automation surprises. In Handbook of Human Factors and Ergonomics (2nd ed.), G. Salvendy, Ed. Wiley, New York, NY, 1926–1943.

[8] Chen, J. Y. C., Procci, K., Boyce, M., Wright, J., Garcia, A., and Barnes, M. 2014. Situation Awareness-based Agent Transparency. Technical Report ARL-TR-6905. U.S. Army Research Laboratory, Aberdeen Proving Ground, MD.

[9] Chen, J. Y. C. and Barnes, M. J. 2014. Human–agent teaming for multirobot control: A review of human factors issues. IEEE Trans. Hum.-Mach. Syst. 44, 1, 13–29.

[10] Endsley, M. R. 1995. Toward a theory of situation awareness in dynamic systems. Hum. Factors 37, 1, 32–64.

[11] van de Merwe, K., Mallam, S., and Nazir, S. 2022. Agent transparency, situation awareness, mental workload, and operator performance: A systematic literature review. Hum. Factors 66, 1, 180–208. DOI: https://doi.org/10.1177/00187208221077804

[12] Wallkötter, S., Tulli, S., Castellano, G., Paiva, A., and Chetouani, M. 2021. Explainable embodied agents through social cues: A review. ACM Trans. Hum.-Robot Interact. 10, 3, Article 27, 24 pages. DOI: https://doi.org/10.1145/3457188

[13] Bhaskara, A., Skinner, M., and Loft, S. 2020. Agent transparency: A review of current theory and evidence. IEEE Trans. Hum.-Mach. Syst. 50, 3, 215–224. DOI: https://doi.org/10.1109/THMS.2020.2965529

[14] Cai, M., Jin, Q., Zhou, J., et al. 2025. How transparency shapes the quality of human-robot interaction: An examination of trust, perception, and workload. Int. J. Soc. Robot. 17, 1335–1362. DOI: https://doi.org/10.1007/s12369-025-01255-0

[15] Kahneman, D. 1973. Attention and Effort. Prentice-Hall, Englewood Cliffs, NJ.

[16] Baddeley, A. 2003. Working memory: Looking back and looking forward. Nat. Rev. Neurosci. 4, 10, 829–839. DOI: https://doi.org/10.1038/nrn1201

[17] Cowan, N. 2001. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. Behav. Brain Sci. 24, 1, 87–114.

[18] Wickens, C. D. 2002. Multiple resources and performance prediction. Theor. Issues Ergon. Sci. 3, 2, 159–177.

[19] Bartlett, F. C. 1932. Remembering: A Study in Experimental and Social Psychology. Cambridge University Press, Cambridge, UK.

[20] Rumelhart, D. E. 1980. Schemata: The building blocks of cognition. In Theoretical Issues in Reading Comprehension, R. J. Spiro, B. C. Bruce, and W. F. Brewer, Eds. Lawrence Erlbaum Associates, Hillsdale, NJ, 33–58.

[21] Schank, R. C. and Abelson, R. P. 1977. Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures. Lawrence Erlbaum Associates, Hillsdale, NJ.

[22] Mandler, J. M. 1984. Stories, Scripts, and Scenes: Aspects of Schema Theory. Lawrence Erlbaum Associates, Hillsdale, NJ.

[23] Bransford, J. D. and Johnson, M. K. 1972. Contextual prerequisites for understanding: Some investigations of comprehension and recall. J. Verbal Learn. Verbal Behav. 11, 6, 717–726.

[24] Goffman, E. 1974. Frame Analysis: An Essay on the Organization of Experience. Harper & Row, New York, NY.

[25] Goffman, E. 1983. The interaction order. Am. Sociol. Rev. 48, 1, 1–17.

[26] Nishida, H. 1999. A cognitive approach to intercultural communication based on schema theory. Int. J. Intercult. Relat. 23, 5, 753–777.

[27] Nishida, H. 2005. Cultural schema theory. In Theorizing About Intercultural Communication, W. B. Gudykunst, Ed. Sage, Thousand Oaks, CA, 401–418.

[28] Suchman, L. A. 1987. Plans and Situated Actions: The Problem of Human–Machine Communication. Cambridge University Press, Cambridge, UK.

[29] Goodwin, C. 2000. Action and embodiment within situated human interaction. J. Pragmat. 32, 10, 1489–1522.
[30] Dourish, P. 2004. What we talk about when we talk about context. Pers. Ubiquitous Comput. 8, 1, 19–30.
[31] Turner, J. C. 1994. Social categorization and the self–concept: A social cognitive theory of group behavior. Adv. Group Process. 11, 77–122.
[32] Thompson, S., Candon, K., and Vázquez, M. 2025. The Social Context of Human-Robot Interactions. arXiv preprint arXiv:2508.13982.
[33] Clark, H. H. 1996. Using Language. Cambridge University Press, Cambridge, UK.
[34] Levinson, S. C. 2006. On the human "interaction engine". In Roots of Human Sociality: Culture, Cognition and Interaction, N. J. Enfield and S. C. Levinson, Eds. Berg, Oxford, UK, 39–69.
[35] Biddle, B. J. 1986. Recent developments in role theory. Annu. Rev. Sociol. 12, 67–92.
[36] Taylor, S. E. and Crocker, J. 1981. Schematic bases of social information processing. In Social Cognition: The Ontario Symposium, Vol. 1, E. T. Higgins, C. P. Herman, and M. P. Zanna, Eds. Lawrence Erlbaum Associates, Hillsdale, NJ, 89–134.
[37] Huang, C.-M. and Mutlu, B. 2013. Modeling and evaluating narrative gestures for humanlike robots. In Proceedings of Robotics: Science and Systems (Berlin, Germany, June 2013).
[38] Tulving, E. 1985. How many memory systems are there? Am. Psychol. 40, 4, 385–398.
[39] Zola-Morgan, S. and Squire, L. R. 1990. The primate hippocampal formation: Evidence for a time-limited role in memory storage. Science 250, 4978, 288–290.
[40] Fiske, S. T. and Neuberg, S. L. 1990. A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In Advances in Experimental Social Psychology, Vol. 23, M. P. Zanna, Ed. Academic Press, New York, NY, 1–74.
[41] Weick, K. E. 1995. Sensemaking in Organizations. Sage, Thousand Oaks, CA.
[42] Strauss, C. and Quinn, N. 1997. A Cognitive Theory of Cultural Meaning. Cambridge University Press, Cambridge, UK.
[43] Tonkin, M., Vitale, J., Herse, S., Williams, M.-A., Judge, W., and Wang, X. 2018. Design Methodology for the UX of HRI: A Field Study of a Commercial Social Robot at an Airport. In Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI '18), Chicago, IL, USA, March 5–8, 2018. ACM, New York, NY, 407–415. DOI: https://doi.org/10.1145/3171221.3171270
[44] Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., and Yang, Q. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23). ACM, New York, NY. DOI: https://doi.org/10.1145/3544548.3581388
[45] Kim, Y., Lee, J., Kim, S., Park, J., and Kim, J. 2024. Understanding Users' Dissatisfaction with ChatGPT Responses: Types, Resolving Tactics, and the Effect of Knowledge Level. In Proceedings of the 29th International Conference on Intelligent User Interfaces (IUI '24), March 18–21, 2024, Greenville, SC, USA. ACM, New York, NY, 385–404. DOI: https://doi.org/10.1145/3640543.3645148
[46] Passi, S. and Vorvoreanu, M. 2022. Overreliance on AI: Literature Review. Microsoft Technical Report MSR-TR-2022-12. Microsoft Corporation.
[47] Nass, C. and Moon, Y. 2000. Machines and Mindlessness: Social Responses to Computers. J. Soc. Issues 56, 1, 81–103. DOI: https://doi.org/10.1111/0022-4537.00153
[48] Komatsu, T., Kurosawa, R., and Yamada, S. 2012. How Does the Difference Between Users' Expectations and Perceptions About a Robotic Agent Affect Their Behavior? Int. J. Soc. Robot. 4, 2, 109–116. DOI: https://doi.org/10.1007/s12369-011-0122-y
[49] Pelikan, H. R. M., Reeves, S., and Cantarutti, M. N. 2024. Encountering Autonomous Robots on Public Streets. In Proceedings of the 2024 ACM/IEEE International Conference on Human–Robot Interaction (HRI '24) (Boulder, CO, USA, March 11–14, 2024). ACM, New York, NY, 561–571. DOI: https://doi.org/10.1145/3610977.3634936
[50] Gehle, R., Pitsch, K., and Wrede, S. 2014. Signaling Trouble in Robot-To-Group Interaction: Emerging Visitor Dynamics with a Museum Guide Robot. In Proceedings of the Second International Conference on Human-Agent Interaction (HAI '14) (Tsukuba, Japan, October 29–31, 2014). ACM, New York, NY, 361–368. DOI: https://doi.org/10.1145/2658861.2658887
[51] Triebel, R., Arras, K., Alami, R., et al. 2016. SPENCER: A socially aware service robot for passenger guidance and help in busy airports. In Field and Service Robotics: Results of the 10th International Conference (FSR 2015), D. S. Wettergreen and T. D. Barfoot, Eds. Springer Tracts in Advanced Robotics, Vol. 113. Springer, Cham, 607–622.
[52] Marchionini, G. 2006. Exploratory search: From finding to understanding. Commun. ACM 49, 4, 41–46.
[53] Tankelevitch, L., Kewenig, V., Simkute, A., Scott, A. E., Sarkar, A., Sellen, A., and Rintel, S. 2024. The Metacognitive Demands and Opportunities of Generative AI. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA. ACM, New York, NY. DOI: https://doi.org/10.1145/3613904.3642902
[54] Spitale, M., Parreira, M. T., Stiber, M., Axelsson, M., Kara, N., Kankariya, G., Huang, C.-M., Jung, M., Ju, W., and Gunes, H. 2024. ERR@HRI 2024 Challenge: Multimodal Detection of Errors and Failures in Human-Robot Interactions. arXiv preprint arXiv:2407.06094.
[55] Firmino de Souza, D., Sousa, S., Kristjuhan-Ling, K., Dunajeva, O., Roosileht, M., Pentel, A., Mõttus, M., Özdemir, M. C., and Gratšjova, Ž. 2025. Trust and Trustworthiness from Human-Centered Perspective in Human–Robot Interaction

(HRI)—A Systematic Literature Review. Electronics 14, 8, 1557. DOI: https://doi.org/10.3390/electronics14081557
[56] McGrath, M. J., Mousavizadeh, M., and Habib, K. 2024. Users do not trust recommendations from a large language model. Front. Comput. Sci. 6, 1456098. DOI: https://doi.org/10.3389/fcomp.2024.1456098
[57] Wolf, A. and Maier, C. 2024. ChatGPT usage in everyday life: A motivation-theoretic mixed-methods study. Int. J. Inf. Manag. 77, 102821. DOI: https://doi.org/10.1016/j.ijinfomgt.2024.102821
[58] Esterwood, C. and Robert Jr, L. P. 2023. Three Strikes and you are out!: The impacts of multiple human–robot trust violations and repairs on robot trustworthiness. Comput. Hum. Behav. 142, 107658. DOI: https://doi.org/10.1016/j.chb.2023.107658
[59] Ayoub, F., Kerr, A., and Villing, R. C. 2025. An Exploration of Trust in Human-Robot Interaction: From Measurement to Repair Strategies and Design Principles. In Human-Friendly Robotics 2024, 58–72. DOI: https://doi.org/10.1007/978-3-031-81688-8_5
[60] Lipton, Z. C. 2016. The mythos of model interpretability. arXiv preprint arXiv:1606.03490.
[61] Doshi-Velez, F. and Kim, B. 2017. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.
[62] Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. Artif. Intell. 267, 1–38.
[63] Levinson, S. C. and Torreira, F. 2015. Timing in turn-taking and its implications for processing models of language. Front. Psychol. 6, Article 731.
[64] Vaisey, S. 2009. Motivation and justification: A dual-process model of culture in action. Am. J. Sociol. 114, 6, 1675–1715.
[65] Neisser, U. 1976. Cognition and Reality: Principles and Implications of Cognitive Psychology. W. H. Freeman, San Francisco, CA.
[66] Bar, M. 2004. Visual objects in context. Nat. Rev. Neurosci. 5, 8, 617–629.
[67] Tonkin, M. 2021. Socially responsible design for social robots in public spaces. Ph.D. thesis. University of Technology Sydney, Sydney, Australia.
[68] Ambady, N. and Rosenthal, R. 1993. Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. J. Pers. Soc. Psychol. 64, 3, 431–441.
[69] Willis, J. and Todorov, A. 2006. First impressions: Making up your mind after a 100-ms exposure to a face. Psychol. Sci. 17, 7, 592–598.
[70] Tversky, A. and Kahneman, D. 1974. Judgment under uncertainty: Heuristics and biases. Science 185, 4157, 1124–1131.
[71] Sweller, J. 1988. Cognitive load during problem solving: Effects on learning. Cogn. Sci. 12, 2, 257–285.
[72] Abelson, R. P. 1981. Psychological status of the script concept. Am. Psychol. 36, 7, 715–729.
[73] Clark, H. H. and Brennan, S. E. 1991. Grounding in communication. In Perspectives on Socially Shared Cognition, L. B. Resnick, J. M. Levine, and S. D. Teasley, Eds. American Psychological Association, Washington, DC, 127–149.
[74] Kumar, S., Edan, Y., and Bensch, S. 2024. Enhancing robot understandability - a model to estimate varying levels of discrepancy. Preprint. DOI: https://doi.org/10.21203/rs.3.rs-5514182/v1
[75] Friston, K. 2010. The free-energy principle: A unified brain theory? Nat. Rev. Neurosci. 11, 2, 127–138.
[76] Clark, A. 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behav. Brain Sci. 36, 3, 181–204.
[77] Zaki, J. and Ochsner, K. 2009. The need for a cognitive neuroscience of naturalistic social cognition. Ann. N. Y. Acad. Sci. 1167, 16–30.
[78] Meyer, D. E., Kieras, D. E., Lauber, E., Schumacher, E. H., Glass, J., Zurbriggen, E., Gmeindl, L., and Apfelblat, D. 1997. Adaptive executive control: Flexible multiple-task performance without pervasive immutable response-selection bottlenecks. Acta Psychol. 90, 1–3, 163–190.
[79] Kwon, M., Huang, S. H., and Dragan, A. D. 2018. Expressing robot incapability. In Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction. ACM, New York, NY, 87–95.
[80] Lemaignan, S., Warnier, M., Sisbot, E. A., Clodic, A., and Alami, R. 2017. Artificial cognition for social human–robot interaction: An implementation. Artif. Intell. 247, 45–69. DOI: https://doi.org/10.1016/j.artint.2016.07.002
[81] Rich, C., Ponsler, B., Holroyd, A., and Sidner, C. L. 2010. Recognizing engagement in human-robot interaction. In 5th ACM/IEEE International Conference on Human-Robot Interaction, 375–382. IEEE. DOI: https://doi.org/10.1109/HRI.2010.5453163