Honda Research Institute EU
Honda Research Institute Europe GmbH
https://www.honda-ri.de/

A Taxonomy of Mental Model Mismatches for Human–Agent Interaction

Phillip Richter, Anna-Lisa Vollmer, Heiko Wersing

2026

Preprint:

Copyright ACM 2026. This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in ACM Transactions on Human-Robot Interaction, https://doi.org/10.1145/3846186.

A Taxonomy of Mental Model Mismatches for Human-Robot Interaction

PHILLIP RICHTER, Bielefeld University, Germany
HEIKO WERSING, Honda Research Institute Europe, Germany
ANNA-LISA VOLLMER, Bielefeld University, Germany

Mental model mismatches (MMM) in human-robot interaction occur when humans' internal representations of robots diverge from robots' actual properties or behaviors, leading to communication breakdowns, task failures, and reduced trust. These mismatches are subtle and difficult to detect, yet the field lacks shared terminology to identify specific types. This makes it difficult to diagnose problems, design targeted interventions, or compare findings across studies. We present a theoretically grounded taxonomy of sixteen MMM types organized by the Sense-Plan-Act (SPA) architecture into sensing (7 types), planning (5 types), and acting (4 types). Drawing from cognitive neuroscience, psychology, communication studies, and human factors, each type addresses a distinct cognitive divergence that arises when users apply mechanisms evolved for understanding humans to robotic systems. The taxonomy's exhaustiveness is established through a dimensional decomposition: three representational dimensions grounded in the neurocognitive architecture of human perception (What, Where, How) are systematically crossed with the three SPA processing stages, and every cell in the resulting matrix is accounted for by a dedicated type or an explicit justification. Each type is equipped with a formal definition and diagnostic examples. We demonstrate the taxonomy's applicability across two contrasting datasets, TEACh (task-oriented simulation) and SHREC (social therapeutic dialogue), revealing distinct mismatch profiles that reflect the structural differences between the two settings. This taxonomy provides precise terminology enabling researchers to identify specific cognitive mismatches, and a foundation for developing targeted interventions, rather than treating all problems as generic "communication failures."

CCS Concepts: • **Human-centered computing** → **HCI theory, concepts and models**; • **General and reference** → *Surveys and overviews.*

Additional Key Words and Phrases: Mental Model, HRI, Taxonomy, Cognitive Mismatch, Sense-Plan-Act

ACM Reference Format:
Phillip Richter, Heiko Wersing, and Anna-Lisa Vollmer. 2018. A Taxonomy of Mental Model Mismatches for Human-Robot Interaction. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 28 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Mental models have become an increasingly prominent topic in human-robot interaction (HRI) research (see Figure 1) [3, 101]. When humans' mental models of robots diverge from the robots' actual properties or behaviors, mental model mismatches (MMM) arise that can lead to communicative breakdowns, task failures, and reduced trust [48, 85]. These mismatches vary in type, are often subtle, and are difficult to detect. They can manifest in language, gestures, gaze, or physical collaboration, yet they share an underlying misalignment between user expectations and the system's

Authors' Contact Information: Phillip Richter, Bielefeld University, Bielefeld, Germany, p.richter@uni-bielefeld.de; Heiko Wersing, Honda Research Institute Europe, Offenbach am Main, Germany, heiko.wersing@honda-ri.de; Anna-Lisa Vollmer, Bielefeld University, Bielefeld, Germany, anna-lisa.vollmer@uni-bielefeld.de.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
Manuscript submitted to ACM

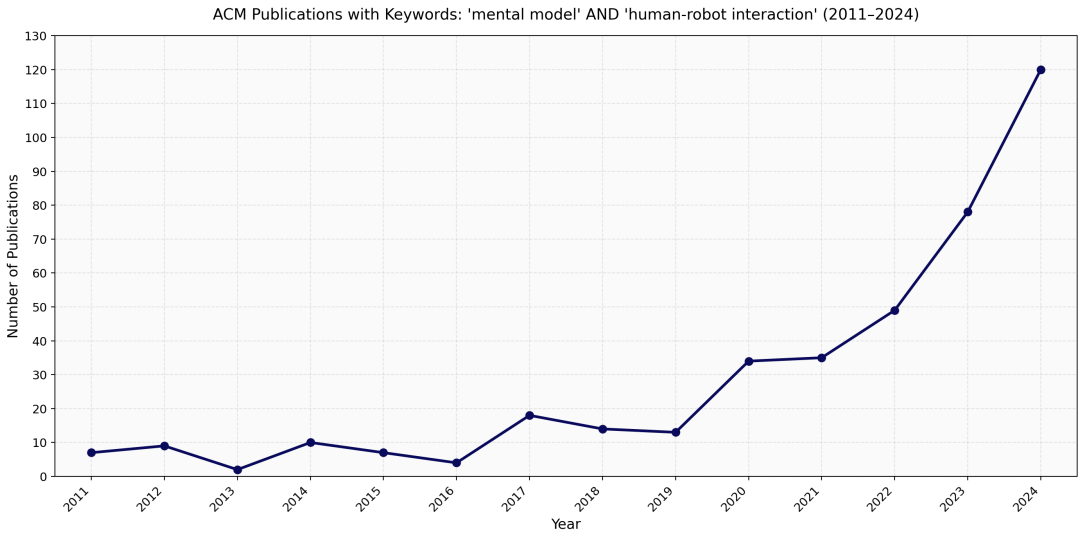

Fig. 1. Number of ACM publications containing the keywords *"mental model"* AND *"human-robot interaction"* (2011 to 2024), retrieved via ACM Digital Library search.

reality [46]. Unlike human-human interaction, where shared cognitive foundations provide common ground [16], HRI involves fundamentally different cognitive architectures [92]: users project human-like reasoning, perception, and capabilities onto robots that operate through entirely different mechanisms [41, 82]. Despite this growing attention, the field still lacks a shared vocabulary to describe specific types of mismatches. Without such terminology, it is difficult to diagnose problems, design targeted interventions, or compare findings across studies [46, 99].

Figure 2 makes this concrete. The user asks for "the green cup," but the robot cannot perceive colour, so the property the user relied on for disambiguation is invisible to it. If the robot then grabs the wrong cup or fails to act at all, an observer would naturally describe the failure as a referential ambiguity. The root cause, however, is perceptual, and the two failure modes require fundamentally different interventions. This is the kind of distinction a taxonomy of mismatches needs to make.

In response, we propose a taxonomy of mental model mismatches that is independent of tasks and modalities, based on cross-disciplinary research. Drawing from theories and findings in cognitive neuroscience, psychology, communication studies, and human factors, our approach adapts these insights for HRI contexts. We organize sixteen mismatch types along the Sense-Plan-Act (SPA) architecture [88, 96] and ground their exhaustiveness in three representational dimensions derived from the neurocognitive architecture of human perception [33, 86, 110]. We demonstrate the taxonomy's applicability by applying it to interaction data from the TEACh (Task-Driven Embodied Agents That Chat) [75] and SHREC (Social Human-Robot Embodied Conversation) [59] datasets, showing proper coverage across different interaction contexts.

Accordingly, this work makes three main contributions:

- A taxonomy of sixteen MMM types organized along the Sense-Plan-Act architecture, with exhaustiveness established through a dimensional decomposition that crosses three neurocognitively grounded representational dimensions (What, Where, How) with three processing stages (Sense, Plan, Act), providing precise, task-independent terminology to identify and communicate distinct mismatches in human-robot interaction.

- Diagnostic operationalization of the taxonomy through formal definitions and examples, enabling systematic annotation of interaction failures.
- A demonstration of the taxonomy's applicability across two contrasting datasets (task-oriented simulation and social therapeutic dialogue), revealing distinct mismatch profiles that reflect the structural differences between the two settings.

The taxonomy shall support identifying problems, mapping them to the right point in the robot's processing pipeline and should then facilitate the design of targeted specific interventions. We name a candidate intervention with each mismatch type in Section 4.

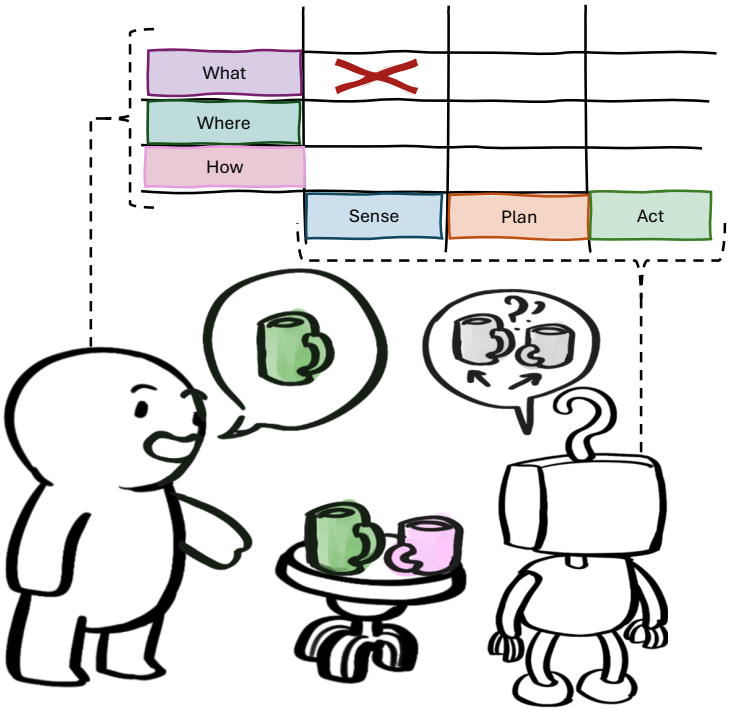

Fig. 2. The user asks for "the green cup," but the robot cannot perceive colour. It detects both cups but, with the colour cue invisible to it, cannot tell them apart and is left with two candidates. The MMM matrix on the right diagnoses the failure at the intersection of What (the property "green") and Sense (the perception stage): a Sense-Object-What Mismatch. An observer might describe this as the robot getting confused about which cup was meant, but the taxonomy presented in Section 4 traces the root cause back to an earlier stage: the failure lies in what the robot can perceive, not in how it interprets the reference, and the two require fundamentally different interventions.

2 Mental Models and Mismatches

Mental models are internal representations that enable humans to simulate reality, predict outcomes, and guide behavior [22, 50, 51]. These dynamic cognitive structures function as working representations that humans construct and update based on perception and experience, preserving key relational and causal properties of external systems [1, 50, 54]. In HRI, users construct mental models of robots including their capabilities, behaviors, and internal processes [51, 101].

These models are particularly prone to inaccuracy because users apply cognitive mechanisms evolved for understanding other humans, attributing human-like reasoning and intentionality to robotic systems [26, 70, 105]. Importantly, mental models are shaped not only by the robot's behavior but also by its physical design and appearance, which set initial expectations before any interaction occurs [79]. Because most users have limited prior experience with robots, their initial models are often based on media portrayals, anthropomorphic assumptions, or analogies to familiar devices [27, 47, 64], and these models update gradually as direct experience accumulates.

Mental model mismatches occur when a human's internal representation of a robot systematically diverges from the robot's actual properties, behaviors, or intended design [42, 43, 85, 101]. Unlike random errors or temporary confusion, these mismatches tend to be systematic: they recur predictably across similar situations rather than occurring as isolated incidents [52], involve a consistent discrepancy between predicted and actual robot behavior [14, 84], and persist despite disconfirming evidence, in line with general patterns of belief perseverance in human cognition [2, 97]. Critically, these mismatches are not user errors but misalignments between the user's mental model and the robot's operational model, shaped by design decisions on the robot side as much as by cognitive tendencies on the human side [68].

Three families of cognitive mechanisms explain why these mismatches arise [46, 60, 90, 94]. The first is mental state inference: users predict robot behavior by automatically attributing beliefs, desires, and intentions to agents exhibiting goal-directed behavior [20, 80, 95, 107], while overemphasizing dispositional over situational factors [32, 87]. The second is communication and grounding: users assume shared vocabularies, contextual understanding, and signal interpretability that robots typically lack, including assumptions about which communicative modalities the robot can receive [16, 17, 34, 53]. The third is trust calibration and affordance perception: users form systematic misjudgments about perceptual range, physical constraints, or manipulation competence [31, 44, 60, 76, 109]. These mechanisms address different aspects of cognition, supported by converging evidence from social neuroscience on the interactive nature of social cognition [94], and identified as distinct failure points in human-automation research.

The consequences of these mismatches span multiple temporal scales. Immediate effects include communication failures, errors in task execution, and user confusion during ongoing interactions [46]. Intermediate consequences include decreased trust in the robot [93], increased cognitive load due to monitoring the system's behavior [76], and reduced willingness to utilize the robot's advanced capabilities. Long-term impacts can lead to system abandonment [78], negative attitudes toward robotics, and reduced willingness to engage with autonomous robots in general [37]. Having a shared vocabulary to describe these mismatches is therefore essential for diagnosing problems, developing targeted interventions, and comparing findings across studies. Developing such a vocabulary requires understanding what existing frameworks already cover and where the diagnostic gap remains.

3 Related Work

Tabrez et al. [101] provide a comprehensive survey of mental models in HRI, mapping how models form, update, and influence interaction outcomes. Andrews et al. [3] review mental models across HRI contexts, focusing on their role in trust and collaboration. Both works establish the importance of mental models and the conditions under which they develop. Our taxonomy builds on this foundation by providing specific categories for *what goes wrong* when mental models are inaccurate.

While these surveys map how models form and evolve, a complementary question is how robots can model human mental states in return. Tavella et al. [103] propose a Theory of Mind (ToM) framework for HRI that examines how robots can model human mental states. Talamadupula et al. [102] demonstrate how robots can use mental modeling and plan recognition to coordinate with human teammates. Earlier work by Scassellati [91] laid foundations for implementing

ToM capabilities in humanoid robots. Our taxonomy addresses the complementary direction: the human's understanding of the robot. Together, the two perspectives cover both sides of the cognitive coordination problem. A robot equipped with ToM capabilities could use our categories to determine *which* aspect of the user's model needs correction.

Understanding mismatches also raises the question of how to prevent them. Wachowiak et al. [111] develop a taxonomy of system transparency in HRI, categorizing how robots communicate their internal states to users. Work on robot explanations [41, 99] addresses how to make robot behavior interpretable. These frameworks describe *interventions* to prevent or resolve mismatches, while our taxonomy describes the *targets* of those interventions. For example, an interaction prone to vocabulary mismatches needs vocabulary clarification, while one prone to physical capability mismatches needs capability disclosure.

Berzuk and Young [9] develop a framework for understanding expectation discrepancy in social HRI, tracing how robot appearance, media exposure, and prior experience shape the expectations users bring to an interaction. Their framework addresses the *sources* of misaligned expectations, while our taxonomy addresses the *types* of mismatches that result.

Closely linked to transparency is the question of trust. Hancock et al. [37, 38] identify factors influencing trust calibration in HRI. Tian and Oviatt [106] offer a taxonomy of social errors in human-robot interaction. Schaefer et al. [93] develop trust measurement instruments. Trust miscalibration is one of the cognitive *mechanisms* that produces several of our mismatch types (particularly those involving physical capabilities and manipulation), while relationship degradation is a *consequence* of persistent mismatches. These frameworks operate at different levels of the same causal chain.

Beyond trust, broader frameworks address how coordination quality can be measured. Hoffman [45] develops metrics for evaluating collaborative fluency. Jiang and Arkin [49] propose a framework for mixed-initiative interaction. Tsiakas et al. [108] provide a taxonomy for robot-assisted training. These frameworks measure interaction *quality* and describe coordination *patterns*. Our taxonomy complements them by explaining *why* coordination fails: when fluency scores drop, our categories can identify whether the cause is perceptual, linguistic, or related to planning expectations.

One specific modality that has received dedicated attention is spatial behavior. Work on spatial HRI [66, 98, 112] addresses physical positioning, distance regulation, and spatial comfort. Our taxonomy addresses the cognitive mismatch aspect of spatial interaction (misaligned reference frames), while proxemics frameworks address broader spatial coordination patterns.

Beyond HRI-specific work, the broader HCI and human factors literature offers error classification frameworks that address complementary aspects. Reason's classification of human error [83] distinguishes slips, lapses, and mistakes based on cognitive processing level. Nielsen's usability heuristics [72] identify interface-level problems such as visibility of system status or error prevention. These frameworks address errors in human cognition or interface design at a general level. Our taxonomy is more specific: it targets the particular cognitive mismatches that arise when users apply human-oriented mental models to robotic systems, a problem that general error frameworks do not differentiate.

In summary, existing frameworks address how mental models *form* (surveys), how robots can *model humans* (ToM), how to *prevent* mismatches (transparency), what *consequences* mismatches produce (trust), how to *measure* interaction quality (coordination metrics), and how to *manage space* (proxemics). Our taxonomy adds a diagnostic vocabulary: a classification of *what specific cognitive mismatches occur* between human expectations and robot reality. These frameworks can be used in combination. For instance, one could use our taxonomy to diagnose a mismatch type, a transparency framework to design the intervention, and a trust metric to evaluate the outcome.

4 The MMM Taxonomy

We present sixteen mismatch types organized along the Sense-Plan-Act (SPA) architecture [71, 88]. This section first introduces the taxonomy's scope and organizing principles (Section 4.1). It then presents each type with its definition and examples (Sections 4.2 to 4.4). Finally, it establishes the theoretical grounding and exhaustiveness of the classification (Section 4.5). A full diagnostic reference for all sixteen types is provided in Appendix A.

4.1 Scope and Overview

Mental model mismatches are divergences between a human's cognitive representation of a robot and the robot's actual properties or behaviors. Our taxonomy addresses collaborative task-oriented human-robot interaction, where humans and robots work together toward shared goals in a dynamic physical environment. This setting is where mismatches about what, where, and how most directly affect coordination, and where the three representational dimensions introduced below are all actively engaged. Section 4.6 details the phenomena that fall outside this scope and the frameworks that address them.

In such environments, any instruction that coordinates human-robot action draws on three kinds of information: which entities are present and what properties they have, where those entities are located relative to each other, and how the state of the environment changes over time through actions and events. "Put the red cup on the top shelf" specifies an entity and its property (red cup), a spatial target (top shelf), and a state change to be achieved (moving the cup there). We refer to these three kinds of information as the *representational dimensions* of the taxonomy: identity and properties (*What*), spatial position and relations (*Where*), and state change and action (*How*).

We organize mismatches by where in the robot's processing pipeline they occur, using the SPA architecture. SPA constitutes the three primitives of robotics [71]: a robot *senses* its environment through sensors [29], *plans* by maintaining a world model and selecting actions [88], and *acts* through physical movement and communication [96]. Every robot architecture, whether hierarchical, reactive, or hybrid, is defined by how it organizes these three stages [5, 13]. This decomposition is consistent with Parasuraman, Sheridan, and Wickens' four-stage model of human-automation interaction [77], which maps information acquisition to Sense, information analysis and decision selection to Plan, and action implementation to Act.

Within both Sense and Act, we identify two processing channels: an *object channel* for detecting, localizing, and tracking physical entities, and a *communication channel* for understanding or producing communicative signals. This parallel structure draws on the theoretical view that language comprehension shares key properties with physical object perception [7, 62]. The communication channels in Sense and Act exhibit a systematic symmetry: both require that the communication modality exists before any content can be processed. In Sense, the robot must be able to receive a communicative signal before it can interpret its content. In Act, the robot must be able to produce a communicative signal before the user can interpret it.

This yields the taxonomy's structure: sixteen mismatch types distributed across Sense (seven types), Plan (five types), and Act (four types). The types map onto the three representational dimensions, but not uniformly: some cells contain a dedicated type, others are spanned by cross-dimensional preconditions (SCNA, PSTD, ACNA) or by integrated types whose distinguishing axis is not the representational dimension. The two Plan World Context types (PWCS, PWMM), for example, are distinguished along a temporal axis (current versus past) rather than along What/Where/How. Section 4.5 presents the consolidated matrix in Table 2.

The structure of the next three sections mirrors the SPA architecture. Section 4.2 covers Sense mismatches in two subcategories: an object channel with three types and a communication channel with four types. Section 4.3 covers Plan mismatches in three subcategories: design scope with one type, world context with two types, and planning algorithm with two types. Section 4.4 covers Act mismatches in two subcategories: physical execution with two types and communication output with two types. Section 4.5 then provides the neurocognitive grounding for the three dimensions and presents the consolidated matrix. For each mismatch type we provide typical examples and candidate interventions.

4.2 Sense Mismatches

Sense mismatches occur when users hold incorrect assumptions about the robot's perceptual capabilities. We organize seven Sense types into two channels: an object channel for perceiving physical entities in the environment, and a communication channel for processing human communicative signals. Both channels process information along all three representational dimensions. They do so through different mechanisms that reflect the difference between physical perception and language comprehension. The following subsections define each channel and its types, each presented with a definition and examples.

4.2.1 Object Channel. The object channel concerns the robot's perception of physical entities and events. Robots sense through optical, acoustic, mechanical, thermal, chemical, and magnetic modalities [29], each constrained by range, resolution, accuracy, response time, and operating conditions. Users, drawing on trust calibration heuristics and anthropomorphic projection [25, 60], systematically form miscalibrated assumptions about these capabilities. The three types in this subsection populate the What, Where, and How cells of the Sense object row in the matrix.

The three types form a branching structure: detection is a precondition for both localization and state change tracking. One cannot localize or track changes in an entity that has not been detected yet. Localization and state change tracking are orthogonal to each other: a robot may correctly track an object's position without noticing that its state has changed, or notice that water is boiling without precisely localizing the pot. We use "sequential dependency" to mean that a later type typically presupposes success at an earlier stage. We use "orthogonal" to mean that two types address independent dimensions and can fail independently of each other. These relationships serve as annotation guidelines rather than strict logical constraints.

Sense-Object-What Mismatch (SOWT). The user's assumption about the robot's ability to detect an entity diverges from its actual detection capability. The robot fails to recognize that an entity is present, misidentifies it, or fails to perceive a relevant property. The mismatch may also run in the opposite direction: the user assumes the robot cannot detect something it actually can. In ambiguous cases, SOWT is identified by whether the gap can be closed through perception. It applies when extended search or user guidance ultimately resolves the failure. It does not apply when the underlying gap remains open despite such attempts.

Examples: A cup stands next to the robot but outside its camera's field of view, and the user assumes the robot can see it. A robot visually detects an object but classifies a cup as a bowl. *Intervention:* The robot can reposition to bring the object into view, or signal that it cannot see what the user referenced, and explicitly name the detected object.

Sense-Object-Where Mismatch (SOWE). The user's assumption about the robot's ability to determine an entity's spatial position diverges from its actual localization capability. The robot detects the entity but cannot correctly determine where it is. The mismatch may also run in the opposite direction: the user incorrectly predicts the robot's spatial estimate. SOWE presupposes that detection (SOWT) has succeeded [24, 53].

Examples: A robot recognizes a cup but misjudges its distance due to insufficient depth resolution. The user and robot operate in different spatial reference frames, leading to disagreement about "left" and "right." *Intervention:* The robot can take a second viewpoint to refine the estimate, or make its own reference frame explicit so left and right are shared.

Sense-Object-How Mismatch (SOHW). The user's assumption about the robot's ability to register a change of state over time diverges from its actual capability. The robot has detected the entity and correctly represents its identity and current attributes (SOWT satisfied), but fails to perceive that the entity's state has transitioned from one condition to another while the interaction unfolds. The mismatch lies in *How*, not *What*: what is missed is not the entity's identity or current attribute but the dynamic transition itself (the ventro-dorsal perception of state change, Section 4.5.1). The mismatch may also run in the opposite direction: the user assumes the robot cannot track a change that it actually can. SOHW presupposes that the entity has been detected (SOWT satisfied) and is orthogonal to SOWE. The mismatch is sensory: the change occurred but the robot's perceptual system did not register it.

Examples: The robot correctly perceives a pot of water on the stove, but does not register that the water has begun to boil, because the state transition fell outside its perceptual sampling. A cup the robot had perceived as empty has since been filled, and the robot continues to treat it as empty because it did not perceive the change. *Intervention:* The robot can re-check the entity rather than perceiving it only once, or use context knowledge to anticipate the change, for example that water in a pot on a hot stove will begin to boil.

4.2.2 Communication Channel. While the object channel concerns the robot's perception of physical entities, the communication channel concerns its perception of human communicative signals. The communication channel is modality-agnostic: it applies to any communicative signal the user produces, whether spoken language, gesture, gaze, facial expression, or body posture. The channel begins with the question of whether the robot can receive the communicative modality at all. It then proceeds through three content-processing stages in sequential dependency: entity semantics, referential binding, and instruction semantics. This staged organization is consistent with neurocognitive models of language comprehension that posit distinct, ordered processing phases [30, 36]. This placement in Sense reflects the view that communication comprehension can be understood as a perceptual process, following the same detection-to-interpretation logic as physical object perception [7, 62, 81]. The three content types populate the What, Where, and How cells of the Sense communication row in the matrix.

In collaborative task-oriented HRI situated in shared physical environments, referential binding ultimately resolves to spatial individuation: identifying which among physically co-present instances is meant [18, 58]. While disambiguation cues may be property-based ("the red one") or relational ("the one near the sink"), they all serve to select a specific entity at a specific location in the shared workspace. Cases where references depend on interaction history rather than the current spatial scene fall outside the communication channel and are addressed in Section 4.3.2.

Sense-Communication-Not-Available Mismatch (SCNA). The user communicates through a modality that the robot cannot receive or process as a communicative signal. The cause is either that the sensory hardware is absent or that the modality is not implemented as a communication input channel. Unlike the content processing types, SCNA does not address a single representational dimension but operates across all three. The absence of a communication input modality prevents any communicative content from being received. We call such types *cross-dimensional preconditions*: their failure blocks all three representational dimensions simultaneously, making dimensional subdivision uninformative. Plan-Scope-Task-Design Mismatch (PSTD, Section 4.3.1) and Act-Communication-Not-Available Mismatch (ACNA, Section 4.4) share this property. ACNA is also the output-side counterpart of SCNA, playing the same gating role for

the robot's communicative output. SCNA is a precondition for content processing: only when the modality is received does the question of content become relevant.

A critical distinction from SOWT is the following. The SOWT type concerns whether the robot detects a physical entity in the environment, whereas SCNA concerns whether the robot interprets a human behavior as a communicative act. A robot may visually detect a user's hand (SOWT satisfied) without recognizing a pointing gesture as a referential signal (SCNA failure).

Examples: A user points at an object, but the robot has no gesture recognition capability. A user speaks to a robot that has no microphone or speech recognition. *Intervention:* The robot can tell the user which input channels it has, so the user switches to one it can receive.

Sense-Communication-What Mismatch (SCWT). The user employs a term, attribute, or concept that the robot does not have in its vocabulary. The meaning itself is unknown to the robot. SCWT presupposes that the communicative modality is received (SCNA satisfied) [15, 16, 39, 55, 104].

Examples: A user says "hand me the carafe" and the robot has no representation of "carafe." A user instructs "make the sauce spicier" and the robot cannot map "spicy" to any property. *Intervention:* The robot can ask what the unknown word means and learn it, or offer the terms it does know.

Sense-Communication-Where Mismatch (SCWE). The user references an entity that the robot cannot uniquely identify, despite knowing the term, because multiple candidates are physically co-present in the shared environment. SCWE presupposes that the term is known (SCWT satisfied) [4, 16, 35].

Examples: A user says "the cup" when three cups are visible on the table. A user says "the bowl next to the stove" when two bowls are near the stove. *Intervention:* The robot can ask which one is meant, or name the candidates it is choosing between.

Dependencies between the object and communication channels can arise. If the robot detects only one of three co-present cups due to a perceptual limitation (SOWT), the referential ambiguity is artificially resolved because only one candidate exists in the robot's world model. In such cases, the root cause is perceptual rather than referential, and the taxonomy classifies at the earliest failure point (SOWT).

Sense-Communication-How Mismatch (SCHW). The user issues a communicative act that the robot cannot map to an action. The mismatch concerns action, the *How* dimension. The robot may know the terms and resolve the references, but cannot turn the utterance into something to do. This also covers the recognition that no action is required: the robot may treat a non-instructive utterance, such as an acknowledgment, as a command. SCHW presupposes that terms are known (SCWT satisfied). Where the utterance contains referential expressions, it also presupposes that references are resolved (SCWE satisfied). Some acts such as "I'm cold" or "thanks" contain no referential expression, so SCWE does not apply and SCHW applies once the terms are understood [34]. We use "instruction" broadly for any communicative act with task-relevant implications, from explicit commands to implicit requests to utterances that should trigger no action.

Examples: A user says "get the kitchen ready for dinner." The robot knows every word but has no decomposition of the directive into the concrete actions it implies (lower inferential demand). A user says "I'm cold," and the robot does not infer the implicit request to bring a blanket or adjust the temperature (higher inferential demand, indirect speech act). A user says "thanks," and the robot treats the acknowledgment as a new command rather than recognizing that no action is required. *Intervention:* The robot can state the action it inferred and confirm it before acting, rather than executing on a guess.

4.3 Plan Mismatches

Once sensory input has been processed, the robot integrates information into its world model and uses it to select actions. Plan mismatches arise when users' assumptions diverge from the robot's actual computational processes for maintaining world knowledge or selecting actions [88]. These mismatches are driven primarily by attribution and ToM mechanisms: users automatically infer human-like beliefs, goals, and reasoning in agents exhibiting goal-directed behavior [20, 25]. The exception is PSTD. It reflects miscalibrated expectations about the robot's purpose, not mental-state inference. Table 1 groups it accordingly.

Before the robot engages in any planning, the task must fall within its intended design scope. The remaining Plan types then address the robot's world knowledge and action selection. Together, the five Plan types populate the Plan rows of the matrix in different ways: as a cross-dimensional precondition, as integrated cells, and as dimension-specific types.

4.3.1 Design Scope. **Plan-Scope-Task-Design Mismatch (PSTD).** The user expects the robot to count a task category as part of its purpose, but the task lies outside the robot's intended design scope. PSTD asks whether the robot is designed for this kind of task at all, prior to any question about objects, locations, or actions involved in executing it. Like SCNA, PSTD is a cross-dimensional precondition: it operates across all three representational dimensions, because a task outside the design scope fails regardless of which entities, spatial configurations, or state changes are involved. ACNA (Section 4.4) shares this same property. PSTD is a precondition for the other Plan types: only tasks that pass this check enter the planning stages [37, 69, 73, 76].

Examples: A household assistance robot is asked to diagnose a medical symptom. A kitchen robot is asked to do the laundry. *Intervention:* The robot can state what it is designed for and redirect the user to a task within that scope.

Once a task passes the design scope check, the remaining Plan types address two aspects of how the robot processes that task: what it knows about the current state of the world, and how it uses that knowledge to select actions.

4.3.2 World Context. The robot's world model represents the current state of the environment as an integrated structure encompassing entity properties, spatial relations, and state information. This includes the robot itself: its internal states (battery level, joint temperature, storage capacity) are part of the world state. Two types address this representation, distinguished not by representational dimension but by temporal status: whether the information was never acquired or was once available but has since been lost. Both types integrate across all three representational dimensions, because the world model is a unified structure and forgotten content can belong to any dimension. Splitting either type by dimension would create subtypes that require the same intervention and that annotators could not reliably distinguish.

Plan-World-Current-State Mismatch (PWCS). The user assumes the robot knows a fact about the current world state, but this information is missing from the current robot's world model state. Perception would have been possible in principle (excluding SOWT, SOWE, and SOHW), but the fact is not represented in the current state: it was outside the robot's sensory range, requires inference the robot cannot perform, was not communicated by the user, or constitutes background world knowledge the robot lacks [23, 80, 88].

Examples: A user assumes the robot knows the fridge contains milk, but the robot has not opened the fridge and was not told. A user assumes the robot knows its own dust bag is full, but the robot has no fill-level sensor. *Intervention:* The robot can go and get the missing fact, for example open the fridge, or ask the user instead of assuming it knows.

Plan-World-Memory Mismatch (PWMM). The user assumes the robot remembers information from previous interactions, earlier states, or past context, but this information is no longer available. The information was once present

in the robot's model but has been lost through memory resets, capacity limits, session boundaries, or context decay. PWMM is orthogonal to PWCS: PWCS concerns information the robot never had, while PWMM concerns information that was once available but has since been lost. The two types are distinguished by temporal status rather than by representational dimension [10, 61]. Whether the robot is aware of its own memory gap is a separate question. A robot that detects the gap and communicates it can prevent the mismatch from causing a coordination failure. A robot that is unaware of the gap cannot.

Examples: A user says "like last time" after a system reset and the robot has no record of the previous session. A user references a preference expressed earlier in the conversation that has dropped from the robot's context window. *Intervention:* The robot can flag early that it does not remember, so the user does not rely on a shared history that is gone.

4.3.3 Planning Algorithm. Beyond what the robot knows about the world, mismatches arise in how it uses that knowledge to select actions. The planning algorithm operates exclusively along the *How* dimension: it transforms world knowledge into action selection. Planning consumes What and Where information from the World Context. Mismatches in that information are classified as PWCS or PWMM, not as planning-algorithm types. When the algorithm selects among alternative objects for the same functional role (e.g., choosing a large pot over a small one), this constitutes a method choice rather than a What-dimension mismatch: the question is not which objects exist but how to use them. Two types address different aspects of action selection, treated as orthogonal dimensions.

Plan-Algorithm-Method Mismatch (PAMT). The robot understands the task and plans at an appropriate level of abstraction, but selects a different method, strategy, or parameter values than the user expects. PAMT encompasses tool selection, execution order, optimization criteria, and quantifiable parameters such as temperature, speed, or force. The distinction between "different method" and "different parameter values" represents a continuum rather than a categorical boundary. PAMT presupposes that the instruction was understood (not SCHW) and that the robot plans at an appropriate level of abstraction [25].

Examples: A user expects the robot to boil a potato whole, but the robot's schema for "boil a potato" includes a slicing step. A robot heats water to 60°C when the user expected 80°C. *Intervention:* The robot can preview its intended method before an irreversible step, so the user can correct it in time.

Plan-Algorithm-Horizon Mismatch (PAHZ). The user's assumption about the robot's planning horizon or level of abstraction diverges from its actual planning capability. The user issues a high-level goal expecting autonomous decomposition, but the robot requires step-by-step guidance. Alternatively, the user provides detailed instructions to a robot capable of autonomous planning that acts ahead of the user's guidance [21, 100]. A third case also applies: the user provides detailed step-by-step instructions, but the robot's planning system can only accept high-level goals and has no interface for receiving fine-grained commands. All three cases concern the level of abstraction at which the robot can operate.

Examples: A user says "set the table" expecting the robot to retrieve plates, cutlery, and glasses and arrange them, but the robot can only execute one action at a time. A user says "make a salad" and the robot asks "what's first?" because it cannot decompose the high-level goal. *Intervention:* The robot can preview the decomposition it plans, so a gap between its horizon and the user's expectation shows before execution.

4.4 Act Mismatches

Planning determines what the robot intends to do. Act mismatches concern the final stage: whether the robot can execute those intentions, both physically and communicatively. These mismatches occur when users' assumptions about robot execution capabilities diverge from actual performance. The physical types (APWE, APHW) are driven by miscalibrated affordance judgments and trust miscalibration [31, 60, 74], while the communication-output types (ACNA, ACSC) arise from grounding assumptions about which signals the robot can produce and how they will be interpreted, and are grouped accordingly in Table 1.

4.4.1 Physical Mismatches. Physical execution involves only Where and How. The What dimension is absent because the motor execution system operates over target coordinates and force parameters, not over object identities. Errors that look like physical wrong-object actions originate upstream, in perception (SOWT, SOWE) or reference resolution (SCWE). Within the physical channel, the two types follow a sequential dependency: manipulation typically applies only when reachability is satisfied.

Act-Physical-Where Mismatch (APWE). The user holds an incorrect assumption about the robot's kinematic limits. The robot cannot reach a required pose, position, or movement. Alternatively, the user incorrectly assumes the robot cannot perform a movement it is actually capable of. Users systematically misjudge reachability based on their own body schema [31].

Examples: A shelf is too high for the robot's arm to reach. A user asks the robot to pick up something from under a low table, but the arm cannot reach the required pose. *Intervention:* The robot can show its reach limit through motion, reaching and stopping short, so the user sees the boundary rather than being told.

Act-Physical-How Mismatch (APHW). The user holds an incorrect assumption about the robot's object interaction capability. The robot reaches the object (APWE is satisfied) but cannot correctly handle it. APHW presupposes reachability, and requires force control, dexterity, and surface adaptation that users misjudge based on familiar affordance cues [31, 74, 89]. The distinction from APWE is whether the failure occurs before or after contact with the object. APWE means the robot cannot bring its end effector to the required position at all. APHW means the robot reaches the object but fails during the physical interaction itself.

Examples: A wet glass slips from the robot's gripper. The robot grips a fragile object with too much force and damages it. *Intervention:* The robot can ask about the object's properties, such as weight or material, before gripping, rather than discovering the limit by dropping it.

4.4.2 Communication Output. Robots also act by producing communicative signals, and mismatches arise here as well. Two types address this channel: the first concerns whether the communication modality exists at all, the second concerns whether the user correctly interprets a signal that the robot produces.

Act-Communication-Not-Available Mismatch (ACNA). The user expects the robot to possess a communication modality that it does not have. The cause is either that the hardware is absent or that the modality is not implemented. The user waits for or attempts to initiate communication through a channel that the robot is structurally or functionally unable to use [28]. ACNA is the output-side counterpart of SCNA. Only when the modality exists does the question of signal content become relevant. Like SCNA and PSTD, ACNA is a cross-dimensional precondition: the absence of a communication modality prevents any signal content from being transmitted, regardless of representational dimension. ACNA should not be confused with misinterpretation of signals the robot does produce. ACNA concerns whether the robot can produce the signal at all, not whether the user correctly understands what a produced signal means.

Examples: A user waits for a verbal response from a robot that has no speech synthesis capability. A user expects facial expressions from a robot without an articulated face. *Intervention:* The robot can signal through a channel it does have which outputs to expect, so the user stops waiting for one it lacks.

Act-Communication-Signal-Content Mismatch (ACSC). The user misinterprets what the robot signals through its behavior. The robot communicates through a modality it possesses, but the user assigns incorrect meaning to the signal [12, 21]. We integrate ACSC across all three representational dimensions (*What*, *Where*, and *How*). This is a design decision motivated by the observation that the robot's communication output typically does not contain separate subsystems for signaling entity properties, spatial references, and action states. A user who misinterprets a status light (What), misreads a pointing gesture (Where), or mistakes a nod for task completion instead of command receipt (How) experiences the same kind of mismatch. The intervention in all three cases is signal redesign. ACSC presupposes that ACNA is satisfied.

Examples: A robot nods its head as an acknowledgment of receiving a command, but the user interprets the nod as agreement with the instruction's content. A robot pauses to recalculate its path, but the user interprets the pause as the robot being stuck or having finished. *Intervention:* When the user does not respond to a signal as intended, the robot can follow it with a second, clearer one, for example confirm a nod in words.

4.5 Theoretical Grounding and Exhaustiveness

We have taken a deductive approach for the definition of an exhaustive matrix of mismatch type categories. The Sense-Plan-Act standard robot behavior decomposition forms one matrix dimension. The other matrix dimension is deduced from the representational segregation of What, Where, and How, which is rooted in the neurocognitive architecture of human perception. Let us explain this conceptual grounding in the following.

The datasets in Section 5 then test whether this taxonomy covers errors in real interaction data. They are not used to build the categories.

4.5.1 Neurocognitive Grounding of the Representational Dimensions. The three-dimensional decomposition used throughout the taxonomy converges with one of the most robust findings in cognitive neuroscience. The human visual system processes information along functionally distinct pathways that map onto precisely these dimensions. The ventral stream supports object recognition and categorization, identifying entities and their properties [110]. The dorsal stream, rather than forming a single pathway, comprises two functionally and anatomically distinct subsystems [86]. The dorso-dorsal stream mediates spatial processing, localization, and visually guided reaching. The ventro-dorsal stream mediates action understanding, grasping, and the perception of dynamic state changes. This three-way decomposition was established through double dissociations between ventral stream damage (visual agnosia, the inability to recognize objects despite intact vision), dorso-dorsal damage (optic ataxia, the inability to guide reaching movements toward seen objects), and ventro-dorsal damage (limb apraxia, the inability to perform skilled actions despite intact motor function) [11, 33]. Subsequent reviews have confirmed this decomposition [57, 67].

Grounded cognition theory provides the bridge from perceptual architecture to mental models more broadly. Barsalou [7, 8] demonstrated that cognition is grounded in perceptual and motor simulations: conceptual representations partially reactivate the same sensory-motor systems used during direct experience. Because visual perception is organized along What, Where, and How pathways, cognitive representations grounded in these systems inherit this tripartite structure. This extends to language comprehension, which constructs situation models indexed on entities, spatial relations, and causal-action sequences [114], and recruits the same neural pathways [81]. Landau and

Jackendoff [58] demonstrated that the What/Where distinction maps directly onto spatial language: object naming encodes detailed geometric properties (ventral stream), while spatial expressions encode only coarse spatial relations (dorsal stream). In language comprehension, the How dimension extends this mapping to verbs encoding motor interactions and affordances.

Each dimension has a clear test. *What* is the identity and properties of an entity, independent of time. *Where* is its position and which co-present instance is meant. *How* is the dynamic part: whether a state change is perceived, and how an instruction becomes an action. This resolves cases that look ambiguous. The robot may see a cup and label it correctly, so What is satisfied, but miss that the cup was filled, so How fails in relation to a targeted action.

This neurocognitive grounding explains why the taxonomy's representational dimensions are not arbitrary. They reflect the architecture of the human cognitive system that constructs mental models in the first place. The parallel dimensional structure observed across the Sense channels (Sections 4.2.1 and 4.2.2) is predicted by the theoretical view that language comprehension shares key properties with physical object perception [7, 62].

Each mismatch type is most directly motivated by one of the three cognitive mechanisms introduced in Section 2. These are mental state inference, communication and grounding, and trust calibration and affordance perception. Table 1 associates every type with the mechanism from which it predominantly arises. The assignment follows where the mismatch sits in the architecture. Communication types arise from grounding assumptions. Plan-stage reasoning types arise from mental state inference. Object-perception and physical-execution types arise from trust calibration and affordance perception. The mechanisms are not a second basis for the taxonomy. The taxonomy comes from the SPA and What/Where/How axes. The mechanisms only supply the prior evidence that a mismatch at a given location is meaningful. Some types admit a secondary reading. A user who misreads a robot's signal (ACSC) commits both a grounding failure and an act of mental state attribution. We assign each type to its dominant mechanism and note such readings in the type definitions.

Cognitive mechanism	**Associated mismatch types**
Mental state inference and attribution (ToM) [20, 80, 87] *"expecting the robot to share one's knowledge and to reason and plan like a person"*	PWCS, PWMM (world model); PAMT, PAHZ (planning algorithm)
Communication and grounding [16, 34] *"expecting the robot to understand the words and refer to the right things in the shared scene"*	SCNA, SCWT, SCWE, SCHW (Sense communication); ACNA, ACSC (Act communication)
Trust calibration and affordance perception [31, 60] *"expecting what the robot can perceive and do, and what it was built for"*	SOWT, SOWE, SOHW (object sensing); PSTD (design scope); APWE, APHW (physical execution)

Table 1. Mapping from the three cognitive mechanisms introduced in Section 2 to the mismatch types they predominantly motivate. Each type is assigned to its dominant mechanism. By this point all sixteen types have been defined, so the mapping connects the mechanisms in Section 2 to the types in Sections 4.2 to 4.4.

4.5.2 The Exhaustiveness Argument and the Consolidated Matrix. The taxonomy's exhaustiveness follows from the systematic combination of its two scoping dimensions. Following Bailey's property space methodology [6], which formalizes the principle that crossing mutually exclusive, collectively exhaustive dimensions produces a classification that is exhaustive by construction, we consider all three representational dimensions at each SPA processing stage.

Table 2 presents the consolidated matrix. For every combination where a mismatch is possible, the taxonomy provides a dedicated type. Where a cell has no dedicated type, an explicit justification is provided.

<table>
<tr><th colspan="2"></th><th>What
Identity, properties</th><th>Where
Location, individuation</th><th>How
State change, action</th></tr>
<tr><td rowspan="3">Sense</td><td>Object ch.</td><td>SOWT
Not detected / property not perceived</td><td>SOWE
Detected, localized wrong</td><td>SOHW
State change not perceived</td></tr>
<tr><td rowspan="2">Comm. ch.</td><td colspan="3">SCNA – Input modality not available (cross-dim. precondition)</td></tr>
<tr><td>SCWT
Term unknown</td><td>SCWE
Referent ambiguous</td><td>SCHW
Action not mappable</td></tr>
<tr><td rowspan="3">Plan</td><td>Scope</td><td colspan="3">PSTD – Task outside design scope (cross-dim. precondition)</td></tr>
<tr><td>World ctx.</td><td colspan="3">PWCS / PWMM
Current state unknown (PWCS) or past information lost (PWMM)
Integrated: world model is a unified structure; types distinguished by temporal status, not by dimension</td></tr>
<tr><td>Algorithm</td><td colspan="2">Consumed from PWCS; strategic use classified as PAMT</td><td>PAMT / PAHZ
Method / horizon</td></tr>
<tr><td rowspan="3">Act</td><td>Physical</td><td>Motor system operates over coordinates, not object identities</td><td>APWE
Cannot reach position</td><td>APHW
Manipulation fails</td></tr>
<tr><td rowspan="2">Comm. out</td><td colspan="3">ACNA – Modality does not exist (cross-dim. precondition)</td></tr>
<tr><td colspan="3">ACSC
Signal content misread
Integrated by design: signals carry meaning as indivisible units</td></tr>
</table>

Table 2. The MMM matrix: sixteen mismatch types across representational dimensions (columns) and processing stages (rows). The two axes are colour-coded: SPA stages along the left edge (blue for Sense, orange for Plan, green for Act) and representational dimensions along the top edge (purple for What, teal for Where, pink for How). The same axis colours are used in Figure 2 to mark the diagnostic cell. Body cells are white. Cells spanning multiple columns contain either cross-dimensional preconditions (SCNA, PSTD, ACNA) or integrated types (PWCS/PWMM, ACSC). Grey cells mark dimensions where the processing stage has no type at all, because the architecture lacks the relevant representational machinery. The full diagnostic reference for each type is provided in Appendix A.

A general principle governs these justifications: the dimensional resolution follows the robot's operational architecture. Where the robot has architecturally distinct subsystems along the representational dimensions, separate types are warranted. Where the robot's architecture does not decompose along these dimensions, integration is the correct modeling choice.

The justifications for the special cases follow the SPA order. In the Sense communication channel, SCNA functions as a cross-dimensional precondition: the absence of an input modality blocks all three dimensions simultaneously. In Plan, PSTD serves the same role: a task outside the design scope fails regardless of what entities, locations, or actions are involved. The two World Context types, PWCS and PWMM, integrate across all three dimensions because the world model stores entity properties, spatial relations, and state information as a unified structure. The Planning Algorithm row contains no dedicated What or Where types because the algorithm consumes this information from the world model rather than representing it independently. Strategic decisions about object selection or spatial ordering are classified as PAMT. In Act Physical, the What dimension has no type because the motor execution system operates over coordinates and force parameters, not over object identities. In Act Communication Output, ACNA is the output-side counterpart to SCNA and functions as a cross-dimensional precondition. ACSC integrates all three dimensions because the robot's communication output produces signals as indivisible units rather than maintaining separate subsystems for each dimension.

4.6 Scope Boundaries

The taxonomy focuses on systematic cognitive mismatches during collaborative task execution. Several categories of interaction phenomena fall outside this scope, either because they involve different cognitive mechanisms or because they are addressed by dedicated frameworks.

Social-emotional dynamics. We exclude social-emotional relationship building because, in the task-oriented collaborative contexts we address, success primarily depends on cognitive coordination, though the relative importance of social-emotional factors varies by context [10, 61].

Social pragmatics. We exclude mismatches arising from sarcasm, irony, humor, or politeness conventions. While a robot that interprets "great job" literally when the user intended sarcasm experiences a communicative failure, the root cause lies in social signal interpretation rather than in the task-coordination mechanisms our taxonomy addresses. These social-communicative mismatches are covered by dedicated frameworks such as Tian and Oviatt's taxonomy of social errors in HRI [106]. Our taxonomy addresses the task-oriented complement: when the robot fails to correctly process what the user wants done, rather than how the user socially frames it.

Normative expectations. We exclude assumptions about what a robot should or should not do based on social, ethical, or policy constraints rather than capability [65]. In practice, normative and capability constraints can produce similar behavioral signatures (a robot that cannot reach a location and a robot that is not permitted to enter it both fail to move there), but the root causes and appropriate interventions differ. Normative mismatches represent an important adjacent area that the current taxonomy does not address.

Cultural variation. We exclude cultural variations because the fundamental cognitive mechanisms that produce mismatches remain consistent across populations, though their expression varies [63].

Multi-agent scenarios. We also exclude interactions involving multiple humans coordinating with a single robot, or scenarios where the identity of the "user" is ambiguous. The taxonomy assumes a single human interacting with a single robot. Extending the framework to multi-party interaction, where users may hold mutually inconsistent mental models that the robot must reconcile, is an important direction but introduces coordination challenges that go beyond the scope of individual cognitive mismatches.

4.7 Worked Example

To illustrate how the taxonomy diagnoses a concrete interaction error, we walk through a dialogue from the TEACh dataset (Section 5). In TEACh, a human *commander* instructs a *driver* who simulates robot-like sensing and action.

The commander opens with a clear task: *"Can you boil a potato for me."* The driver responds without hesitation: *"of course."* This first exchange already rules out one diagnosis. If the instruction had not been understood, we would expect a clarification question or silence, not a confident affirmation.

The commander then adds context: *"You can find one on the chair by the table, and there is a pot on the hob already. You may need to fill it with water."* The driver navigates directly to the potato and picks up a knife. So far the driver has located the correct object and selected a tool that is plausibly task-relevant. There is no perceptual or referential problem visible.

Then the driver slices the potato (Figure 3a) and places a single slice into the pot (Figure 3b). At this point, an observer would naturally describe the failure as "the robot did not understand the instruction." The naive diagnosis is therefore Sense-Communication-How Mismatch (SCHW): the driver supposedly failed to extract what action was required.



(a) The driver slices the potato on the table.



(b) A single slice is placed in the pot.



(c) After correction, the whole potato is in the pot.

Fig. 3. Worked example from TEACh. The driver sliced the potato (a) and placed a single slice in the pot (b), confident that this was the correct execution. Four minutes later, the commander corrected the misalignment, and the driver then placed the whole potato in the pot (c).

Four minutes later, the driver itself rules this diagnosis out. It asks: *"how many potatoes do you want boiled?"* This question only makes sense if the driver understood that the goal was to boil potatoes. The commander then notices the misalignment and corrects: *"Oh could you please boil a whole potato, not a slice my apologies."* The driver now places the whole potato in the pot (Figure 3c).

The mismatch lies further downstream, in the planning algorithm. The driver's schema for "boil a potato" includes a slicing step that the commander did not intend. This is a Plan-Algorithm-Method Mismatch (PAMT). The robot understands the task and plans at the appropriate level of abstraction, but applies a different method than the user expects.

SCHW and PAMT call for different interventions. SCHW would call for vocabulary clarification or instruction rephrasing. PAMT calls for the commander to specify method constraints ("boil the potato whole, do not slice it") or for the system to confirm its intended steps before executing irreversible actions.

The diagnostic value becomes clearer in contrast with a similar case from the same dataset. The commander instructs *"put all ladle in one drawer."* The driver responds *"laddle?"*, then *"what thats?"*, then *"english word?"* The commander must explain *"like spoon"* before the driver can act at all. Here the failure is visible immediately, in the first response. The driver cannot proceed because the term itself is unknown. This is a Sense-Communication-What Mismatch (SCWT).

Both cases would, on a surface level, be described as "the robot didn't understand." The taxonomy reveals opposite root causes. In the potato case, the driver understood immediately and acted with confidence, but applied the wrong method. In the ladle case, the driver could not act at all because the term was unknown. The interventions follow accordingly: constraining the method versus expanding the vocabulary.

4.8 Cascading and Co-Occurring Mismatches

The taxonomy treats each error as a single classification, identifying the root cause at the earliest processing stage. This is a deliberate design choice: in cascading situations, a failure at an earlier stage triggers downstream failures that would not have occurred independently. Consider a user who says "put the carafe on the counter." If the robot does not

know the term "carafe" (SCWT), it cannot bind the reference (SCWE would not arise), cannot plan the action (PAHZ is preempted), and cannot execute it (APWE/APHW become irrelevant). From the user's perspective, the robot simply "did nothing," but the root cause is a single vocabulary gap at the earliest processing stage. Identifying this root cause is the most productive starting point for intervention.

However, mismatches can also *co-occur* independently when they arise from different processing stages. A robot might simultaneously have a limited planning horizon (PAHZ) and weak grip force (APHW). In a task like "set the table," both mismatches manifest through separate mechanisms. The taxonomy as currently applied does not fully capture such co-occurring mismatches, as it assigns a single type per error instance. This means that in cases of genuine co-occurrence, the annotation captures the earliest or most salient mismatch but may miss independent failures at later stages. We consider this an acceptable trade-off for a first-generation annotation tool: single-type classification keeps the annotation process tractable and achieves consistent agreement across annotators, while the taxonomy's structure preserves the conceptual distinction between cascading and co-occurring patterns. Extending the annotation scheme to support multi-label classification is an important direction for future work.

The root-cause principle also resolves cases where the manifestation point differs from the origin of the mismatch. For example, a robot announces "task complete" when the task is not finished. The user correctly interprets the signal (not ACSC) but forms a false belief. The root cause is that the robot's world model incorrectly represents the task as complete (PWCS), and the communication faithfully transmits this incorrect belief. Classifying at the root cause (PWCS) rather than the manifestation point (communication) identifies the most productive intervention: improving the robot's task-state tracking rather than its communication design.

A related pattern is *masking*: a more salient mismatch hides a less visible one. If the robot cannot reach a shelf (APWE), the user never discovers that it also lacks the dexterity to handle the objects there (APHW). Masking complicates diagnosis because resolving one mismatch may reveal another, and it implies that mismatch distributions observed in user studies may systematically undercount types hidden behind more salient failures.

5 Application: Demonstrating the Taxonomy on Two Contrasting Datasets

To demonstrate that the taxonomy can be applied to real interaction data, we use two contrasting datasets: task-oriented manipulation in simulation (TEACh [75]) and social therapeutic dialogue with physical robots (SHREC [59]). This study is an application and reliability check rather than a validation: it tests whether the taxonomy covers errors arising in real interaction data and whether independent annotators apply it consistently. It makes no claim of ecological completeness, and it does not benchmark performance against an external ground truth, since no labelled ground truth for mental model mismatches exists.

5.1 Datasets

TEACh (Task-Driven Embodied Agents That Chat) simulates collaborative HRI in AI2-THOR virtual environments [56, 75]. A *commander* with a high-level goal instructs a *driver* who operates under robot-like constraints (restricted field of view, limited object recognition, simplified manipulation), making the driver a viable robot proxy. Since TEACh provides no predefined error annotations, we identified interaction breakdowns from the dataset ourselves. The breakdown was anchored to the first turn at which a clear mismatch signal was visible, typically a driver response. Signals included clarification questions ("where?", "which?"), capability statements ("I can't"), wrong-object actions, repeated failures of the same action (three or more), explicit confusion markers, and prolonged silent search. We annotated the commander turn only when the commander contradicted or corrected mid-dialogue, or when the commander misread a driver

signal. When multiple signals appeared, we used the earliest one. Sampling continued until 150 valid breakdowns were obtained.

SHREC (Social Human-Robot Embodied Conversation) contains authentic therapeutic interactions with physical Jibo robots in wellness coaching contexts [59]. We use the Wellness-Home subset, in which Jibo was deployed in participants' homes as a positive psychology coach communicating through speech, facial expressions, and head movements (no manipulation). Among the SHREC source studies, Wellness-Home offered the most complete and consistent error annotations and the broadest distribution of failure modes, making it the most suitable subset for testing taxonomy coverage. SHREC provides predefined error annotations; we randomly sampled annotations and classified each one using the taxonomy. Of 215 sampled error annotations, 57 had to be excluded because the underlying transcript data was corrupted (turn-mixing, truncation, or missing tags), the system experienced session-level malfunctions (freezing, looping greetings, stuck states), or the annotated error fell outside the captured transcript window. Sampling continued until 150 valid cases were obtained.

5.2 Annotator Agreement

To check whether the taxonomy yields consistent classifications across annotators, we ran an agreement check on both datasets. The first author, a PhD researcher involved in developing the taxonomy, served as the primary annotator and coded both datasets. For the agreement check, two second annotators independently classified the IRR sets, one per dataset: one coded the TEACh set, the other the SHREC set. Both were master's-level research assistants in computer science, and neither had been involved in developing the taxonomy. A calibration phase preceded the agreement pass, in which all three annotators jointly discussed representative cases drawn from outside the IRR test set to establish a shared understanding of the taxonomy. Independent classification then proceeded on each IRR set, with each annotator coding in isolation. Each error was presented as a dialogue excerpt with contextual information about the robot's state and capabilities. Annotators assigned exactly one mismatch type per error or marked it as unclassifiable if none of the sixteen types applied. We computed Cohen's κ [19] on all sixteen type labels plus unclassifiable as nominal categories. Each reported agreement value is computed between the primary annotator and one independent annotator applying the taxonomy from its definitions alone.

On TEACh ($n = 150$), observed agreement was 86.7% with $\kappa = 0.813$. On SHREC ($n = 150$), observed agreement was 89.3% with $\kappa = 0.842$. The agreement values suggest the taxonomy yields consistent classifications across annotators in both settings, with κ in the same range despite very different distributions. The reported κ values are computed on the independent annotations. Disagreements were afterward resolved through discussion, but this reconciliation does not enter the reported agreement.

5.3 Results and Observations

On TEACh, the analysis subset of 150 errors yielded a distribution (Figure 4, Table 3a) concentrated in Sense mismatches (70.7%), with SOWT dominating at 48.0% and SCWE as the second most frequent type (12.7%). Plan accounts for 11.3% and Act for 10.7%. Unclassifiable cases account for 7.3%. Three types are absent: SCNA, APWE, and ACNA.

On SHREC, the distribution (Figure 4, Table 3b) is concentrated in Act mismatches (52.7%), driven entirely by ACSC. SCHW is the second most frequent type at 18.7%, followed by PAMT at 8.0%. Sense accounts for 26.7% and Plan for 12.0%. Unclassifiable cases account for 8.7%. Five types are absent: SCWT, SCWE, APWE, APHW, and ACNA.

In TEACh, the communication channel contributes substantially through SCWE: even when objects are visible, commanders frequently reference them in ways the driver cannot resolve. In SHREC, nearly all ACSC cases concern

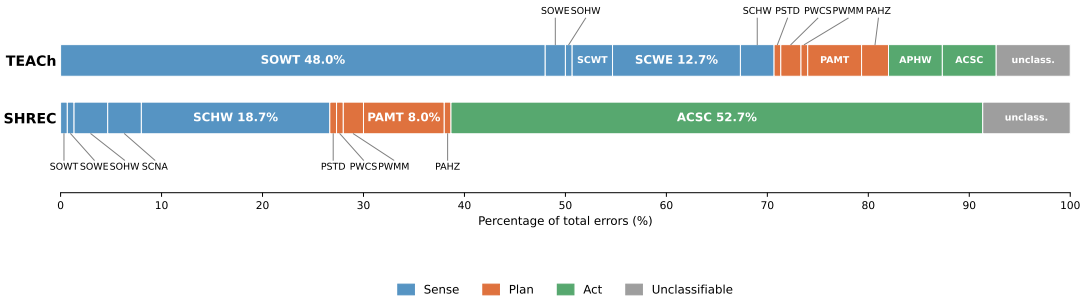

Fig. 4. Distribution of MMM types across TEACh and SHREC, shown as 100% stacked bars over all observed errors. Segments are coloured by SPA category (Sense, Plan, Act); the gray segment at the right of each bar represents errors that could not be classified by the taxonomy. Percentages are relative to the total number of identified errors per dataset. Types not observed in a given dataset are absent from its bar: in TEACh, SCNA, APWE and ACNA did not occur; in SHREC, SCWT, SCWE, APWE, APHW and ACNA did not occur. The contrast between the two datasets is visible at a glance: TEACh is dominated by Sense mismatches (SOWT and SCWE), SHREC by Act (ACSC) and the communication content type SCHW. Detailed counts are reported in Table 3.

SPA	Type	Count	%
Sense	SOWT	72	48.0
	SOWE	3	2.0
	SOHW	1	0.7
	SCNA	–	–
	SCWT	6	4.0
	SCWE	19	12.7
	SCHW	5	3.3
Plan	PSTD	1	0.7
	PWCS	3	2.0
	PWMM	1	0.7
	PAMT	8	5.3
	PAHZ	4	2.7
Act	APWE	–	–
	APHW	8	5.3
	ACNA	–	–
	ACSC	8	5.3
Classified		139	92.7
Unclassifiable		11	7.3
Total		**150**	

(a) TEACh

SPA	Type	Count	%
Sense	SOWT	1	0.7
	SOWE	1	0.7
	SOHW	5	3.3
	SCNA	5	3.3
	SCWT	–	–
	SCWE	–	–
	SCHW	28	18.7
Plan	PSTD	1	0.7
	PWCS	1	0.7
	PWMM	3	2.0
	PAMT	12	8.0
	PAHZ	1	0.7
Act	APWE	–	–
	APHW	–	–
	ACNA	–	–
	ACSC	79	52.7
Classified		137	91.3
Unclassifiable		13	8.7
Total		**150**	

(b) SHREC

Table 3. Distribution of MMM types in TEACh (left) and SHREC (right) on the analysis subsets (n = 150 for both). Dashes indicate types not observed. Percentages are relative to the total errors in each analysis subset.

the user perceiving the robot's delayed or absent response as a signal carrying meaning the robot did not intend. This pattern would not have surfaced in TEACh, where text-based exchange removes timing as a meaningful dimension of communication.

Across both datasets combined, fourteen of the sixteen taxonomy types appear at least once. APWE and ACNA are absent from both datasets, but for principled reasons. AI2-THOR discretises physical interaction with no continuous

notion of reach, and Jibo has no manipulation capability, ruling out APWE in both. ACNA is similarly precluded. TEACh is text-based on both sides, and SHREC participants engage in wellness coaching with a robot they have been told can speak, so neither setting affords the kind of capability misassumption that would generate ACNA. Both types would be expected to surface in less guided deployments, particularly with mobile manipulator robots in unconstrained physical environments or with robots lacking a modality the user assumes is available. Even within types that do appear, the masking pattern introduced in Section 4.8 likely skews counts further: later-stage types (e.g., APHW after APWE, SCHW after SCWT) can remain hidden behind more salient earlier-stage failures, so the observed frequencies should be read as lower bounds rather than direct estimates of true incidence.

The unclassifiable cases warrant a closer look. The proportion is broadly similar in both datasets (7.3% in TEACh, 8.7% in SHREC), but their composition differs in ways that map onto the scope boundaries laid out in Section 4.6. In SHREC, the surviving unclassifiable cases fall into three patterns: social-form-only violations where Jibo's plan and perception are intact but the script lacks an expected politeness beat (no acknowledgment of a participant's disclosure); cases where a topically-aware acknowledgment is present but the user perceives it as too shallow for the substance of their share; and post-activity cases where the user's signal arrives after Jibo's session-ending utterance, leaving open whether the system ignored the input or had architecturally ended the session. The first pattern reflects the social-pragmatics scope boundary; the third reflects an observability limit of transcript-only annotation. In TEACh, unclassifiable cases concentrate on simulator-level artefacts (physics edge cases, object-snap behavior) and on cases where the commander's own beliefs about the environment were incorrect, which the taxonomy's user-robot scope does not cover.

6 Discussion

6.1 Implications for Research and Design

It is important to note that the mismatches identified by this taxonomy are not user errors. They are misalignments between the user's mental model and the robot's operational model, and responsibility for these misalignments is shared. The robot designer is always implicitly present in HRI: the robot's vocabulary, planning algorithms, and physical capabilities reflect design decisions that shape which mismatches arise [68, 73]. A Sense-Communication-What Mismatch, for instance, indicates a gap between the vocabulary the designer implemented and the vocabulary users naturally employ. A Plan-Algorithm-Method Mismatch indicates that the designer's chosen method does not match the method users expect. The taxonomy is agnostic about attribution: it identifies where in the processing pipeline the mental model and the operational model diverge, and this diagnosis can inform design improvements just as readily as user onboarding.

Currently, researchers report interaction failures using varied, often ad hoc descriptions such as "the robot didn't understand" or "communication breakdown." The taxonomy offers a shared vocabulary: a "communication breakdown" can now be identified as a SCNA (unavailable input modality), a SCWT (unknown term), a SCWE (ambiguous reference), or a SCHW (unmappable instruction), each with different theoretical implications and empirical signatures. This could enable cross-study comparison that is currently difficult when every study defines its own failure categories.

The taxonomy also opens up vulnerability prediction as a research direction. The SPA structure suggests, for example, that systems with manipulation-heavy tasks may show elevated Act mismatch rates, longitudinal interactions may show concentrations of Plan-World-Memory Mismatches, and dynamic environments may show Sense-Object-How Mismatches. These are testable hypotheses that researchers could pre-register in future studies, rather than discovering

patterns only post hoc. If multiple research groups adopt the same categories, the field can begin to identify which mismatch types are most persistent across systems, which are most amenable to design interventions, and which dominate in specific application domains.

Beyond research, the taxonomy can also support practical system development. When a system fails in user testing, the natural tendency is to focus on the most visible failure mode. The SPA structure encourages designers to systematically check whether less obvious mismatch types might also be contributing, and whether failures concentrate in perception, reasoning, or execution. What makes this practical is the operationalization through formal definitions with adjacent-type contrasts, which give designers concrete decision rules. The taxonomy also encourages thinking about mismatches *before* deployment. By reviewing the sixteen types against a system's known architecture, designers can identify likely vulnerability profiles in advance and design interfaces that mitigate them. A system prone to Plan-Algorithm-Horizon Mismatches can be designed to elicit step-by-step instructions. A system prone to Plan-World-Memory Mismatches can communicate its memory limitations during onboarding.

The most advanced application of the taxonomy would be an internal estimation on the robot side that estimates human mental model mismatches based on the sequence of interactions. Such a system might involve clarifying questions and explanations from the robot, aimed at actively identifying and correcting mismatches. An interaction "supervisor" of this kind has been investigated by Richter et al. [85] in a simple interaction game scenario.

6.2 Limitations of the Taxonomy

The taxonomy presupposes that the robot's operation can be decomposed into Sense, Plan, and Act stages. This assumption holds for most classical and modular robot architectures, but it becomes increasingly difficult to sustain for end-to-end trained systems, where perception, reasoning, and action selection collapse into a single learned mapping. With the growing prevalence of LLM-based and vision-language-action models in robotics, this is not a peripheral concern but a question of how the framework remains relevant as architectures change. The representational dimensions (What, Where, How) are grounded in human cognitive architecture rather than in the robot's implementation, and remain applicable: a user can still misjudge what such a system recognizes, where it locates entities, or how it expects state to change. The SPA stages, by contrast, lose their direct architectural referent, which weakens the diagnostic-to-intervention link the taxonomy provides for modular systems. Knowing that a mismatch lies in "planning" is less actionable when there is no separable planning component to redesign. This is less a rejection of the taxonomy's categories than a reminder that it describes mismatches at the level of an operational model, and the operational model has to be available for the mapping to apply.

Beyond the question of how many types to assign, there is the question of what to do once a type is identified. The taxonomy identifies *what* went wrong and *where* in the architecture it occurred, but it does not prescribe specific interventions. The connection between mismatch types and cognitive mechanisms (Section 2) suggests intervention directions, but validated strategies for each type remain to be developed. Knowing that an error is a Sense-Communication-What Mismatch does not yet tell a designer whether vocabulary expansion, synonym resolution, visual grounding, or user rephrasing prompts will be most effective.

The intentional scope boundaries laid out in Section 4.6 mean the taxonomy may miss relationally motivated mismatches in contexts where social-emotional dynamics are primary, such as companion robots for elderly care or therapeutic applications. Critical perspectives in HRI further caution that taxonomic approaches risk detaching interaction failures from their relational and sociocultural context [40, 113]. Our taxonomy addresses this concern

partially through the shared-responsibility framing and the explicit role of design decisions, but it does not capture the full relational dynamics that shape how mismatches are experienced by individual users.

All mismatch types were inferred from behavioral evidence rather than direct cognitive measurement. We observe what the user said and what the robot did, and infer the underlying mismatch from the gap between the two. This creates uncertainty about the actual mental processes involved. Complementary methods such as think-aloud protocols, eye tracking, or physiological measures could provide converging evidence.

6.3 Future Work

Several directions follow from the taxonomy's current state. The most immediate practical need is linking each mismatch type to validated intervention strategies, transforming the taxonomy from a diagnostic tool into a complete design framework. For each type, controlled studies could compare candidate interventions and establish which approaches are most effective under which conditions.

Extending the annotation scheme to support multi-label classification would enable systematic study of co-occurring and cascading mismatches. The temporal processing chain of the taxonomy predicts that earlier-stage mismatches should cascade more frequently, which is a testable hypothesis. Application to full-spectrum embodied interactions, involving mobile manipulator robots with speech, gesture, gaze, and physical interaction, would test the taxonomy in settings where all sixteen types can manifest within a single session. Such settings would also enable testing context-dependent predictions: whether Sense-Communication-Not-Available Mismatches concentrate with robots that cannot receive non-verbal input, whether Sense-Object-How Mismatches concentrate in dynamic environments, and whether Act-Communication-Not-Available Mismatches concentrate with minimally expressive platforms.

Longitudinal studies could reveal how mismatch distributions evolve as users gain experience. Some types may diminish with exposure (Sense-Object-What Mismatches as users learn camera limitations), while others may persist (Plan-Algorithm-Method Mismatches if users cannot observe the robot's decision process) or even increase (Plan-World-Memory Mismatches as users build longer interaction histories).

7 Conclusion

This work addresses mental model mismatches in HRI through a theoretically grounded taxonomy of sixteen mismatch types. Drawing from cognitive neuroscience, psychology, and communication research, we organized mismatches along the Sense-Plan-Act architecture and established their exhaustiveness through a dimensional decomposition that crosses three neurocognitively grounded representational dimensions (What, Where, How) with three processing stages. Each type is equipped with formal definitions and diagnostic examples that support systematic annotation. Applied to two contrasting datasets, the taxonomy produced nearly inverse distributions (task-oriented manipulation concentrated in Sense mismatches, social dialogue concentrated in Act mismatches) while reaching consistent annotator agreement in both settings, indicating that it discriminates between interaction contexts rather than collapsing failures into a single category.

The taxonomy is designed as a task-independent tool for the HRI community, applicable to collaborative interaction contexts where humans and robots work together toward shared goals. By providing precise terminology for distinct cognitive mismatches, it enables researchers and designers to diagnose specific interaction failures and provides a basis for developing targeted interventions, rather than treating all problems as generic "communication failures."

Acknowledgments

Phillip Richter gratefully acknowledges the financial support from Honda Research Institute Europe for the project 'Reducing Mental Model Mismatch for Cooperative Robot Teaching'.

References

[1] Robert B Allen. 1997. Mental models and user models. In *Handbook of human-computer interaction*. Elsevier, 49–63.
[2] Craig A Anderson, Mark R Lepper, and Lee Ross. 1980. Perseverance of social theories: The role of explanation in the persistence of discredited information. *Journal of personality and social psychology* 39, 6 (1980), 1037.
[3] Robert W Andrews, J Mason Lilly, Divya Srivastava, and Karen M Feigh. 2023. The role of shared mental models in human-AI teams: a theoretical review. *Theoretical Issues in Ergonomics Science* 24, 2 (2023), 129–175.
[4] Mira Ariel. 2014. *Accessing noun-phrase antecedents*. Routledge.
[5] Ronald C Arkin. 1998. *Behavior-based robotics*. MIT press.
[6] Kenneth D Bailey. 1994. *Typologies and taxonomies: An introduction to classification techniques*. Vol. 12. Sage.
[7] Lawrence W Barsalou. 1999. Perceptual symbol systems. *Behavioral and brain sciences* 22, 4 (1999), 577–660.
[8] Lawrence W Barsalou. 2008. Grounded cognition. *Annu. Rev. Psychol.* 59, 1 (2008), 617–645.
[9] James M Berzuk and James E Young. 2023. Clarifying social robot expectation discrepancy: developing a framework for understanding how users form expectations of social robots. In *Companion of the 2023 ACM/IEEE international conference on human-robot interaction*. 231–233.
[10] Timothy W Bickmore and Rosalind W Picard. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)* 12, 2 (2005), 293–327.
[11] Ferdinand Binkofski and Laurel J Buxbaum. 2013. Two action systems in the human brain. *Brain and language* 127, 2 (2013), 222–229.
[12] Cynthia Breazeal. 2003. Emotion and sociable humanoid robots. *International journal of human-computer studies* 59, 1-2 (2003), 119–155.
[13] Rodney Brooks. 2003. A robust layered control system for a mobile robot. *IEEE journal on robotics and automation* 2, 1 (2003), 14–23.
[14] Judee K Burgoon, Joseph A Bonito, Paul Benjamin Lowry, Sean L Humpherys, Gregory D Moody, James E Gaskin, and Justin Scott Giboney. 2016. Application of expectancy violations theory to communication with and judgments about embodied agents during a decision-making task. *International Journal of Human-Computer Studies* 91 (2016), 24–36.
[15] Joyce Y Chai, Lanbo She, Rui Fang, Spencer Ottarson, Cody Littley, Changsong Liu, and Kenneth Hanson. 2014. Collaborative effort towards common ground in situated human-robot dialogue. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. 33–40.
[16] Herbert H Clark and Susan E Brennan. 1991. Grounding in communication. (1991).
[17] Herbert H Clark and Edward F Schaefer. 1989. Contributing to discourse. *Cognitive science* 13, 2 (1989), 259–294.
[18] Herbert H Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition* 22, 1 (1986), 1–39.
[19] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.
[20] Daniel C Dennett. 1989. *The intentional stance*. MIT press.
[21] Anca D Dragan, Kenton CT Lee, and Siddhartha S Srinivasa. 2013. Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 301–308.
[22] John Duncan. 2025. Construction and use of mental models: Organizing principles for the science of brain and mind. *Neuropsychologia* 207 (2025), 109062.
[23] Mica R Endsley. 1995. A taxonomy of situation awareness errors. *Human factors in aviation operations* 3, 2 (1995), 287–292.
[24] Nicholas Epley, Boaz Keysar, Leaf Van Boven, and Thomas Gilovich. 2004. Perspective taking as egocentric anchoring and adjustment. *Journal of personality and social psychology* 87, 3 (2004), 327.
[25] Nicholas Epley, Adam Waytz, and John T Cacioppo. 2007. On seeing human: a three-factor theory of anthropomorphism. *Psychological review* 114, 4 (2007), 864.
[26] Connor Esterwood and Lionel P Robert. 2023. The theory of mind and human–robot trust repair. *Scientific Reports* 13, 1 (2023), 9877.
[27] Kerstin Fischer. 2011. How people talk with robots: Designing dialog to reduce user uncertainty. *Ai Magazine* 32, 4 (2011), 31–38.
[28] Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. 2003. A survey of socially interactive robots. *Robotics and autonomous systems* 42, 3-4 (2003), 143–166.
[29] Jacob Fraden. 2004. *Handbook of modern sensors: physics, designs, and applications*. Springer.
[30] Angela D Friederici. 2002. Towards a neural basis of auditory sentence processing. *Trends in cognitive sciences* 6, 2 (2002), 78–84.
[31] James J Gibson. 2014. *The ecological approach to visual perception: classic edition*. Psychology press.
[32] Daniel T Gilbert and Patrick S Malone. 1995. The correspondence bias. *Psychological bulletin* 117, 1 (1995), 21.
[33] Melvyn A Goodale and A David Milner. 1992. Separate visual pathways for perception and action. *Trends in neurosciences* 15, 1 (1992), 20–25.
[34] Herbert P Grice. 1975. Logic and conversation. In *Speech acts*. Brill, 41–58.
[35] Barbara J Grosz, Aravind Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational linguistics* 21, 2 (1995), 203–225.
[36] Peter Hagoort. 2013. MUC (memory, unification, control) and beyond. *Frontiers in psychology* 4 (2013), 416.

[37] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human factors* 53, 5 (2011), 517–527.

[38] Peter A Hancock, Theresa T Kessler, Alexandra D Kaplan, John C Brill, and James L Szalma. 2021. Evolving trust in robots: specification through sequential and comparative meta-analyses. *Human factors* 63, 7 (2021), 1196–1229.

[39] Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena* 42, 1-3 (1990), 335–346.

[40] Katherine Harrison and Ericka Johnson. 2023. Affective corners as a problematic for design interactions. *ACM Transactions on Human-Robot Interaction* 12, 4 (2023), 1–9.

[41] Thomas Hellström and Suna Bensch. 2018. Understandable robots-what, why, and how. *Paladyn, Journal of Behavioral Robotics* 9, 1 (2018), 110–123.

[42] Lukas Hindemith, Jan Philip Göpfert, Christiane B Wiebel-Herboth, Britta Wrede, and Anna-Lisa Vollmer. 2021. Why robots should be technical: Correcting mental models through technical architecture concepts. *Interaction Studies* 22, 2 (2021), 244–279.

[43] Lukas Hindemith, Christiane B Wiebel-Herboth, Heiko Wersing, Britta Wrede, and Anna-Lisa Vollmer. 2025. Improving HRI through robot architecture transparency. *International Journal of Social Robotics* (2025), 1–21.

[44] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors* 57, 3 (2015), 407–434.

[45] Guy Hoffman. 2019. Evaluating fluency in human–robot collaboration. *IEEE Transactions on Human-Machine Systems* 49, 3 (2019), 209–218.

[46] Shanee Honig and Tal Oron-Gilad. 2018. Understanding and resolving failures in human-robot interaction: Literature review and model development. *Frontiers in psychology* 9 (2018), 861.

[47] Aike C Horstmann and Nicole C Krämer. 2019. Great expectations? Relation of previous experiences with social robots in real life or in the media and expectancies based on qualitative and quantitative assessment. *Frontiers in psychology* 10 (2019), 939.

[48] Jung-Hoon Hwang, KangWoo Lee, and Dong-Soo Kwon. 2005. The role of mental model and shared grounds in human-robot interaction. In *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.* IEEE, 623–628.

[49] Shu Jiang and Ronald C Arkin. 2015. Mixed-initiative human-robot interaction: definition, taxonomy, and survey. In *2015 IEEE International conference on systems, man, and cybernetics.* IEEE, 954–961.

[50] Philip N Johnson-Laird. 1980. Mental models in cognitive science. *Cognitive science* 4, 1 (1980), 71–115.

[51] Philip Nicholas Johnson-Laird. 1983. *Mental models: Towards a cognitive science of language, inference, and consciousness.* Number 6. Harvard University Press.

[52] Philip N Johnson-Laird. 2010. Mental models and human reasoning. *Proceedings of the National Academy of Sciences* 107, 43 (2010), 18243–18250.

[53] Boaz Keysar, Dale J Barr, Jennifer A Balin, and Jason S Brauner. 2000. Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological science* 11, 1 (2000), 32–38.

[54] David E Kieras and Susan Bovair. 1984. The role of a mental model in learning to operate a device. *Cognitive science* 8, 3 (1984), 255–273.

[55] Sara Kiesler. 2005. Fostering common ground in human-robot interaction. In *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.* IEEE, 729–734.

[56] Eric Kolve et al. 2017. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474* (2017).

[57] Dwight J Kravitz, Kadharbatcha S Saleem, Chris I Baker, and Mortimer Mishkin. 2011. A new neural framework for visuospatial processing. *Nature Reviews Neuroscience* 12, 4 (2011), 217–230.

[58] Barbara Landau and Ray Jackendoff. 1993. "What" and "where" in spatial language and spatial cognition. *Behavioral and Brain Sciences* 16, 2 (1993), 217–265.

[59] Dong Won Lee, Yubin Kim, Denison Guvenoz, Sooyeon Jeong, Parker Malachowsky, Louis-Philippe Morency, Cynthia Breazeal, and Hae Won Park. 2025. The Human Robot Social Interaction (HSRI) Dataset: Benchmarking Foundational Models' Social Reasoning. *arXiv e-prints* (2025), arXiv–2504.

[60] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[61] Iolanda Leite, Carlos Martinho, and Ana Paiva. 2013. Social robots for long-term interaction: a survey. *International Journal of Social Robotics* 5, 2 (2013), 291–308.

[62] Alvin M Liberman and Ignatius G Mattingly. 1985. The motor theory of speech perception revised. *Cognition* 21, 1 (1985), 1–36.

[63] Velvetina Lim, Maki Rooksby, and Emily S Cross. 2021. Social robots on a global stage: establishing a role for culture during human–robot interaction. *International Journal of Social Robotics* 13, 6 (2021), 1307–1333.

[64] Manja Lohse. 2011. Bridging the gap between users' expectations and system evaluations. In *2011 RO-MAN.* IEEE, 485–490.

[65] Bertram Malle, Kerstin Fischer, James Young, Ajung Moon, and Emily Collins. 2020. Trust and the discrepancy between expectations and actual capabilities of social robots. In *Human-robot interaction: Control, analysis, and design.* Cambridge Scholars Press, 1–23.

[66] Ross Mead and Maja J Matarić. 2017. Autonomous human–robot proxemics: socially aware navigation based on interaction potential. *Autonomous Robots* 41, 5 (2017), 1189–1201.

[67] A David Milner and Melvyn A Goodale. 2008. Two visual systems re-viewed. *Neuropsychologia* 46, 3 (2008), 774–785.

[68] David A Mindell. 2015. *Our robots, ourselves: Robotics and the myths of autonomy.* Viking.

[69] Marvin Minsky. 1974. A framework for representing knowledge. (1974).

[70] Wenxuan Mou, Martina Ruocco, Debora Zanatto, and Angelo Cangelosi. 2020. When would you trust a robot? a study on trust and theory of mind in human-robot interactions. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN).* IEEE, 956–962.

[71] Robin R Murphy. 2019. *Introduction to AI robotics*. MIT press.
[72] Jakob Nielsen. 1994. *Usability engineering*. Morgan Kaufmann.
[73] Don Norman. 2013. *The design of everyday things: Revised and expanded edition*. Basic books.
[74] Donald A Norman. 1999. Affordance, conventions, and design. *interactions* 6, 3 (1999), 38–43.
[75] Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. 2022. Teach: Task-driven embodied agents that chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 2017–2025.
[76] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human factors* 39, 2 (1997), 230–253.
[77] Raja Parasuraman, Thomas B Sheridan, and Christopher D Wickens. 2000. A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans* 30, 3 (2000), 286–297.
[78] Betsy Phillips and Hongxin Zhao. 1993. Predictors of assistive technology abandonment. *Assistive technology* 5, 1 (1993), 36–45.
[79] Aaron Powers and Sara Kiesler. 2006. The advisor robot: tracing people's mental model from a robot's physical attributes. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*. 218–225.
[80] David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences* 1, 4 (1978), 515–526.
[81] Friedemann Pulvermüller. 2005. Brain mechanisms linking language and action. *Nature reviews neuroscience* 6, 7 (2005), 576–582.
[82] Preeti Ramaraj. 2021. Robots that help humans build better mental models of robots. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. 595–597.
[83] James Reason. 1990. *Human error*. Cambridge university press.
[84] Minjin Rheu, Yue Dai, Jingbo Meng, and Wei Peng. 2024. When a chatbot disappoints you: Expectancy violation in human-chatbot interaction in a social support context. *Communication Research* 51, 7 (2024), 782–814.
[85] Phillip Richter, Heiko Wersing, and Anna-Lisa Vollmer. 2024. Reducing Mental Model Mismatch with Intention-Based Feedback in Human-Robot Teaching. In *Proceedings of the 12th International Conference on Human-Agent Interaction*. 399–401.
[86] Giacomo Rizzolatti and Massimo Matelli. 2003. Two different streams form the dorsal visual system: anatomy and functions. *Experimental brain research* 153, 2 (2003), 146–157.
[87] Lee Ross. 1977. The intuitive psychologist and his shortcomings: Distortions in the attribution process. In *Advances in experimental social psychology*. Vol. 10. Elsevier, 173–220.
[88] Stuart Russell and Peter Norvig. 2009. *Artificial Intelligence: A Modern Approach* (3rd ed.). Prentice Hall Press, USA.
[89] Erol Şahin et al. 2007. To afford or not to afford: A new formalization of affordances toward affordance-based robot control. *Adaptive Behavior* 15, 4 (2007), 447–472.
[90] Nadine B Sarter and David D Woods. 1995. How in the world did we ever get into that mode? Mode error and awareness in supervisory control. *Human factors* 37, 1 (1995), 5–19.
[91] Brian Scassellati. 2002. Theory of mind for a humanoid robot. *Autonomous Robots* 12, 1 (2002), 13–24.
[92] Brian M Scassellati. 2001. *Foundations for a Theory of Mind for a Humanoid Robot*. Ph. D. Dissertation. Massachusetts Institute of Technology.
[93] Kristin E Schaefer. 2016. Measuring trust in human robot interactions: Development of the "trust perception scale-HRI". In *Robust intelligence and trust in autonomous systems*. Springer, 191–218.
[94] Leonhard Schilbach et al. 2013. Toward a second-person neuroscience1. *Behavioral and brain sciences* 36, 4 (2013), 393–414.
[95] Natalie Sebanz, Harold Bekkering, and Günther Knoblich. 2006. Joint action: bodies and minds moving together. *Trends in cognitive sciences* 10, 2 (2006), 70–76.
[96] Bruno Siciliano, Oussama Khatib, and Torsten Kröger. 2008. *Springer handbook of robotics*. Vol. 200. Springer.
[97] Jana Siebert and Johannes Ulrich Siebert. 2023. Effective mitigation of the belief perseverance bias after the retraction of misinformation: Awareness training and counter-speech. *Plos one* 18, 3 (2023), e0282202.
[98] Phani Teja Singamaneni et al. 2024. A survey on socially aware robot navigation: Taxonomy and future challenges. *The International Journal of Robotics Research* 43, 10 (2024), 1533–1572.
[99] Hang Su, Wen Qi, Jiahao Chen, Chenguang Yang, Juan Sandoval, and Med Amine Laribi. 2023. Recent advancements in multimodal human–robot interaction. *Frontiers in Neurorobotics* 17 (2023), 1084000.
[100] Richard S Sutton, Doina Precup, and Satinder Singh. 1999. Between MDPs and semi-MDPs. *Artificial intelligence* 112, 1-2 (1999), 181–211.
[101] Aaquib Tabrez, Matthew B Luebbers, and Bradley Hayes. 2020. A survey of mental modeling techniques in human–robot teaming. *Current Robotics Reports* 1, 4 (2020), 259–267.
[102] Kartik Talamadupula, Gordon Briggs, Tathagata Chakraborti, Matthias Scheutz, and Subbarao Kambhampati. 2014. Coordination in human-robot teams using mental modeling and plan recognition. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2957–2962.
[103] Federico Tavella et al. 2024. Towards a computational model for higher orders of Theory of Mind in social agents. *Frontiers in Robotics and AI* 11 (2024), 1468756.
[104] Stefanie Tellex et al. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 25. 1507–1514.
[105] Sam Thellman, Maartje De Graaf, and Tom Ziemke. 2022. Mental state attribution to robots: A systematic review of conceptions, methods, and findings. *ACM Transactions on Human-Robot Interaction (THRI)* 11, 4 (2022), 1–51.

[106] Leimin Tian and Sharon Oviatt. 2021. A taxonomy of social errors in human-robot interaction. *ACM Transactions on Human-Robot Interaction (THRI)* 10, 2 (2021), 1–32.
[107] Michael Tomasello and Malinda Carpenter. 2007. Shared intentionality. *Developmental science* 10, 1 (2007), 121–125.
[108] Konstantinos Tsiakas, Vangelis Karkaletsis, and Fillia Makedon. 2018. A taxonomy in robot-assisted training: current trends, needs and challenges. In *Proceedings of the 11th PErvasive Technologies Related to Assistive Environments Conference*. 208–213.
[109] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *science* 185, 4157 (1974), 1124–1131.
[110] Leslie G Ungerleider. 1982. Two cortical visual systems. *Analysis of visual behavior* 549 (1982), chapter–18.
[111] Lennart Wachowiak, Andrew Coles, Gerard Canal, and Oya Celiktutan. 2024. A taxonomy of explanation types and need indicators in human–agent collaborations. *International Journal of Social Robotics* 16, 7 (2024), 1681–1692.
[112] Michael L Walters, Kerstin Dautenhahn, René Te Boekhorst, Kheng Lee Koay, Dag Sverre Syrdal, and Chrystopher L Nehaniv. 2009. An empirical framework for human-robot proxemics. In *Procs of new frontiers in human-robot interaction: symposium at the AISB09 convention*. 144–149.
[113] Katie Winkle, Donald McMillan, Maria Arnelid, Katherine Harrison, Madeline Balaam, Ericka Johnson, and Iolanda Leite. 2023. Feminist human-robot interaction: Disentangling power, principles and practice for better, more ethical HRI. In *Proceedings of the 2023 ACM/IEEE international conference on human-robot interaction*. 72–82.
[114] Rolf A Zwaan and Gabriel A Radvansky. 1998. Situation models in language comprehension and memory. *Psychological bulletin* 123, 2 (1998), 162.

A Diagnostic Reference for the Sixteen MMM Types

Table 4 provides the full diagnostic reference for all sixteen mismatch types, including definitions, the test question that distinguishes each type from its neighbours, and representative examples. This appendix is intended for annotators applying the taxonomy to interaction data and for designers reviewing a system against the full set of mismatch categories. The structural relationships between types (sequential dependencies and orthogonal dimensions within subcategories) are noted in the first column.

SPA / Subcategory	Mismatch Type	Definition & Test Question	Examples
SENSE Object Channel SOWT→SOWE SOWT→SOHW SOWE⊥SOHW	SOWT (Sense-Object-What)	Robot fails to detect entity, misidentifies it, or fails to perceive a relevant property. Addresses *What*. *Test: Would a different sensor or configuration fix it? Yes → SOWT.*	Cup outside field of view; object unrecognizable in dark; cup classified as bowl; weight not sensed without tactile sensor
	SOWE (Sense-Object-Where)	Robot detects entity but determines its spatial position incorrectly. Addresses *Where*. Presupposes SOWT satisfied. *Test: Detects but locates incorrectly? Yes → SOWE.*	Misjudges distance; depth too coarse; reference frame mismatch
	SOHW (Sense-Object-How)	Robot detects entity but fails to perceive a state change over time. Addresses *How*. Presupposes SOWT satisfied; orthogonal to SOWE. *Test: Detected entity but missed state change? Yes → SOHW.*	Water boils unnoticed (identity intact, transition missed); empty cup filled while unobserved; door closed while not looking
SENSE Comm. Channel SCNA→SCWT →SCWE →SCHW	SCNA (Sense-Communication-Not-Available)	User communicates through a modality the robot cannot receive or process as communicative input. Cross-dimensional precondition. *Test: Input modality unavailable? Yes → SCNA.*	Pointing without gesture recognition; speaking to microphoneless robot; gaze reference without gaze tracking
	SCWT (Sense-Communication-What)	User employs a term the robot does not know. Addresses *What*. Presupposes SCNA satisfied. *Test: Does the robot know the word? No → SCWT.*	"Carafe" unknown; "spicy" unmappable; "ladle" prompts "what's that?"
	SCWE (Sense-Communication-Where)	Term known but robot cannot identify which co-present instance is meant. Addresses *Where*. Presupposes SCWT satisfied. *Test: Knows term, not which one? Yes → SCWE.*	"The cup" with three present; "that one over there" with multiple candidates
	SCHW (Sense-Communication-How)	Robot cannot correctly map communicative act to appropriate response, including recognizing that no action is required. Addresses *How*. Presupposes SCWT satisfied and, where applicable, SCWE satisfied. *Test: Understands what to do? No → SCHW.*	"Get the kitchen ready" not decomposed; "I'm cold" not recognized as request; "thanks" treated as command
PLAN Design Scope	PSTD (Plan-Scope-Task-Design)	Task category falls outside robot's intended purpose. Cross-dimensional precondition for Plan. *Test: Task category in design scope? No → PSTD.*	Kitchen robot asked to do laundry; logistics robot expected to chat
PLAN World Context PWCS⊥PWMM	PWCS (Plan-World-Current-State)	Current world state absent from world model; not caused by sensor failure. Integrates *What+Where+How* (world model is a unified structure). *Test: Missing info not from sensor failure? Yes → PWCS.*	Fridge contents unknown; glass fragility unknown; oven heat not inferred; robot unaware own dust bag is full
	PWMM (Plan-World-Memory)	Information was once known but has been lost. Integrates *What+Where+How* (forgotten content can belong to any dimension). Distinguished from PWCS by temporal status. *Test: Knew before but forgot? Yes → PWMM.*	"Like last time" after reset; user name forgotten; preference lost
PLAN Planning Algorithm PAMT⊥PAHZ	PAMT (Plan-Algorithm-Method)	Robot understands task but chooses different method, strategy, or parameters. Addresses *How*. *Test: Right task, different strategy? Yes → PAMT.*	Wipes randomly not in rows; water 60°C vs. 80°C; slices potato before boiling
	PAHZ (Plan-Algorithm-Horizon)	Planning horizon or abstraction level diverges from user's assumption. Addresses *How*. *Test: More look-ahead would fix it? Yes → PAHZ.*	Expects full table-setting plan; "make a salad" → "what's first?"
ACT Physical APWE→APHW	APWE (Act-Physical-Where)	Incorrect assumption about robot's kinematic limits; cannot reach required position. Addresses *Where*. *Test: Could perform if it could reach? Yes → APWE.*	Shelf too high; too wide for doorway; arm cannot reach under table
	APHW (Act-Physical-How)	Robot reaches object but handling fails. Addresses *How*. Presupposes APWE satisfied. *Test: Reaches but fails at handling? Yes → APHW.*	Wet glass slips; bottle cap twist too weak; fragile object crushed
ACT Comm. Output ACNA→ACSC	ACNA (Act-Communication-Not-Available)	User expects a communication modality the robot cannot produce. Cross-dimensional precondition. *Test: Expected modality doesn't exist? Yes → ACNA.*	Expects verbal reply from mute robot; expects facial expressions without face
	ACSC (Act-Communication-Signal-Content)	User misinterprets what robot signals via existing modality. Integrates *What+Where+How* (design decision: signals carry meaning as indivisible units). Presupposes ACNA satisfied. *Test: Modality exists but meaning misread? Yes → ACSC.*	Nod read as agreement (meant attention); LED color misread; pause read as stuck (meant recalculating)

Table 4. Diagnostic reference for the sixteen MMM types with definitions, test questions, and examples. Types within subcategories follow proposed sequential dependencies (→) or address orthogonal dimensions (⊥).