Theory of Mind in Clarifying Dialogue: Inferring Another Person's Interpretation from Relational History

Wanhong HUANG
huangwanhong@serendip.ngo

Abstract

Clarifying dialogue can supply evidence about another person's interpretation while changing the interpretation being investigated. This paper develops a relational-history account of the resulting inference problem. It separates current coordination, retrospective attribution, and the prospective formation of understanding, while treating relations, histories, interpretive conditions, and response profiles as jointly constrained. A finite model retains paired initial and current candidate states through inquiry-induced transitions. Exact recovery of an earlier target requires constancy across the initial histories compatible with the complete dialogue transcript. A bounded adaptive-inquiry recursion characterizes when an allowed sequence of questions can achieve that recovery, and a constructive example shows how successful explanation can erase the discriminating opportunity while securing present mastery. Further examples distinguish historical ambiguity, response ambiguity, contextual error, and participants' interpretations of an investigator's expectations. The philosophical analysis connects first-person authority, interpersonal understanding, and the formation of shared meaning without identifying predictive success with access to another person's experience. A proposed study separates immediate correct responding, transfer, and retrospective attribution. The results are conditional formal consequences and conceptual distinctions; their psychological adequacy and the comparative merits of the proposed inquiry procedure remain empirical questions.

Keywords: theory of mind; clarifying dialogue; relational history; inquiry; historical identification

Research Note

This paper develops a bounded argument through conceptual distinctions, stated assumptions, conditional derivations, worked examples, and objections. Its length follows the requirements of that argument. Formal statements concern the specified models; the application of those models to persons and institutions requires independent interpretation and evidence. Constructed examples are explicitly identified and carry illustrative, rather than observational, status.

The manuscript follows the repository's paper template for typography, numbered formal statements, section structure, citations, and navigable contents. The associated project files preserve research provenance, source verification, derivation records, unresolved objections, and validation results. Philosophical commitments about the ultimate nature of social facts remain suspended where the descriptive framework is sufficient for the argument.

Authorship and Research Assistance

Wanhong Huang directed the research questions and authorized this paper's preparation. An AI research assistant helped organize the material, develop explicit formal constructions, check primary sources, and prepare the text. Agent-added arguments and formulations remain subject to the author's review. Source positions, repository proposals, and model consequences are distinguished throughout the manuscript.

Use and Licence

Copyright © 2026 Wanhong Huang. This work is licensed under Creative Commons Attribution–NonCommercial 4.0 International, following the repository template. Third-party material remains subject to its own applicable rights. The author encourages careful criticism and responsible application; this ethical request adds no condition to the licence. Reuse implies no authorial endorsement.

DRAFT

Contents

1 Introduction **4**

2 Antecedents and Scope of the Contribution **5**

3 Targets and Relational-History Descriptions **6**

4 Joint Compatibility and Inferential Warrant **7**
4.1 Joint admissibility and spurious alternatives 8

5 Historical Summaries and Target Recovery **8**

6 Inquiry-Induced Transitions and Earlier Interpretations **9**
6.1 A finite probabilistic extension 11
6.2 Visual comparison of historical and current targets 12

7 Bounded Adaptive Clarification **12**
7.1 An explicit adaptive inquiry tree 13

8 Constructed Dialogues and Competing Histories **14**
8.1 Instruction and the timing of acquisition 14
8.2 Relational caution and response selection 14
8.3 Repair and retrospective disagreement 14

9 Nested Interpretation and Observer-Authored Records **15**

10 Evidence Design and Evaluation **16**
10.1 Targets and observable measures 16
10.2 Comparison and error analysis 16
10.3 Falsification and revision 17

11 Philosophical Objections and Interpretive Limits **17**
11.1 First-person authority and historical fallibility 17
11.2 Interactional formation and representational abstraction 18
11.3 Practical understanding and the demand for certainty 18
11.4 Predictive success and interpersonal recognition 18

12 Discussion and Research Directions **19**

DRAFT

1 Introduction

Clarifying dialogue is both an opportunity for interpersonal understanding and an event in the history that understanding concerns. This paper investigates the resulting problem for theory of mind: the warranted attribution of an interpretation to another person when the act of asking can change that person's interpretation. The analysis proceeds from a distinction among historical attribution, current coordination, and prospective learning to a finite model of adaptive inquiry. It then examines conversational examples, experimental assessment, and philosophical objections.

Consider an instructor who asks a learner to explain a diagram. The learner hesitates. The instructor supplies a contrast, points to a boundary, and offers an example. The learner subsequently explains the diagram fluently. This sequence can support a practical judgment that the participants may proceed. Its bearing on the learner's understanding before the explanation remains a different question. The later response could express an interpretation already available but initially difficult to articulate. It could also express an interpretation formed during the exchange. A record containing only the successful final answer leaves these histories compatible.

The same structure appears in ordinary conversation. A friend explains what a remark meant after another person objects to it. A manager reformulates an instruction while investigating an apparent misunderstanding. An interviewer clarifies the categories within which a respondent is being asked to describe an experience. In each case, the exchange can change the categories, expectations, or practical stakes through which a response becomes possible. The difficulty concerns the time and object of attribution. It persists even when all participants are sincere and the final response is accurate.

The paper's central position is that a theory-of-mind account for these settings should retain three features together. First, candidate interpretations belong to jointly constrained relational histories: an interpretation, the circumstances in which it was formed, and the manner of its expression may restrict one another. Second, an attribution should name its temporal target. Third, an inquiry action can be part of the transition from an earlier configuration to a later one. These features allow a model to distinguish learning about a person from helping to constitute a new shared understanding, including exchanges that accomplish both.

The contribution is deliberately scoped. Existing work already investigates planning histories, communicative grounding, interactional transformation, and social relationships. Section 2 identifies those antecedents and the limits of the comparison. The formal contribution begins with a coupled representation that preserves an initial candidate while updating its current state. A target is historically identified when all surviving initial candidates agree about it. A bounded recursion then determines whether a permitted adaptive inquiry can secure that agreement under specified transition and observation rules. A small countermodel shows why a successful explanatory action can establish present mastery while leaving prior mastery ambiguous.

This formal distinction has philosophical consequences without settling the metaphysics of mind. Another person's interpretation can be accessible through interaction while the warrant for a particular historical attribution remains limited. The categories used to describe the person may themselves develop through the interaction. At the same time, a jointly achieved vocabulary can preserve disagreement about an earlier event. Current agreement therefore needs a temporal description before it can enter an explanation of what occurred.

The methodology combines conceptual analysis, finite constructions, and adversarial examples. A mathematical result establishes a consequence of a declared model. A constructed dialogue illustrates the conceptual conditions under which that result matters. A proposed study states how the account could be confronted with observations. These roles remain separate throughout. Exact recovery is used to expose an inferential boundary; many practical conversations instead require a calibrated, revisable judgment sufficient for a particular task. Section 10 explains how partial success can be evaluated without converting uncertain historical attribution into certainty.

DRAFT

2 Antecedents and Scope of the Contribution

This section positions the proposed account against five relevant lines of work. The comparison identifies the claims supported by the inspected sources, then states the narrower issue developed here. It also establishes limits on novelty: relational and formative accounts of social cognition already exist.

Baker, Saxe, and Tenenbaum model action understanding through inverse planning, with approximately rational planning supplying a generative account of observed action. Their discussion includes temporally structured behavior, changing goals, and subgoals [1]. This antecedent matters because an inverse account can already use a trajectory rather than an isolated movement. The present question concerns a further complication: the observer's clarification can enter the process generating the subsequent trajectory. The model must then preserve which earlier interpretation a later response is being used to support.

Clark and Brennan examine grounding through the evidence participants establish for understanding sufficient for their current purposes [3]. This purpose-relative orientation makes a useful starting point for distinguishing conversational continuation from a retrospective conclusion. Participants may reasonably coordinate their next action on the basis of an exchange even when the exchange leaves an earlier misunderstanding unresolved. The current paper develops that distinction by giving the two targets different maps on the same candidate history space.

De Jaegher and Di Paolo's account of participatory sense-making explicitly addresses how interaction can transform individual sense-making and generate meaning through coordination, while considering the autonomy of interaction and participants [4]. The formative dimension of dialogue therefore has a clear antecedent. The additional issue pursued here is the inferential warrant for an earlier-target claim after such a transformation. The finite construction is a device for examining that issue; a finite state description leaves much of lived interaction outside its representation.

Orne's analysis of demand characteristics concerns participants' interpretation of the experimental situation and its cues [5]. It provides a reason to include the participant's understanding of the investigator's expectations when interpreting experimental responses. The nested descriptions introduced later are bounded candidates for that purpose. Their depth and empirical usefulness require assessment; a more elaborate description can merely redescribe an investigator's preferred interpretation.

Finally, Chen and colleagues' recent preprint explicitly studies inverse planning in social relationships, using relational constraints to interpret social interaction [2]. The inspected abstract establishes its relevance and preprint status. Detailed comparison with its complete models and reported experiments remains a literature obligation. The present work therefore makes a scoped contribution claim about the constructions it states, while priority and comprehensive coverage remain open.

These antecedents suggest a disciplined division of intellectual labor. Inference requires a model of the relation between observable conduct and a target interpretation. Grounding requires a purpose against which communicative sufficiency can be assessed. Interactional formation requires a transition description through which the participants' understanding can change. Experimental interpretation requires attention to the situation participants believe they occupy. These are compatible requirements; combining their names would leave the substantive inferential problem untouched.

The paper gives that problem a precise form. Suppose two initially different interpretations produce the same final answer after explanation. A model that retains only the final state cannot determine which initial interpretation occurred. A model that retains the explanation and the initial-to-current relation can represent the ambiguity directly. Further inquiry can resolve it only if the permitted observations still separate the relevant initial candidates. This is the burden carried by the new formal construction.

The mathematical device itself belongs to elementary set-based reasoning about functions and partitions. The proof is supplied in full, and the paper claims no novelty for the general factorization principle. The research contribution lies in specifying the target, the inquiry-induced transition, the admissible histories, and the evidential obligations that a dialogue application must meet. An application gains substance when those specifications rule out an otherwise tempting attribution or justify a better inquiry sequence.

DRAFT

Table 1: Temporal targets in a clarifying exchange

Target	Question represented	Suitable evidence and limitation
Earlier interpretation	Interpretation available before a specified inquiry	Earlier productions and complete inquiry history; later success can arise through formation.
Current interpretation	Interpretation presently available for a task	Current explanation and transfer; a successful response may still have several compatible mechanisms.
Shared coordination	Sufficient alignment for a named joint activity	Reciprocal repair and coordinated action; agreement can remain local.
Prospective formation	Change produced by a specified exchange	Comparative exposure and later assessment; target and timing require stable definitions.

3 Targets and Relational-History Descriptions

The conceptual architecture fixes the object of inference before introducing an update rule. It distinguishes interpretations from utterances, temporal targets from present purposes, and descriptive categories from commitments about the ultimate nature of mental or social facts. Table 1 summarizes the target distinctions used throughout.

Definition 3.1 (Interpretation target)**.** An interpretation target is a specified distinction among a person's possible ways of taking an expression, action, situation, or relation, indexed to a time and a practical or explanatory purpose. Its formal representation is a map from admissible configurations to a declared value domain.

This definition avoids treating an answer as a transparent container of a mental state. For example, taking "finish the report" to mean "complete the analysis" differs from taking it to mean "send the current version." The distinction can matter before any overt response. Yet an observer needs evidence to apply it. The same sentence may be produced under both interpretations, and a person may understand both readings while remaining uncertain which one another person intended.

Four descriptive components help organize a candidate configuration:

$$z = (R, H, I, P) \in Z \subseteq \mathcal{R} \times \mathcal{H} \times \mathcal{I} \times \mathcal{P}. \tag{1}$$

Here R represents relevant present relations, H a relevant history, I interpretive conditions or distinctions, and P a response profile under specified inquiries. A response profile can include several possible expressions; it need not be a deterministic psychological disposition. The finite deterministic model introduced below is a further restriction for the purpose of a conditional derivation.

The subset sign in (1) performs conceptual work. A history in which a participant has never encountered a convention can be incompatible with an interpretation requiring that convention, unless another learning route is represented. A response produced under fear of sanction can belong to a different candidate configuration from a response produced under confidence that correction is welcome. Taking arbitrary combinations of independently listed components can manufacture candidate persons who have no coherent history within the account.

Joint admissibility also prevents a simple error about relationships. A relationship is relevant when it affects which histories, interpretations, or response possibilities are plausible for the target. Merely adding a label such as "friend" or "supervisor" supplies little explanatory content. The model should state whether the relation concerns shared vocabulary, authority, expected consequences of disagreement, remembered obligations, or another specified mechanism. A relationship label may summarize these features when the summary is adequate for the target.

The six descriptive categories used elsewhere in the research program remain available here. Persons and institutions are entities in the description; fluency and authority can be attributed properties; trust and

DRAFT

instruction are relations; a present interpretation is represented within a state; an answer or explanation is an event; learning and negotiation are processes. One exchange can admit all six descriptions. Their overlap is analytically useful because an event's occurrence and a process's development answer different questions. The taxonomy suspends a stronger claim that these categories divide reality into mutually exclusive kinds.

A person's experience also remains part of the relational history under consideration. An investigator's record of that experience is a second object, with its own author, time, and basis. The statement "the participant accepted the explanation" might describe a spoken assent, an investigator's inference, or a participant's later retrospective report. These versions can coexist in one record only if their status is preserved. Otherwise, the observer's description can silently become the person's purported mental history.

The central methodological requirement is consequently local specification. An inquiry begins by naming the target distinction, its time, its intended use, and the candidate histories relevant to it. The subsequent mathematics tests whether those candidates can be separated under the declared inquiry model. It cannot supply the original warrant for choosing the candidates or the authority to impose an interpretation on the person described.

4 Joint Compatibility and Inferential Warrant

This section develops the static compatibility problem that precedes adaptive dialogue. It defines candidate filtering, distinguishes disagreement from inconsistency, and derives an elementary constraint on historical summaries. The purpose is to make the inferential burden visible before inquiry is allowed to transform the candidate state.

Let Z be a nonempty finite set of jointly admissible configurations. An evidence record e is interpreted through a compatibility predicate $C(e, z) \in \{0, 1\}$. The surviving candidates and their target values are

$$Z_e = \{z \in Z : C(e, z) = 1\}, \qquad Q_e = \{q(z) : z \in Z_e\}. \tag{2}$$

Here $q : Z \to V$ is the target map. Compatibility is a declared relation between evidence and a model; treating it as perfect access to another person's experience would exceed the definition. In an application, C can depend on the reliability of a recording, the interpretation of a response, and the scope of the candidate mechanisms.

Definition 4.1 (Compatibility-based identification)**.** For $Z_e \neq \varnothing$, the target is identified when Q_e contains exactly one value. It is underidentified when Q_e contains at least two values. An empty Z_e is a model–evidence conflict requiring investigation.

The nonemptiness condition is indispensable. If every candidate contradicts the evidence, the statement that all surviving candidates agree is vacuously true in elementary logic. Calling this identification would reward a failed model. The interpretation should instead examine missing candidates, coding errors, unreliable evidence, or incompatible assumptions. Underidentification and conflict therefore require different responses: the former calls for discrimination among surviving possibilities; the latter calls for repair of the candidate account or its evidential interface.

A model may represent beliefs, intentions, and interpretive possibilities jointly even when a particular target concerns only one component. Suppose two histories yield the same current trust relation and the same verbal answer. In one history the participant acquired a convention through prior collaboration; in the other the convention was introduced by the interviewer. The target "can now use the convention" may be constant across the candidates. The target "understood the convention before the interview" may vary. Adding every detail of a person's biography would be wasteful, but discarding the acquisition history would destroy the latter distinction.

Structural comparisons also need aligned meanings. If a candidate represents a person's attributed relations by an edge set $\widehat{E}$ and the comparison account by E^*, then

$$E_{\text{add}} = \widehat{E} \setminus E^*, \qquad E_{\text{omit}} = E^* \setminus \widehat{E} \tag{3}$$

DRAFT

separates added and omitted relations only after vertices, edge types, time indices, and the evidential status of E^* have been aligned. A shared label such as "trust" can hide different targets. Counting edges before alignment can penalize a richer vocabulary or mistake a disputed comparison account for established truth.

Equation (3) is therefore an audit device, rather than a general score for understanding another person. Added relations and omitted relations can have different practical consequences. Attributing consent where only compliance is observed differs from overlooking a previously explained preference. Weighting those errors requires a purpose and a normative account. A lower total count can conceal the one error that matters for the intended use.

Probabilistic assessment is compatible with this architecture but requires additional commitments. A prior over Z, an observation likelihood, and an updating convention can assign different weights to compatible candidates. Exact compatibility-based identification is stronger than a high posterior probability. Conversely, a prior concentrated on one candidate can produce a confident answer without providing new discrimination in the evidence. The finite results below use possibility sets to keep this distinction visible. They leave uncertainty weighting and calibration to later empirical development.

The substantive work lies in justifying the candidate domain and its updates. A large candidate set that contains arbitrary stories can make every target unresolvable. A narrow set chosen to favor one interpretation can manufacture certainty. An adequate account documents why candidates enter, why some are excluded, and which observations could overturn those decisions. This discipline gives the formal language a critical function: it displays where an attribution depends on an evidential restriction or a modeling choice.

4.1 Joint admissibility and spurious alternatives

A finite example clarifies why joint admissibility matters before any update occurs. Let $H = \{h_0, h_1\}$ encode two possible formation histories and $I = \{i_0, i_1\}$ two interpretation states. Suppose the independently defended model admits only $Z = \{(h_0, i_0), (h_1, i_1)\}$. Evidence establishing h_0 then identifies i_0 within that model.

The product of the marginal projections is $\{h_0, h_1\} \times \{i_0, i_1\}$. It includes the additional pair (h_0, i_1). If the analyst replaces Z by this product, the same evidence leaves the interpretation ambiguous. The lost constraint has created an apparent inferential obstacle. In another application, an unwarranted constraint can create apparent certainty. The warrant for the admitted pairs is therefore substantive.

The example leaves open how the history and interpretation are related. Their covariance in an admissible set is compatible with several causal explanations and with a definitional restriction. A causal claim requires a transition or intervention account. The subset notation in (1) represents compatibility; it supplies no causal direction by itself.

For dialogue, this distinction can prevent two opposite errors. An analyst can treat a participant's interpretation as independent of the circumstances through which it became available, thereby admitting incoherent combinations. An analyst can also make a relationship label determine the interpretation by assumption, thereby excluding plausible disagreement. The candidate audit should inspect both the omitted constraints and the imposed constraints.

Joint modeling consequently means more than retaining many variables. It requires stating which combinations remain possible and why. The finite inference procedures become informative only after this conceptual and evidential work has been done.

5 Historical Summaries and Target Recovery

This section proves the criterion used to evaluate a compressed history or a dialogue transcript. It then examines the criterion's scope: exact recovery of one declared target is weaker than a complete reconstruction and stronger than adequate performance under an unspecified practical standard.

DRAFT

Let $\sigma : Z \to S$ be a summary map and $q : Z \to V$ a target. The kernel of a map is the equivalence relation it induces:

$$\ker \sigma = \{(z, z') \in Z^2 : \sigma(z) = \sigma(z')\}. \tag{4}$$

Two configurations in the same summary class are indistinguishable to a decoder whose only input is that summary.

Proposition 5.1 (Exact recovery through a summary)**.** *There exists a decoder $d : \sigma(Z) \to V$ satisfying $q = d \circ \sigma$ if and only if* $\ker \sigma \subseteq \ker q$.

Proof. Suppose $q = d \circ \sigma$. If $\sigma(z) = \sigma(z')$, applying d to both sides gives $q(z) = q(z')$. This proves the inclusion. Conversely, assume the inclusion. For $s \in \sigma(Z)$ choose any z with $\sigma(z) = s$ and define $d(s) = q(z)$. If another representative z' is chosen, then $(z, z') \in \ker \sigma$, hence $q(z) = q(z')$. The definition is consequently independent of the representative and gives the required factorization. □

The proof identifies the precise failure mode. When two target-distinct histories share a summary, a decoder must give them one output. At least one history receives the wrong target value. More sophisticated computation on the same summary cannot remove that conflict. Additional evidence, a changed target, or a restricted candidate domain can remove it, but each changes a premise of the recovery problem.

A present-state summary can therefore be adequate for one inquiry and inadequate for another. In the diagram example, a summary indicating current mastery may suffice to choose the next exercise. The same summary can merge histories with different earlier mastery. A disagreement about the summary's adequacy often becomes clearer when the two participants state their target maps. One is asking about continuing instruction; the other is asking what the learner understood before receiving help.

Historical sufficiency also differs from the exhaustive preservation of history. Suppose the target concerns which of two instructions was understood. A record that preserves the instruction actually received and the participant's earlier paraphrase might separate all target-distinct candidates even while omitting irrelevant room details. The proposition requires preservation of the target distinction, rather than a complete transcript of existence. Its practical difficulty is that future targets are partly unknown.

For a declared family $\mathcal{Q}$ of targets, simultaneous exact recovery requires

$$\ker \sigma \subseteq \bigcap_{q \in \mathcal{Q}} \ker q. \tag{5}$$

This follows by applying Proposition 5.1 separately to each target. Enlarging $\mathcal{Q}$ can demand a finer summary. A repository or institution should consequently state which future questions motivated its record design and which remain outside the design.

A further limitation concerns the evidential status of the domain. If the actual history lies outside Z, exact recovery on Z offers no guarantee about that history. Similarly, a perfectly sufficient summary of an inaccurate source record can reproduce the source's error faithfully. Sufficiency concerns preservation of distinctions already represented; truth and coverage require separate warrants. This distinction becomes especially important when one observer's interpretation is repeatedly summarized by others.

The criterion is useful precisely because its strength and limits are clear. It rules out an exact attribution from an insufficient record under fixed assumptions. It leaves open whether a probability estimate, a broader answer such as "either prior understanding or learning during the exchange," or a task-specific decision is adequate. In later sections, stopping inquiry will therefore be distinguished from achieving exact historical identification.

6 Inquiry-Induced Transitions and Earlier Interpretations

The static model now becomes a model of dialogue. This section fixes the timing of elicitation and observation, derives a coupled update, and constructs a case in which present learning coexists with

DRAFT

historical ambiguity. The construction isolates a possibility; whether a given question has this effect is an empirical matter.

Assumption 6.1 (Finite elicitation model)**.** Let Z and the action set A be finite. Each inquiry a has a deterministic transition $T_a : Z \to Z$ and a post-transition observation $O_a : Z \to Y_a$. These maps are known within the model. The initial target is $q_0 : Z \to V$. Allowed actions may depend on the surviving candidates.

The order of the maps matters. The model represents an action that may change the configuration before the observed answer is produced. A setting in which an answer is recorded before any change requires a different decomposition or an expanded state that preserves that answer. Treating the order as implicit would conceal the very question under investigation.

At the beginning, let the admissible initial candidates be $Z_0 \subseteq Z$, $Z_0 \neq \varnothing$. Preserve initial and current coordinates together:

$$B_0 = \{(z_0, z_0) : z_0 \in Z_0\} \subseteq Z \times Z. \tag{6}$$

For any nonempty paired candidate set B, an action a and observed answer y yield

$$B_{a,y} = \{(z_0, T_a(z)) : (z_0, z) \in B,\ O_a(T_a(z)) = y\}. \tag{7}$$

The first coordinate records which initial candidate remains possible. The second records its possible present state following the inquiry.

Equation (7) follows in three steps. Apply the action to each candidate's current coordinate. Retain the initial coordinate because the retrospective target still concerns it. Filter the resulting pairs by the actual answer. A model that keeps only current states performs an additional projection after this update. That projection can merge distinct initial candidates even when their shared current state is known exactly.

Define the surviving initial-target set and current-target set by

$$Q_0(B) = \{q_0(z_0) : (z_0, z) \in B\}, \qquad Q_t(B) = \{q_t(z) : (z_0, z) \in B\}. \tag{8}$$

The time-indexed map q_t must be specified. Reusing the same numerical values for the two maps leaves their temporal references distinct.

Example 6.2 (Explanation and retrospective ambiguity)**.** Let $Z = \{n, k\}$, where n represents absence and k presence of mastery of a specified distinction. Let $q_0(n) = 0$, $q_0(k) = 1$. An explanatory inquiry a has $T_a(n) = T_a(k) = k$ and $O_a(k) = 1$. Starting from $B_0 = \{(n, n), (k, k)\}$, a correct answer gives

$$B_{a,1} = \{(n, k), (k, k)\}.$$

For a current-mastery map with $q_t(k) = 1$, the current-target set is $\{1\}$. The initial-target set remains $\{0, 1\}$.

The example's strength comes from its narrowness. A single transition rule constructs two histories with the same current state and different initial targets. That suffices to refute the general inference from successful post-explanation performance to prior mastery. The binary vocabulary and perfect teaching rule simplify the witness. They are unnecessary as general psychological claims and would require validation in any application.

Repeated explanation in this example preserves the ambiguity. Once both paired candidates have current state k, any further inquiry whose behavior depends only on that current state generates the same answer from both. Initial differences can still be recovered from an independent earlier record. That record constitutes additional input and should enter the paired model explicitly. A participant's later recollection can also be evidence, with an observation model allowing its reliability and possible reconstruction to be examined.

The model offers a way to represent a productive conversation without forcing it into a purely diagnostic role. The instructor can reasonably prefer the explanation because present understanding is the immediate educational objective. The resulting record should then support a statement such as "mastery demonstrated after clarification," with the earlier-target question remaining open. Precision about timing preserves the achievement of the interaction and the limit of the retrospective inference simultaneously.

DRAFT

6.1 A finite probabilistic extension

The deterministic construction can be extended to uncertain transitions without discarding the initial coordinate. This subsection supplies the finite calculation and a numerical witness. The extension is a proposed model, and its transition probabilities require independent estimation or justification.

Let $b(z_0, z)$ be a probability distribution on initial/current pairs. For each allowed inquiry a, let $R_a(z', y \mid z)$ be a joint transition-and-observation kernel satisfying

$$R_a(z', y \mid z) \geq 0, \qquad \sum_{z',y} R_a(z', y \mid z) = 1.$$

The current state is assumed sufficient for the kernel's dependence on the represented past. When earlier history affects the response beyond the current description, the state must be expanded or the kernel's arguments changed.

The joint probability of the initial state, next state, and answer follows by summing over the previous current state:

$$\widetilde{b}_{a,y}(z_0, z') = \sum_z b(z_0, z) R_a(z', y \mid z).$$

Let $L_a(y) = \sum_{z_0,z'} \widetilde{b}_{a,y}(z_0, z')$. For $L_a(y) > 0$, the normalized update is

$$b_{a,y}(z_0, z') = \frac{\sum_z b(z_0, z) R_a(z', y \mid z)}{\sum_{\bar{z}_0,\bar{z},\bar{z}'} b(\bar{z}_0, \bar{z}) R_a(\bar{z}', y \mid \bar{z})}. \tag{9}$$

The numerator enumerates mutually exclusive previous states; the denominator is the probability of the observed answer. Their construction supplies the justification for each term. A zero denominator requires a model–observation conflict response rather than an arbitrary normalized belief.

Earlier and current target distributions are obtained by different marginals:

$$\Pr(q_0 = v \mid a, y) = \sum_{q_0(z_0)=v} \sum_{z'} b_{a,y}(z_0, z'), \qquad \Pr(q_t = v \mid a, y) = \sum_{q_t(z')=v} \sum_{z_0} b_{a,y}(z_0, z').$$

Marginalizing out z_0 before later inference can lose the association needed to interpret the current response retrospectively. Keeping the joint distribution preserves that association without assuming that the two targets coincide.

Take prior probabilities 0.6 for initial state n and 0.4 for k in Example 6.2. Perfect explanation followed by the same correct answer has likelihood one under both initial states. Equation (9) then retains probabilities 0.6 and 0.4 for the initial coordinate, while assigning probability one to current state k. The result expresses certain current mastery together with unchanged uncertainty about prior mastery.

Now suppose a nondestructive probe produces answer 1 with probability 0.2 from n and 0.8 from k. After that answer, the probability of initial k is

$$\frac{0.4(0.8)}{0.6(0.2) + 0.4(0.8)} = \frac{8}{11}.$$

This is increased support, while both initial states remain possible. If two probes occur before explanation, are conditionally independent given the unchanged state, and both yield 1, the corresponding probability is $0.4(0.8)^2/[0.6(0.2)^2 + 0.4(0.8)^2] = 32/35$. Conditional independence is an additional assumption; repeated questions in real dialogue can change memory, confidence, and the interpreted task.

If the same probes occur only after perfect explanation has merged the current states, their likelihoods are identical across the retained initial histories. They then add no discrimination about the initial target within this model. Timing and transition assumptions, rather than response count alone, determine the update's historical meaning.

The extension also explains why exact recovery can be a demanding practical benchmark. When every answer has positive probability under every initial candidate, finite evidence can leave all candidates possible while changing their weights substantially. A probabilistic stopping rule should state an error tolerance, the consequences of error, and a calibration obligation. Those choices are further elements of inquiry design, distinct from the exact finite recursion.

DRAFT

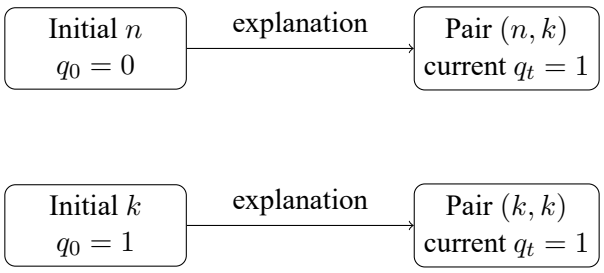

Figure 1: Constructed explanation histories with equal current mastery and distinct earlier targets

6.2 Visual comparison of historical and current targets

Figure 1 displays the two histories in Example 6.2. The arrows concern the current coordinate. The initial-target labels remain attached to their respective histories after convergence. This makes the distinction visible without requiring the reader to infer it from a marginal state alone.

The figure also locates a modeling choice. If an earlier record is later retrieved, it can distinguish the initial labels even after present states converge. If the retained evidence consists only of further responses generated from the shared current state, the model preserves the ambiguity. The distinction concerns available evidence about the histories, rather than a claim that the histories have become identical.

7 Bounded Adaptive Clarification

This section derives a criterion for successful inquiry when later questions may depend on earlier answers. It begins with the allowed actions, proves the finite recursion, and compares a separating probe with an explanatory action. The criterion concerns exact recovery under the declared model. Practical preferences among admissible policies require further objectives.

Let $A(B) \subseteq A$ be the actions available at a nonempty paired state B. This set may reflect actual conversational possibilities, time, participant agreement, and justified restrictions on inquiry. A policy is a rule selecting either an action in $A(B)$ or a terminal target report after each attainable transcript. Its depth is the largest number of inquiries on any attainable branch. A policy identifies the initial target if every terminal report equals $q_0(z_0)$ for every initial candidate producing that transcript.

Write $W_0(B)$ for the statement $|Q_0(B)| = 1$. Recursively define

$$W_{d+1}(B) = W_0(B) \ \vee \ \Big[\exists a \in A(B)\, \forall y \in Y_a : \big(B_{a,y} \neq \varnothing \implies W_d(B_{a,y})\big)\Big]. \tag{10}$$

Only attainable answers impose successor obligations. Empty successors represent branches excluded by the model, rather than successful evidence states. The first disjunct permits immediate stopping when the target is already constant.

Proposition 7.1 (Finite adaptive recovery)**.** *Under Assumption 6.1, $W_d(B)$ holds if and only if there is a deterministic allowed policy that identifies the initial target from B in at most d inquiries.*

Proof. For $d = 0$, the policy can only report immediately. Such a report is correct for every candidate exactly when $Q_0(B)$ is a singleton. Assume the equivalence at depth d. If $W_{d+1}(B)$ holds through its first disjunct, immediate reporting succeeds. Otherwise choose the action supplied by the second disjunct. For each attainable answer, the inductive hypothesis supplies a successful policy of depth at most d from $B_{a,y}$. Attach these policies to the first action, obtaining a policy of depth at most $d + 1$.

Conversely, suppose such a policy exists. If it stops immediately, its common correct report implies $W_0(B)$. If it acts, call its first action a. Each attainable answer must lead to a successful continuation of depth at most d, because the complete policy is correct on every attainable branch. The inductive hypothesis gives $W_d(B_{a,y})$ on each such branch. The second disjunct of (10) follows. □

DRAFT

The proof is constructive. When the recurrence is true, a witnessing action and its successor witnesses describe a successful inquiry tree. When it is false, every allowed first action has at least one attainable branch on which the remaining budget fails, unless the target was already constant. This is a model-relative obstruction, not evidence that the person is intrinsically unintelligible.

For a fixed policy π from a diagonal initial set, let $\rho_\pi : Z_0 \to \mathcal{T}$ map an initial candidate to its complete action-and-answer transcript, including the stopping point. Proposition 5.1 gives

$$\ker \rho_\pi \subseteq \ker q_0 \tag{11}$$

as the equivalent transcript criterion. Including actions is important: the answer "yes" following different explanatory prompts can have different historical implications. Including the stopping point prevents a truncated prefix from being confused with a completed inquiry.

Return to Example 6.2. Suppose an additional probe p leaves the state unchanged and yields $O_p(n) = 0$, $O_p(k) = 1$. The probe separates B_0 in one step, so $W_1(B_0)$ is true. Explanation alone produces a mixed initial-target successor and fails at every depth if all subsequent actions depend only on the merged current state. A probe followed by explanation can preserve the historical distinction and produce present mastery.

This comparison relies on a strong assumption: the probe is both separating and nondestructive relative to the target. Real questions can cue a distinction, alter confidence, or change the perceived task. The model therefore provides a research obligation to investigate such effects. It provides no reason to demand that ordinary interlocutors always be diagnosed before being helped. The order of actions should follow the declared purpose and the participants' interests.

The recurrence also treats worst-case success rather than expected cost. A policy that identifies every candidate may be lengthy, intrusive, or unavailable. A policy with lower expected effort can leave rare branches unresolved. Introducing costs, probabilities, or partial-credit targets requires explicit extensions. Randomization alone cannot create a deterministic distinction between two candidates with identical responses to every allowed action; additional observations or a different transition model are needed.

7.1 An explicit adaptive inquiry tree

The recursion can be illustrated by a case in which an adaptive sequence adds genuine discrimination. Let the initial states be $z_{00}, z_{01}, z_{10}, z_{11}$, where the two indices encode two specified interpretive distinctions. The target is the ordered pair of indices. All probes in this example leave the state unchanged.

An initial probe a reports the first index. After answer 0, an allowed probe b reports the second index for states in that branch. After answer 1, a different allowed probe c reports the second index for that branch. Suppose the inquiry setting allows a initially and makes b or c available only after the corresponding first answer. This restriction models a branch-specific clarification whose meaning is established by the first exchange.

Initially $W_0(B_0)$ is false because four target values survive. Also $W_1(B_0)$ is false: the only initial probe leaves two values on either branch. After answer 0, probe b has singleton-target successors, so $W_1(B_{a,0})$ is true. After answer 1, the same holds for c. Equation (10) therefore gives $W_2(B_0)$.

This calculation demonstrates both the budget and the action constraint. If b becomes unavailable, the first branch can remain unresolved even when the other branch succeeds. A procedure that reports only its successful branches would overstate worst-case identification. If the research purpose concerns only the first distinction, the target changes and one probe suffices. The shortest inquiry is consequently relative to both the target and the permitted action repertoire.

The example also suggests a concrete way to report an inquiry's limit. The observer can state which branch remains ambiguous and what additional distinction would need to be observed. Such a report is more informative than a global label of misunderstanding. Realizing the required probe in conversation remains an empirical and practical obligation.

DRAFT

8 Constructed Dialogues and Competing Histories

The following examples translate the formal distinctions into conversational episodes. Each is constructed for analysis. Their purpose is to test the account against ambiguity, changing relationships, and successful repair, rather than to provide evidence about their prevalence.

8.1 Instruction and the timing of acquisition

The first example examines a clarification that both teaches and tests. A coordinator tells a volunteer to "close the registration." The volunteer asks whether this means hiding the form or removing all existing entries. The coordinator explains the difference and asks for a description of the next step. The volunteer describes disabling new submissions while retaining the records.

One initial history contains familiarity with the software but uncertainty about the coordinator's intended policy. Another contains unfamiliarity with the software distinction itself. A third contains correct understanding and concern about receiving blame for acting without confirmation. The final description is compatible with all three. A binary mastery variable would merge policy uncertainty, technical understanding, and relational caution. The candidate domain should therefore represent the distinctions relevant to the actual question.

If the coordinator's objective is safe completion of the task, a joint walkthrough can supply sufficient current coordination. If the objective is to explain an earlier delay, the earlier instruction, the volunteer's prior experience, and the authority relation become relevant. Attributing the delay to technical ignorance from the final exchange would outrun the evidence. Attributing it to strategic hesitation would likewise require evidence for that particular history.

A useful record separates the observed question, the supplied explanation, the final paraphrase, and the inference about earlier understanding. Its precision lies in keeping these entries connected while preserving their different roles. A compact final label such as "instruction understood" should carry its time reference and the clarification that preceded it.

8.2 Relational caution and response selection

The second example isolates the relation between an available interpretation and its expression. Two colleagues have established that disagreement is welcome. A third participant has previously experienced correction being treated as disloyalty. When asked whether a proposal is clear, all three say yes. Their identical utterances can belong to different response profiles.

The analyst should avoid converting a background relationship into a complete explanation. A history of unwelcome correction makes caution a candidate mechanism; it leaves open whether that mechanism operated in this episode. Further evidence could include an earlier private question, a later independently offered correction, or differences across contexts. Each observation requires its own interpretation. Private speech is still speech in a relation, and later speech can incorporate newly acquired understanding.

The formal model accommodates these possibilities by expanding Z and altering T_a and O_a. An inquiry delivered by a different person can change the response profile and possibly the interpretation itself. Treating that change as simple noise would discard the mechanism under study. Conversely, attributing every response difference to power would exclude ordinary uncertainty, politeness, memory, and task differences without a test.

8.3 Repair and retrospective disagreement

The third example examines a successful repair that leaves a historical dispute open. A friend takes a remark to be dismissive. The speaker explains an intended reference, acknowledges why another reading was available, and uses a different expression. Both agree on how to communicate in the future. They may still disagree about what the original remark meant in its context.

DRAFT

Table 2: Constructed dialogue cases and the evidence each leaves open

Episode	Supported immediate use	Additional historical question
Registration instruction	Coordinate the next software action	Distinguish prior technical ignorance, policy uncertainty, and relational caution.
Proposal review	Record the observed assent	Distinguish understanding from expression under different anticipated consequences.
Friendship repair	Establish a future communicative convention	Retain separate claims about earlier intention, reception, and contextual meaning.

Here the prospective target is a workable shared convention. The retrospective targets include the speaker's earlier intention, the hearer's earlier interpretation, and the public meaning reasonably available in the situation. These targets concern different objects. A speaker's explanation is evidence about intention and a contribution to repair; it cannot alone establish the hearer's earlier experience or resolve every contextual interpretation.

Table 2 records the distinct inferential failures exposed by the three examples. The examples also show why maximal historical certainty can be an inappropriate practical objective. A friendship or collaborative task may proceed through a revisable agreement that explicitly leaves some history contested.

9 Nested Interpretation and Observer-Authored Records

This section examines two ways an investigator's own activity enters a theory-of-mind account. The first concerns participants interpreting the investigator. The second concerns later readers interpreting an investigator's record. Both require explicit targets and provenance, and both can generate misleading certainty if the levels are collapsed.

Let A be an investigator and B a participant. The notation $m_A(B; q)$ denotes A's model of B with respect to target q. A nested description

$$m_A\big(B;\, m_B(A; q)\big) \tag{12}$$

represents A's model of how B interprets A with respect to q. This is schematic notation for an object of description. An empirical model must specify its domain, observable implications, and closure rule.

For example, an interviewer may believe a participant is choosing an answer that the participant expects the interviewer to reward. This differs from the interviewer expecting the answer personally. It also differs from the participant understanding the stated task correctly. These possibilities can coincide, yet they remain separate hypotheses. The nested notation helps prevent a participant's supposed model of the investigator from being silently substituted for the investigator's actual intention.

Increasing nesting depth carries a cost. Additional levels can explain any unexpected response after the fact if each level is unconstrained. A bounded model should therefore state what closes the recursion. It might compare two explicit candidate expectations, use a finite experimentally disclosed instruction set, or treat deeper expectations as an unresolved residual. Closure by assumption is acceptable for a conditional result when the assumption remains visible.

A discriminating design changes a cue relevant to the participant's interpretation of the investigator while holding the substantive task constant as far as possible. The design still needs a manipulation check: the altered cue might change perceived difficulty, affiliation, or trust along with inferred expectations. An observed difference would establish a response contrast under the assigned conditions. Attributing that difference uniquely to a nested representation requires further evidence.

Observer-authored records add a second chain. Let r_A record A's statement about B, and let $s_C(r_A)$ be a later summary prepared by reader C. The record should retain at least the author, target person,

DRAFT

target time, observation time, asserted content, evidential basis, and status of the assertion. A sentence reporting that A inferred confusion should remain distinguishable from a direct quotation in which B reported confusion.

If several later records merely copy r_A, their number supplies no new independent discrimination among histories already merged by r_A. Formally, when s_C is a deterministic function of that record alone, equal original records necessarily produce equal later summaries. This limited observation leaves room for better interpretation of a rich source and for supplementary evidence. It concerns the preservation of distinctions, rather than the competence of every later reader.

The practical problem appears when a provisional attribution loses its qualification through repeated circulation. "The interviewer considered the participant uncertain" can become "the participant was uncertain," and then a purported explanation of later conduct. Preserving the inference's author and target time makes reassessment possible. A later participant correction should be linked to the earlier record while retaining the fact that the original attribution occurred.

These distinctions support a modest account of relational epistemic power. Control over the question, the response categories, and the official record can affect which interpretations remain available for later use. The present model represents particular mechanisms through action availability and record maps. Whether an institution uses that control asymmetrically, and with which consequences, requires a separate empirical inquiry. The formalism should expose such possibilities without turning them into an unsupported universal explanation.

10 Evidence Design and Evaluation

This section proposes a study program for the distinctions developed above. It specifies observable outcomes, separates formative effects from retrospective accuracy, and identifies checks that can falsify a proposed dialogue model. The program remains prospective; the paper reports constructed examples and formal results rather than collected participant data.

10.1 Targets and observable measures

The first design step is to separate three measurements. Immediate response accuracy concerns an answer produced after a specified prompt. Transfer concerns performance on a newly specified task after the exchange. Retrospective attribution concerns an observer's estimate of what the participant understood before that exchange. Combining these into one "understanding" score would conceal the distinction the study is intended to assess.

A suitable initial study uses a task with two explicit conventions and records the convention to which a participant was exposed before dialogue. Exposure supplies a controlled historical variable, while understanding of the exposure remains an empirical target. A pre-dialogue production can provide further evidence, although producing it may itself encourage reflection. The design should consequently distinguish assigned information, demonstrated earlier performance, and inferred earlier interpretation. Calling any one of these unqualified ground truth would overstate the design.

One experimental factor can vary whether clarification includes an explanatory contrast capable of introducing the relevant distinction. Another can vary whether observers receive the pre-clarification response and the full clarification sequence or only the final response. A later transfer task can test whether current success extends beyond repeating the supplied words. Table 3 separates the intended comparisons.

10.2 Comparison and error analysis

The evaluation should compare at least a final-response model, a static history model, and the coupled initial/current model. Their information inputs and fitting flexibility should be documented. A complex model that receives richer evidence can outperform a simple model for reasons unrelated to its transition

DRAFT

Table 3: Proposed study contrasts and their target interpretations

Contrast	Outcome	Main interpretation obligation
Explanatory versus minimally informative clarification	Immediate and later transfer performance	Specify all differences between prompts and examine whether the nominal probe also teaches.
Full transcript versus final response only	Observers' earlier-target judgments and uncertainty	Separate access to historical evidence from extra reading time and presentation differences.
Same task under different interpersonal cues	Response and interpretation reports	Check whether cues alter expectations, trust, task meaning, or several of these together.

representation. Comparisons using matched inputs help isolate the contribution of retaining the initial coordinate.

For a finite coded task, the evaluator can report the size of the predicted initial-target set, whether it contains the independently assessed target, and the rate at which a model declares conflict. A model that always retains every possibility has broad coverage but little discrimination. A model that always reports one value has apparent precision but can exclude the supported history. Reporting both properties prevents an arbitrary preference for either certainty or permanent ambiguity.

Human-observer judgments can be evaluated for calibration when a probabilistic answer is requested and a suitable reference measure exists. Such evaluation requires repeated comparable tasks and a clear account of reference error. For richer real dialogue, qualitative disagreement analysis may be more appropriate than a single accuracy score. Participant corrections should be preserved as evidence with a stated target and timing.

10.3 Falsification and revision

The deterministic transition model fails as an empirical description if nominally equivalent candidate states systematically yield different post-inquiry outcomes unexplained by its domain. The appropriate revision may expand the state, introduce stochastic transitions, or narrow the claimed application. Expanding the state after every failure can also make the model untestable. A study should predeclare which variations count as evidence against the proposed account.

A stronger challenge would show that the distinction between earlier and current targets adds little explanatory or predictive value under the study's actual purposes. That result would restrict the practical usefulness of the framework even while the finite counterexample remained mathematically valid. Conversely, systematic retrospective overconfidence after successful explanation would motivate a more detailed investigation of the mechanism. The framework supports either outcome by separating the model's logical possibility claim from an empirical prevalence claim.

11 Philosophical Objections and Interpretive Limits

The philosophical assessment examines whether the model misdescribes interpersonal understanding, weakens first-person authority, or makes clarification impractically demanding. Each objection targets a different transition from formal representation to human application.

11.1 First-person authority and historical fallibility

A participant ordinarily has a distinctive position in articulating what an expression means to them, describing an experience, and correcting an attributed intention. The framework can represent that position by giving first-person reports a specified evidential role. Its central distinction concerns how a report is

DRAFT

used across time. A present articulation can express a current interpretation authoritatively while leaving the path through which that interpretation emerged partly uncertain.

This distinction should be applied symmetrically. An observer's historical model also has limitations and can be less well supported than the person's account. A finite candidate list has no independent authority to disqualify a participant's correction. When a report falls outside the domain, an empty update identifies a conflict between the model and evidence. It creates a reason to examine the domain and its assumptions.

A risk remains when uncertainty is selectively invoked to discount one person's testimony while another person's interpretation is treated as transparent. The proposed record structure exposes the asymmetry by preserving author, target, basis, and status for each attribution. Deciding the evidential weight of competing reports requires a substantive inquiry; the set-based formalism leaves that judgment visible.

11.2 Interactional formation and representational abstraction

An objection from the formative character of dialogue holds that fixed candidate states miss the emergence of meanings participants could not previously describe. The objection is serious. The finite model assumes a declared domain and transition repertoire, so its strongest theorem applies only within that closure. A conversation that creates a new target distinction may require an enlarged domain or a revised target map.

The appropriate response is to preserve the revision explicitly. Earlier targets can remain indexed to their earlier vocabulary, with a separately justified translation to the new vocabulary when such a translation exists. Assuming a complete translation in advance would erase an important form of novelty. A partial correspondence can record where the earlier and later distinctions align and where they remain incomparable.

This limitation does not make all formal analysis empty. The binary teaching example establishes a clear failure of a proposed inference even in a favorable closed model. Richer emergence creates further obligations, rather than rescuing that inference automatically. Formal abstraction is useful when it reveals a dependency and remains answerable to the phenomenon it simplifies.

11.3 Practical understanding and the demand for certainty

Many conversations proceed without reconstructing another person's historical interpretation exactly. Coordination can rely on a clarification, an apology, a demonstration, or a provisional agreement. The framework accommodates this by distinguishing the purpose of the exchange from a researcher's retrospective target. A policy can stop because sufficient current coordination has been reached even when $W_d(B)$ remains false.

The converse risk is an endless demand for further explanation imposed on one participant. An analyst should specify whether another question is available, proportionate, and useful for the agreed purpose. Restricting $A(B)$ can make historical recovery impossible under the remaining options. That outcome describes a limit of the inquiry, rather than a deficiency in the person or a warrant for unlimited questioning.

11.4 Predictive success and interpersonal recognition

Accurate prediction can coexist with misunderstanding of the reasons a person regards as important. A model might predict assent from an authority relation while omitting the participant's articulated commitments. Another model might represent those commitments faithfully while remaining unable to predict an unusual response. These are distinct achievements that require distinct evaluation criteria.

The present paper treats interpretation as a declared target rather than assuming that behavioral prediction exhausts it. A richer account can include the person's recognized reasons, capacities for correction, and participation in specifying the question. Agreement with an observer remains different from recog-

DRAFT

nition of the person as an interlocutor. The formal apparatus can clarify evidential limits; the ethical quality of the relation requires its own argument.

12 Discussion and Research Directions

The discussion connects the formal results to the paper's original question and identifies the next developments that would change the argument's scope. It distinguishes completed conditional results from open psychological, methodological, and philosophical obligations.

The analysis supports a time-indexed account of clarifying dialogue. A response has an evidential history, and the inquiry that elicited it can be part of the process forming its meaning. Retaining an initial coordinate alongside the current state allows a model to represent learning and retrospective ambiguity together. The factorization criterion identifies when a transcript preserves the target distinction; the adaptive recursion identifies when a permitted finite inquiry can secure that preservation.

The teaching example establishes a limited but important possibility: perfect present performance can coexist with unresolved prior understanding. The result survives any interpretation of the example that preserves its two essential premises, a target-distinct pair of initial candidates and an inquiry that merges their observable current possibilities. Its application to a real exchange requires evidence that those premises describe the exchange adequately.

Several developments would make the framework more useful. The finite probabilistic extension already retains a joint distribution over initial and current states. Its transition and observation kernels now require empirical estimation, comparison, and sensitivity analysis. Approximate recovery should define an error criterion and distinguish calibration from target-set coverage. Open-ended dialogue should examine how the candidate vocabulary changes, including cases where a later interpretation has only a partial correspondence with earlier categories. Each extension needs a substantive reason; additional notation alone would add little.

The study program provides a route to assessing practical value. Its key test concerns whether preserving the clarification sequence and the temporal target improves retrospective judgments while allowing present learning to be recognized. A successful study would still leave broad generalization across educational, interpersonal, and institutional settings open. A null or adverse result would help identify purposes for which simpler records and models suffice.

The philosophical consequence is a disciplined plurality of achievements. An exchange may clarify the speaker's intended meaning, change the hearer's understanding, establish a workable convention, and leave an earlier disagreement unsettled. Those outcomes can be described together without reducing them to a single score. A serious theory-of-mind account should make clear which achievement its evidence supports and which historical claims remain available for correction.

DRAFT

References

[1] Chris L. Baker, Rebecca Saxe, and Joshua B. Tenenbaum. Action understanding as inverse planning. *Cognition*, 2009. Verified primary source.

[2] Alicia M. Chen, Ashley J. Thomas, Joshua B. Tenenbaum, and Rebecca Saxe. Inverse planning of social interactions in relationships, 2026. Preprint, arXiv:2609.01639. Primary source record.

[3] Herbert H. Clark and Susan E. Brennan. Grounding in communication. In *Perspectives on Socially Shared Cognition*, pages 127–149. American Psychological Association, 1991. Primary chapter reproduction.

[4] Hanne De Jaegher and Ezequiel Di Paolo. Participatory sense-making: An enactive approach to social cognition. *Phenomenology and the Cognitive Sciences*, 6:485–507, 2007. Verified primary source.

[5] Martin T. Orne. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, 17:776–783, 1962. Verified primary source.

DRAFT