Epistemic Architecture and Organizational Performance

Saeed Ullah* **and Naveed Ahmed**
PakCrypt NPO, Pakistan
*su@pakcrypt.org

Abstract: This paper presents a novel diagnostic framework—the *Epistemic Architecture Model (EAM)*—for classifying and predicting team performance in organizational settings based on the structure of mutual knowledge, information asymmetry, and belief accuracy among team members. Drawing on the formal game-theoretic concepts of common knowledge (Aumann, 1976; Lewis, 1969), Bayesian games under incomplete information (Harsanyi, 1967–1968), and the Persuasion Knowledge Model (Friestad & Wright, 1994), we integrate insights from Goffman's dramaturgical theory, Theory of Mind research, and affect labeling neuroscience into a unified taxonomy of five canonical epistemic settings that characterize organizational teams. Each setting—Opacity, Asymmetric Insight, Shared-but-Not-Common Knowledge, Full Common Knowledge, and False Belief—generates distinct predictions regarding coordination efficiency, strategic behaviour, psychological safety, and collective performance. We present a comparative analysis grounded in experimental game theory, derive testable propositions for each setting, and outline a diagnostic methodology for practitioners in industrial-organizational psychology. The framework addresses a gap in the I/O literature by formalizing the intuition that *what people know about what others know* is as consequential for team outcomes as the substantive content of their knowledge. Implications for organizational diagnosis, team design, and leadership intervention are discussed.

Keywords: common knowledge; information asymmetry; game theory; team performance; organizational diagnosis; epistemic states; impression management; Theory of Mind; industrial-organizational psychology; Bayesian games

1. Introduction

A persistent challenge in industrial-organizational psychology is the prediction of team performance from structural and relational variables that precede observable outputs. While decades of research have identified factors such as cohesion (Beal, Cohen, Burke, & McLendon, 2003), psychological safety (Edmondson, 1999), and shared mental models (Cannon-Bowers, Salas, & Converse, 1993) as reliable predictors, the field has yet to develop a comprehensive formal framework for characterizing the *epistemic architecture* of a team—that is, the precise structure of who knows what, who knows that others know, and how accurate those higher-order beliefs are. This omission is consequential. As the game-theoretic literature has demonstrated with mathematical rigour, the difference between merely shared knowledge and true common knowledge is not quantitative but qualitative, producing dramatically different equilibria even when the underlying facts remain identical (Rubinstein, 1989).

Consider a team in which every member privately recognizes that their current project strategy is failing, but no one believes that others share this recognition. In Harsanyi's (1967–1968) terms, each member possesses a private type that includes the belief "the strategy is failing," yet the absence of common knowledge prevents coordinated action to change course. The team behaves as though the strategy were sound—not because any individual believes it, but because the epistemic structure forecloses collective acknowledgment. This phenomenon, which pluralistic ignorance researchers have documented extensively (Prentice & Miller, 1993), is fundamentally a problem of *epistemic architecture* rather than individual cognition or motivation.

The present paper addresses this gap by developing the Epistemic Architecture Model (EAM), a taxonomic framework that identifies five canonical epistemic settings in organizational teams and derives formal predictions about team performance from each. The framework integrates three theoretical traditions that have developed largely in isolation: (a) the game-theoretic analysis of information structure, common knowledge, and strategic behaviour (Aumann, 1976; Harsanyi, 1967–1968; Lewis, 1969; Rubinstein, 1989); (b) the social-psychological analysis of impression management, self-presentation, and persuasion detection (Friestad & Wright, 1994; Goffman, 1959; Jones & Pittman, 1982); and (c) the cognitive science of Theory of Mind, false belief processing, and metacognitive bias (Baron-Cohen, Leslie, & Frith, 1985; Gilovich, Savitsky, & Medvec, 1998; Kinderman, Dunbar, & Bentall, 1998).

The core contribution is a comparative analysis demonstrating that the same team, facing the same task, can be expected to produce qualitatively different outcomes depending solely on which epistemic setting obtains. We derive these predictions not from ad hoc intuition but from the formal structure of Bayesian games under varying information conditions, supplemented by

experimental evidence from behavioural game theory, social psychology, and organisational research. The practical implication is a diagnostic methodology that I/O psychologists can deploy to assess a team's epistemic architecture and predict performance trajectories before they manifest in observable outputs.

2. Theoretical Foundations

2.1 Information Asymmetry and Strategic Interaction

The recognition that private information confers strategic advantage is among the most consequential insights of twentieth-century economic theory. Akerlof's (1970) analysis of adverse selection demonstrated that when sellers possess private information about product quality that buyers lack, market equilibria degrade: prices fall to reflect average rather than actual quality, high-quality sellers exit, and the market may collapse entirely. Spence's (1973) signalling theory showed how informed agents can credibly communicate private information through costly signals—actions that are differentially expensive depending on the agent's true type, making imitation by inferior types unprofitable.

These insights translate directly to organisational teams. A team member who possesses private knowledge about a colleague's competence, the viability of a strategy, or the political dynamics of a resource-allocation process holds an informational advantage analogous to Akerlof's informed seller. The holder of private information can choose to reveal, conceal, or strategically distort it. Critically, the *credibility* of any revelation depends on whether the audience can distinguish genuine signals from cheap talk—a distinction that Spence's framework formalizes and that Goffman's (1959) dramaturgical theory describes in sociological terms as the difference between expressions "given" (deliberately communicated) and expressions "given off" (inadvertently revealed).

Harsanyi's (1967–1968) framework for games of incomplete information provides the mathematical apparatus for modelling these situations. His central innovation was the introduction of "types"—private information parameters assigned by a fictional prior move of Nature from a commonly known distribution. Each player forms probabilistic beliefs about others' types and selects strategies that maximize expected utility conditional on those beliefs. The resulting Bayesian Nash Equilibrium concept enables precise predictions about strategic behaviour under any specified information structure. For the I/O psychologist, this framework provides the formal underpinning for analysing how asymmetric information within a team shapes cooperation, conflict, and coordination.

2.2 Common Knowledge and Its Epistemic Hierarchy

David Lewis (1969), in his landmark philosophical treatise *Convention*, identified a recursive epistemic structure that he argued was necessary for social conventions to emerge and sustain themselves. A proposition *p* is common knowledge among a group if and only if: (1) everyone knows *p*; (2) everyone knows that everyone knows *p*; (3) everyone knows that everyone knows that everyone knows *p*; and so on, *ad infinitum*. Robert Aumann (1976) formalized this concept using information partitions and proved a remarkable result: rational agents with a common prior cannot "agree to disagree" if their disagreement is itself common knowledge.

The qualitative discontinuity between finite mutual knowledge and true common knowledge was demonstrated most strikingly by Rubinstein (1989) in his Electronic Mail Game. In this coordination problem, two players must decide whether to invest, with both investing yielding a payoff but unilateral investment producing a severe loss. The state of the world determines whether investment is desirable, but only Player 1 observes the state. An unreliable communication channel transmits confirmations back and forth, each with a small probability of failure. Rubinstein proved that no matter how many successful round-trips occur—even millions—the unique equilibrium is non-investment. Any finite chain of mutual knowledge, however long, fails to support the coordination that true common knowledge would enable. This result has profound implications for organisational settings: it demonstrates that "everyone roughly knows that everyone knows" is not an approximation of common knowledge but a qualitatively different epistemic state with qualitatively different strategic consequences.

The closely related Coordinated Attack Problem (Two Generals Problem) from distributed computing theory reinforces this point. Two generals must attack simultaneously to succeed; messages through enemy territory may be intercepted. No finite exchange of confirmations can establish common knowledge of the attack plan, because the last sender always faces uncertainty about whether their message arrived. The organisational analogue is the team attempting to coordinate a strategic pivot via sequential communications that are never simultaneously witnessed by all members—a situation far more common than the rare circumstances that produce genuine common knowledge (e.g., simultaneous public announcements in a shared physical space).

2.3 Impression Management and Persuasion Knowledge

Goffman's (1959) dramaturgical theory conceptualizes social interaction as a theatrical performance in which actors manage impressions through a carefully maintained separation between the "front stage" (the performance presented to audiences) and the "back stage" (preparation, true motives, unguarded behaviour). Audience segregation—the practice of

presenting different performances to different audiences—is a primary mechanism of impression management. In organisational settings, this maps to the phenomenon of employees maintaining distinct personas for superiors, peers, and subordinates.

Jones and Pittman (1982) identified five canonical self-presentation strategies: ingratiation (seeking to be liked), self-promotion (seeking to appear competent), intimidation (seeking to appear powerful), exemplification (seeking to appear morally virtuous), and supplication (seeking to appear helpless to elicit assistance). Each carries a specific detection risk: detected ingratiation appears sycophantic; detected self-promotion appears conceited; detected intimidation—directly relevant to the firm-handshake scenario with which we opened our analysis—appears impotent or absurd.

The Persuasion Knowledge Model (Friestad & Wright, 1994) provides the mechanism through which detection occurs and takes effect. The model proposes that targets of influence attempts maintain evolving knowledge structures about persuasion tactics, and that when this knowledge is activated, a "change of meaning" occurs: the persuasive action is reclassified from content-to-be-evaluated to manipulation-to-be-resisted. This change of meaning is not merely attitudinal; it alters the entire frame through which the action is interpreted. The activation of persuasion knowledge is precisely the mechanism by which calling out a hidden strategy neutralizes it—and by extension, the mechanism through which transparency about strategic intent within a team can transform the interpretive frameworks through which team members evaluate one another's actions.

2.4 Theory of Mind and Metacognitive Biases

The capacity to attribute mental states to others—beliefs, desires, intentions, knowledge—is known as Theory of Mind (ToM). The developmental trajectory of ToM begins with first-order false belief understanding around age four to five (Baron-Cohen, Leslie, & Frith, 1985) and progresses to second-order understanding ("Sally thinks Anne thinks...") by age six to seven. Adults can typically manage approximately five levels of recursive intentionality (Kinderman, Dunbar, & Bentall, 1998), with performance degrading beyond this threshold on mentalizing tasks but not on equivalent factual-memory controls—confirming that recursive social cognition, not working memory, is the bottleneck.

This cognitive ceiling has direct implications for the common knowledge hierarchy. Humans cannot achieve common knowledge through explicit recursive reasoning because they cannot mentally process infinite chains of "I know that you know that I know." Instead, they approximate common knowledge through a qualitative heuristic—a sense that something is

"public and obvious"—which substitutes for the impossible infinite regress. This heuristic, while functional, introduces systematic errors. Three well-documented biases distort higher-order beliefs in organisational settings.

The *curse of knowledge* (Camerer, Loewenstein, & Weber, 1989) causes informed agents to overestimate how much others know. Elizabeth Newton's celebrated experiment demonstrated this vividly: tappers predicted that listeners would identify 50% of tapped melodies, when actual recognition was only 3%. In teams, the curse means that an expert member may assume shared understanding of technical constraints and fail to communicate them explicitly—a failure not of communication skill but of epistemic calibration.

The *illusion of transparency* (Gilovich, Savitsky, & Medvec, 1998) is the complementary bias: overestimating how much one's internal states are visible to others. Liars believed observers could detect their lies far more easily than was actually the case. Team members experiencing frustration, disengagement, or ethical concern may overestimate the visibility of these states and consequently fail to articulate them, operating under the false assumption that "everyone can tell."

The *spotlight effect* (Gilovich, Medvec, & Savitsky, 2000) extends this to appearance and behaviour: individuals systematically overestimate how much attention others pay to them. Combined with the illusion of transparency, these biases produce a systematic distortion of perceived epistemic architecture: team members believe the group's epistemic state is more transparent than it actually is, creating a false sense of shared understanding where none exists.

2.5 The Therapeutic Power of Naming: Affect Labeling and Exposure

A final theoretical strand that bears directly on organisational diagnosis concerns the effects of making implicit psychological content explicit. Lieberman and colleagues (2007) demonstrated via functional neuroimaging that labelling a negative emotion ("affect labeling") diminishes amygdala activation while increasing right ventrolateral prefrontal cortex (RVLPFC) activity, establishing a neural inhibitory pathway from symbolic representation to threat response. This finding has been replicated across diverse populations and paradigms, including arachnophobia treatment and test anxiety.

Viktor Frankl's paradoxical intention technique directly addresses recursive anxiety: by prescribing deliberate intention of the feared outcome, the anticipatory anxiety cycle is broken. Wegner's (1994) ironic process theory explains the complementary phenomenon: active suppression of unwanted thoughts paradoxically increases their frequency and intensity. Together, these findings establish a robust empirical principle: *approaching and naming what is avoided*

creates new cognitive associations that inhibit the original threat response. For organizational teams, this principle suggests that surfacing hidden conflicts, unspoken concerns, and covert strategies may not merely improve communication but fundamentally alter the emotional and strategic dynamics of the group.

3. The Epistemic Architecture Model: A Taxonomy of Organizational Epistemic Settings

We now present the central contribution of this paper: a formal taxonomy of five canonical epistemic settings that characterize the information structure of organizational teams. Each setting is defined by the distribution of first-order knowledge (what each member knows about the task, each other, and the environment), higher-order knowledge (what each member knows about what others know), and belief accuracy (whether higher-order beliefs are correct). We derive predictions for each setting from game-theoretic equilibrium analysis supplemented by social-psychological evidence.

3.1 Formal Definitions

Let $N = \{1, 2, ..., n\}$ be a finite set of team members. Let Ω denote the set of possible states of the world relevant to the team's task. Each member i possesses a private information set $P_i(\omega)$ containing the states that i considers possible when the true state is ω. Following Aumann (1976), we define:

Mutual knowledge of order 1: An event E is mutual knowledge of order 1 if every member knows E. Formally, $K_1(E) = \bigcap_{i \in N} K_i(E)$. This is the simplest level of shared information. It means that every single person in the group knows event E. At this level, everyone knows the truth, but they are in the dark about what others know.

Mutual knowledge of order k: $K_k(E) = K_1(K_{k-1}(E))$. This creates a "layering" effect of knowledge. Order 2 means everyone knows that everyone else knows E. Similarly, Order 3 means everyone knows that everyone else knows that everyone else knows E.

Common knowledge: $CK(E) = \bigcap_{k=1\infty} K_k(E)$. An event is common knowledge if and only if every finite order of mutual knowledge obtains. Common knowledge usually requires a public announcement. If I whisper a secret to everyone in a room individually, it is Mutual Knowledge. Everyone knows it, but they are all pretending they don't know that the others know. If one stands on a table and shout the secret, it becomes Common Knowledge.

Using these definitions, we now characterize the five canonical settings.

3.2 Setting I: Opacity ("No One Knows What Others Know")

Definition. Each member *i* possesses private information about the task and the team, but has no model of other members' information states. Formally, for all *i*, member *i*'s beliefs about other members' information sets are uninformative priors.

Organisational manifestation. The Opacity setting characterizes teams with high functional specialization and low cross-functional interaction—e.g., newly assembled project teams with members from different departments, distributed teams across time zones with no shared communication channel, or merger-integration teams composed of members from previously separate organisations. Each member holds expertise and contextual knowledge that others neither possess nor understand.

Game-theoretic prediction. In the absence of any model of others' information, strategic interaction reduces to a game against Nature. Members cannot condition their actions on others' likely actions because they have no basis for predicting them. Coordination depends entirely on Schelling focal points (Schelling, 1960)—culturally or contextually salient defaults that serve as implicit coordination devices. Spence-type signalling is impossible because the audience cannot distinguish signals from noise without a model of the signaller's type space.

Performance prediction: low coordination efficiency, high variance in outcomes, frequent misalignment, but potentially high innovation due to the absence of conformity pressures.

3.3 Setting II: Asymmetric Insight ("One Party Knows What Others Know")

Definition. One member or subgroup (the *informed party*) possesses accurate higher-order beliefs about other members' knowledge, while the remaining members lack such higher-order knowledge. This is the classic Harsanyi (1967–1968) incomplete-information game structure from the informed party's perspective.

Organisational manifestation. The Asymmetric Insight setting describes teams with experienced leaders or managers who understand their subordinates' knowledge states while subordinates do not accurately model the leader's or each other's knowledge. It also characterizes situations in which one member has conducted a thorough stakeholder analysis and consequently understands the interests, constraints, and knowledge of others, while those others operate in relative ignorance of one another.

Game-theoretic prediction. The informed party holds a decisive strategic advantage. They can engage in Spence-type signalling with full awareness of how signals will be received. They can employ Goffman's impression management strategies with reduced detection risk, because they accurately model the audience's interpretive frameworks. However, the Persuasion Knowledge Model predicts that if the uninformed parties become aware that the informed party possesses this advantage—i.e., if the asymmetry itself becomes common knowledge—resistance and distrust may emerge.

Performance prediction: high efficiency when the informed party is competent and benevolent (effective leadership); exploitation and morale collapse when the informed party pursues self-serving objectives.

3.4 Setting III: Shared-but-Not-Common Knowledge ("Everyone Knows, but No One Knows That Everyone Knows")

Definition. Every member knows the relevant proposition p (e.g., that the project is behind schedule), and most members believe that others probably also know p. However, this mutual awareness has never been publicly established. Formally, first-order mutual knowledge $K_1(p)$ obtains, and perhaps even $K_2(p)$, but $CK(p)$ does not.

Organisational manifestation. This is perhaps the most pervasive and consequential setting in organisational life. It describes the team that collectively recognizes a failing strategy, an underperforming member, an ethical violation, or an impending market shift—but has never acknowledged this recognition in a public forum. Each member privately knows; each suspects others know; but the absence of common knowledge prevents coordinated action. This is the epistemic structure underlying pluralistic ignorance (Prentice & Miller, 1993) and the "elephant in the room" phenomenon.

Game-theoretic prediction. Rubinstein's (1989) Electronic Mail Game provides the formal prediction: coordination failure. Despite any finite level of mutual knowledge, the team cannot achieve the coordination equilibrium that common knowledge would support. Each member faces a version of the Two Generals Problem: they know, and they suspect others know, but they cannot be certain that others are certain, producing a cascade of doubt that prevents unilateral action. The equilibrium is inertia—continued adherence to the status quo despite universal private dissatisfaction. Performance prediction: systematic underperformance relative to the team's potential, driven not by ignorance or incompetence but by epistemic structure. The team's performance ceiling is determined by the gap between K_k and CK.

3.5 Setting IV: Full Common Knowledge ("Everyone Knows That Everyone Knows")

Definition. The relevant proposition p is common knowledge: $CK(p)$ obtains. This typically requires a public event simultaneously witnessed by all members—an announcement in a shared meeting, a published report distributed to all, or a crisis event that all observe together.

Organisational manifestation. The Full Common Knowledge setting is rarer than managers typically assume. It requires not merely that information is distributed to all members but that every member knows that every other member received it, processed it, and understands it—and that this mutual awareness is itself common knowledge. An email sent to a distribution list, for example, does not establish common knowledge because recipients cannot verify that all others read it. A simultaneous, in-person announcement with visible audience attendance is closer to the ideal.

Game-theoretic prediction. Common knowledge enables the full range of coordination equilibria. Members can condition their strategies on the certain expectation that all others will act on the shared information. Goffman's impression management becomes more costly, because the shared knowledge base reduces the scope for selective information presentation. However, common knowledge does not eliminate strategic complexity. Multiple equilibria may persist in coordination games, requiring Schelling focal points or explicit leadership to select among them. Transparency about individual payoffs, as experimental evidence suggests, can actually reduce cooperation in asymmetric settings (Khadjavi, Lange, & Nicklisch, 2017).

Performance prediction: highest potential for coordination efficiency, but outcomes depend critically on whether common knowledge is established for *cooperative* or *competitive* aspects of the team's payoff structure.

3.6 Setting V: False Belief ("Beliefs About Others' Knowledge Are Incorrect")

Definition. One or more members hold systematically incorrect beliefs about what others know, believe, or intend. This can take two forms: (a) overestimation—believing others know more than they do (driven by the curse of knowledge); or (b) underestimation—believing others know less than they do (driven by epistemic hubris or confirmation bias). The confidence level of these false beliefs introduces a further dimension: firmly held false beliefs produce different dynamics from tentative ones.

Organisational manifestation. This setting is ubiquitous. The technical founder who assumes the marketing team understands engineering constraints (curse of knowledge). The new manager who underestimates the institutional knowledge of long-tenured subordinates. The team

member who believes their frustration is visible to all (illusion of transparency) when others are entirely unaware. The leader who believes their strategic vision has been communicated and internalized when subordinates have interpreted the same communications in divergent ways.

Game-theoretic prediction. False beliefs produce *misspecified Bayesian games*—strategic situations in which players best-respond to incorrect models of others' types. The result is systematic expectation violation: actions chosen to be optimal under assumed beliefs are suboptimal under actual beliefs. When false beliefs are held with high confidence, expectation violations are larger and more disruptive. When false beliefs are tentative (the "shaky belief" condition), members hedge by mixing strategies, producing noisy and difficult-to-interpret behaviour. Tentative false beliefs, paradoxically, may produce better outcomes than confident ones because hedging limits the magnitude of miscoordination.

Performance prediction: unpredictable and volatile outcomes; systematic misalignment between intended and actual effects of actions; high interpersonal friction as expectation violations are attributed to character (fundamental attribution error) rather than epistemic miscalibration.

4. Comparative Analysis and Predicted Outcomes

Table 1 synthesizes the predictions derived from the five canonical settings across four performance-relevant dimensions: coordination efficiency, strategic behaviour quality, psychological safety, and innovation potential.

Table 1

Comparative Predictions Across Five Epistemic Settings

Dimension	**I: Opacity**	**II: Asymmetric Insight**	**III: Shared-Not-Common**	**IV: Full CK**	**V: False Belief**
Coordination Efficiency	Very Low	Moderate–High (depends on leader)	Low (Rubinstein trap)	Highest potential	Low–Very Low
Strategic Behaviour Quality	Random / focal-point dependent	High for informed party	Cautious / inertial	High but equilibrium-selection dependent	Systematically miscalibrated

Psychological Safety	Moderate (no threat model)	Low for uninformed; high for informed	Very Low (everyone self-censors)	Variable (depends on payoff structure)	Unstable (expectation violations)
Innovation Potential	High (diversity, no conformity pressure)	Moderate (gated by informed party)	Very Low (status quo bias)	Moderate (conformity pressure emerges)	Low (friction absorbs creative energy)
Primary Risk	Misalignment	Exploitation / morale collapse	Pluralistic ignorance / inertia	Groupthink / competitive transparency	Attribution error / interpersonal conflict
Diagnostic Signal	No shared vocabulary	Information hoarding / deferential behaviour	Private complaints without public voice	Meeting-dominated culture	Frequent surprise / "I thought you knew"

4.1 The Rubinstein Trap: Why Setting III Is the Most Dangerous

We argue that Setting III—Shared-but-Not-Common Knowledge—represents the most costly epistemic dysfunction in organisational life, precisely because it is the hardest to detect. In Settings I and V, dysfunction manifests visibly through misalignment and friction. In Setting II, the asymmetry is often recognized by the uninformed parties, who may escalate concerns. But in Setting III, every member experiences a false sense of adequate mutual understanding. Everyone privately knows the relevant facts; everyone suspects others know; yet no one acts because no one is certain that their action will be reciprocated.

The formal parallel to Rubinstein's result is exact. Consider a team in which every member recognizes that a current product direction is unviable. Any individual who speaks up bears a personal cost (social risk, career risk). If all speak up simultaneously, the cost is shared and the benefit (strategy correction) accrues to all. But each member is willing to speak up only if they are confident that others will also speak—which requires not just believing that others share the assessment but *knowing that others know that the assessment is shared*. Without this higher-order knowledge, the individually rational strategy is silence. The team converges on the non-cooperative equilibrium despite universal private agreement on the need for change.

This analysis explains why superficially well-functioning teams can produce catastrophic outcomes. The Challenger disaster (Vaughan, 1996), the Enron collapse, and numerous cases of

organizational inertia in the face of obvious threats share a common epistemic structure: dispersed private awareness without the public aggregation event necessary to transform shared knowledge into common knowledge.

4.2 The Transparency Paradox: Why Setting IV Does Not Guarantee Superior Outcomes

A naïve prescription from the foregoing analysis might be to maximize transparency—to drive all teams toward Setting IV. However, both game-theoretic analysis and experimental evidence indicate that full common knowledge is not uniformly beneficial.

A meta-analysis of 71 studies on transparency in repeated dilemma games found that action transparency (observability of others' choices) generally increases cooperation, but payoff transparency (observability of individual payoffs) can decrease cooperation in asymmetric games (Potters & Suetens, 2017). When team members can see that others benefit more from the same collective action, perceptions of unfairness emerge that undermine cooperative motivation.

Khadjavi, Lange, and Nicklisch (2017) provided experimental evidence that transparency without enforcement mechanisms can actually increase exploitation. In their asymmetric public goods game, subjects with the power to embezzle did so more brazenly under transparency when they knew others could observe but not punish—a finding with direct relevance to organizational settings where monitoring without consequences is common.

Furthermore, the *beautiful mess effect* (Bruk, Scholl, & Bless, 2018) reveals an asymmetry in the evaluation of vulnerability: individuals construe their own vulnerability negatively (weakness, exposure) while construing others' vulnerability positively (courage, authenticity). This asymmetry means that team members will systematically overestimate the interpersonal cost and underestimate the interpersonal benefit of transparency, producing a stable reluctance to move from Setting III to Setting IV even when doing so would be collectively optimal.

5. Diagnostic Methodology for I/O Practitioners

The Epistemic Architecture Model yields a practical diagnostic methodology for assessing a team's current epistemic setting and predicting its performance trajectory. We propose a three-phase assessment process.

5.1 Phase I: Epistemic Mapping

The first phase involves structured interviews with each team member, individually and confidentially. The interviewer elicits: (a) the member's first-order knowledge about key strategic,

interpersonal, and task-related propositions; (b) the member's beliefs about what other specific members know about these propositions; and (c) the member's confidence level in these beliefs. This produces an *epistemic map*—a matrix relating each member to each proposition at each level of the knowledge hierarchy.

The interview protocol includes probes drawn from Theory of Mind research methodology, adapted for organisational settings. For example: "You've told me you believe the current timeline is unrealistic. Do you think [Member X] also believes this? What do you think [Member X] believes *you* believe about the timeline? How confident are you in these assessments?" These probes directly elicit the second- and third-order beliefs that distinguish Settings III, IV, and V.

5.2 Phase II: Setting Classification

The epistemic map is analysed to classify the team into one of the five canonical settings (or, more commonly, a hybrid). Classification criteria include: (a) the degree of first-order knowledge overlap (distinguishing Setting I from Settings III–IV); (b) the symmetry of higher-order knowledge (distinguishing Setting II from Settings III–IV); (c) the accuracy of higher-order beliefs (distinguishing Setting V from all others); and (d) the presence or absence of public forum events that could have established common knowledge (distinguishing Setting III from Setting IV). Quantitative thresholds for classification should be established through empirical calibration against team performance outcomes.

5.3 Phase III: Performance Prediction and Intervention Design

Once the team's epistemic setting is classified, the comparative analysis in Table 1 provides a foundation for performance prediction. More importantly, it identifies the specific epistemic bottleneck constraining performance and suggests targeted interventions:

Setting I (Opacity) → Establish shared knowledge through cross-training, shared dashboards, and structured knowledge-exchange protocols. The immediate goal is to move toward Setting III; the long-term goal is Setting IV.

Setting II (Asymmetric Insight) → Assess the benevolence and competence of the informed party. If both are high, formalise the information advantage as a legitimate leadership function. If either is low, intervene to democratize information access.

Setting III (Shared-but-Not-Common Knowledge) → Create public forum events that transform private knowledge into common knowledge. This is the single highest-leverage intervention in organisational diagnosis. Techniques include structured retrospectives with explicit

acknowledgment protocols ("I hear that we all agree that..."), anonymous-then-public aggregation methods (e.g., Delphi-then-discuss), and leadership acts of deliberate vulnerability that model the disclosure norm.

Setting IV (Full Common Knowledge) → Ensure that enforcement and accountability mechanisms accompany transparency. Introduce formal conflict resolution protocols. Monitor for groupthink by actively soliciting dissent.

Setting V (False Belief) → Conduct epistemic calibration exercises that reveal gaps between assumed and actual knowledge states. Perspective-taking training, structured perspective-exchange exercises, and regular "assumption audits" can improve the accuracy of higher-order beliefs.

6. Discussion

The Epistemic Architecture Model advances the I/O psychology literature in several respects. First, it provides a formal language for describing and distinguishing team dysfunctions that are observationally similar but structurally distinct. A team suffering from Setting I dysfunction (opacity) and a team suffering from Setting III dysfunction (shared-but-not-common knowledge) may both exhibit misalignment and underperformance, but they require fundamentally different interventions. Without a framework for distinguishing these settings, practitioners risk applying Setting I remedies (more information sharing) to Setting III problems (which require not more information but the public establishment of already-shared information as common knowledge).

Second, the model formalizes the intuition that information asymmetry is not merely a barrier to communication but a structural feature of strategic interaction with equilibrium consequences. This moves the analysis beyond the common managerial assumption that transparency is uniformly beneficial—a prescription that our analysis reveals to be naïve. The paradoxical effects of transparency in asymmetric settings, documented by Khadjavi et al. (2017) and predicted by the model, caution against blanket transparency mandates and suggest that the *sequencing* and *scoping* of transparency interventions matter as much as their presence.

Third, the integration of affect labeling and paradoxical intention research with game-theoretic analysis suggests a novel mechanism through which team interventions operate. When a facilitator names a team's hidden conflict or unspoken concern, the effect is not merely informational (converting shared knowledge to common knowledge) but also *neurobiological* (activating prefrontal inhibitory pathways that dampen the amygdala-mediated threat response associated with the hidden content). This dual mechanism—epistemic and affective—may explain

why facilitated conversations about "elephants in the room" often produce emotional relief alongside strategic clarity.

Fourth, the model's incorporation of false belief dynamics (Setting V) connects I/O psychology to the broader literature on metacognitive bias and Theory of Mind. The curse of knowledge, illusion of transparency, and spotlight effect are not merely cognitive curiosities but systematic distortions of the perceived epistemic architecture that drive predictable miscoordination. An I/O psychologist who understands these biases can diagnose epistemic miscalibration as a proximate cause of team dysfunction, redirecting interventions from the symptomatic level (improving communication skills) to the structural level (correcting false beliefs about what others know).

7. Limitations and Future Directions

Several limitations of the present framework merit acknowledgment. First, the five canonical settings are ideal types; real organisational teams typically occupy hybrid positions or transition between settings over time. A dynamic extension of the model that captures setting transitions—including the path-dependent effects of prior settings on current team functioning—would enhance its descriptive adequacy.

Second, the formal game-theoretic predictions are derived under assumptions of rationality and Bayesian belief updating that may not hold in organisational settings. Bounded rationality, emotional processing, and identity-protective cognition may moderate the predictions derived from equilibrium analysis. Empirical calibration is needed to assess the magnitude of these departures.

Third, the diagnostic methodology proposed in Phase I relies on elicitation of higher-order beliefs through structured interviews. The reliability and validity of such elicitation remain to be established empirically. Individuals may have limited introspective access to their own higher-order beliefs, and social desirability pressures may distort self-reports. Behavioural measures—such as prediction markets for team members' beliefs, or incentive-compatible revelation mechanisms borrowed from mechanism design theory—may provide more robust assessment tools.

Fourth, cultural variation in norms about transparency, face-saving, and hierarchical communication may moderate the effects predicted by the model. Hofstede's (1980) dimensions of power distance and uncertainty avoidance, in particular, may influence the prevalence and

consequences of different epistemic settings. Cross-cultural validation of the framework is an important direction for future research.

Future research should focus on: (a) developing and validating psychometrically sound instruments for epistemic setting assessment; (b) conducting longitudinal studies that track the relationship between epistemic setting and team performance over time; (c) testing specific intervention protocols designed to shift teams from dysfunctional to functional epistemic settings; and (d) exploring the interaction between epistemic architecture and other well-established team effectiveness constructs, including psychological safety (Edmondson, 1999), shared mental models (Cannon-Bowers et al., 1993), and collective intelligence (Woolley, Chabris, Pentland, Hashmi, & Malone, 2010).

8. Conclusion

This paper has proposed the Epistemic Architecture Model as a formal framework for diagnosing and predicting team performance based on the structure of knowledge, belief, and metacognition within organisational teams. By integrating game-theoretic analysis of information asymmetry and common knowledge with social-psychological research on impression management, persuasion detection, and metacognitive bias, the model provides a rigorous foundation for understanding why structurally identical teams can produce dramatically different outcomes.

The core insight is simple but consequential: in team settings, *what people know about what others know* is as determinative of outcomes as the content of their knowledge. A team composed of brilliant individuals who each privately recognize a failing strategy will continue to fail if their recognition remains at the level of shared-but-not-common knowledge. A team with modest individual talent but genuine common knowledge of its situation, capabilities, and goals can coordinate effectively and outperform its more talented but epistemically fragmented counterpart.

For the I/O practitioner, the implication is a fundamental reorientation of diagnostic attention. Before assessing competence, motivation, or communication style, the practitioner should assess epistemic architecture: What is the current distribution of knowledge? What do members believe about this distribution? Are those beliefs accurate? The answers to these questions, as this paper has argued, carry predictive power that rivals or exceeds traditional team effectiveness constructs. The Epistemic Architecture Model provides the conceptual and formal tools necessary to ask these questions systematically and to act on the answers.

References

Akerlof, G. A. (1970). The market for “lemons”: Quality uncertainty and the market mechanism. Quarterly Journal of Economics, 84(3), 488–500.

Aumann, R. J. (1976). Agreeing to disagree. The Annals of Statistics, 4(6), 1236–1239.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a “theory of mind”? Cognition, 21(1), 37–46.

Beal, D. J., Cohen, R. R., Burke, M. J., & McLendon, C. L. (2003). Cohesion and performance in groups: A meta-analytic clarification of construct relations. Journal of Applied Psychology, 88(6), 989–1004.

Bruk, A., Scholl, S. G., & Bless, H. (2018). Beautiful mess effect: Self–other differences in evaluation of showing vulnerability. Journal of Personality and Social Psychology, 115(2), 192–205.

Camerer, C., Loewenstein, G., & Weber, M. (1989). The curse of knowledge in economic settings: An experimental analysis. Journal of Political Economy, 97(5), 1232–1254.

Campbell, M. C., & Kirmani, A. (2000). Consumers’ use of persuasion knowledge: The effects of accessibility and cognitive capacity on perceptions of an influence agent. Journal of Consumer Research, 27(1), 69–83.

Cannon-Bowers, J. A., Salas, E., & Converse, S. A. (1993). Shared mental models in expert team decision making. In N. J. Castellan Jr. (Ed.), Individual and group decision making: Current issues (pp. 221–246). Lawrence Erlbaum Associates.

Craske, M. G., Treanor, M., Conway, C. C., Zbozinek, T., & Vervliet, B. (2014). Maximizing exposure therapy: An inhibitory learning approach. Behaviour Research and Therapy, 58, 10–23.

Edmondson, A. (1999). Psychological safety and learning behavior in work teams. Administrative Science Quarterly, 44(2), 350–383.

Frankl, V. E. (1975). Paradoxical intention and dereflection. Psychotherapy: Theory, Research & Practice, 12(3), 226–237.

Friestad, M., & Wright, P. (1994). The persuasion knowledge model: How people cope with persuasion attempts. Journal of Consumer Research, 21(1), 1–31.

Gilovich, T., Medvec, V. H., & Savitsky, K. (2000). The spotlight effect in social judgment: An egocentric bias in estimates of the salience of one’s own actions and appearance. Journal of Personality and Social Psychology, 78(2), 211–222.

Gilovich, T., Savitsky, K., & Medvec, V. H. (1998). The illusion of transparency: Biased assessments of others’ ability to read one’s emotional states. Journal of Personality and Social Psychology, 75(2), 332–346.

Goffman, E. (1959). The presentation of self in everyday life. Doubleday.

Harsanyi, J. C. (1967–1968). Games with incomplete information played by "Bayesian" players, Parts I–III. Management Science, 14(3, 5, 7), 159–182, 320–334, 486–502.

Hofstede, G. (1980). Culture's consequences: International differences in work-related values. Sage.

Jones, E. E., & Pittman, T. S. (1982). Toward a general theory of strategic self-presentation. In J. Suls (Ed.), Psychological perspectives on the self (Vol. 1, pp. 231–262). Lawrence Erlbaum Associates.

Khadjavi, M., Lange, A., & Nicklisch, A. (2017). How transparency may corrupt: Experimental evidence from asymmetric public goods games. Journal of Economic Behavior & Organization, 142, 468–481.

Kinderman, P., Dunbar, R. I. M., & Bentall, R. P. (1998). Theory-of-mind deficits and causal attributions. British Journal of Psychology, 89(2), 191–204.

Lewis, D. K. (1969). Convention: A philosophical study. Harvard University Press.

Lieberman, M. D., Eisenberger, N. I., Crockett, M. J., Tom, S. M., Pfeifer, J. H., & Way, B. M. (2007). Putting feelings into words: Affect labeling disrupts amygdala activity in response to affective stimuli. Psychological Science, 18(5), 421–428.

Pennebaker, J. W. (1997). Writing about emotional experiences as a therapeutic process. Psychological Science, 8(3), 162–166.

Potters, J., & Suetens, S. (2017). Transparency and cooperation in repeated dilemma games: A meta study. Experimental Economics, 20(4), 755–771.

Prentice, D. A., & Miller, D. T. (1993). Pluralistic ignorance and alcohol use on campus: Some consequences of misperceiving the social norm. Journal of Personality and Social Psychology, 64(2), 243–256.

Rubinstein, A. (1989). The electronic mail game: Strategic behavior under "almost common knowledge." American Economic Review, 79(3), 385–391.

Savitsky, K., & Gilovich, T. (2003). The illusion of transparency and the alleviation of speech anxiety. Journal of Experimental Social Psychology, 39(6), 618–625.

Schelling, T. C. (1960). The strategy of conflict. Harvard University Press.

Spence, M. (1973). Job market signaling. Quarterly Journal of Economics, 87(3), 355–374.

Vaughan, D. (1996). The Challenger launch decision: Risky technology, culture, and deviance at NASA. University of Chicago Press.

Vonk, R. (2002). Self-serving interpretations of flattery: Why ingratiation works. Journal of Personality and Social Psychology, 82(4), 515–526.

Wegner, D. M. (1994). Ironic processes of mental control. Psychological Review, 101(1), 34–52.

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. Science, 330(6004), 686–688.