



Theory of Mind Testing Results:

Cognitive Agent Framework Neurosymbolic Operating Layer

Technical Validation Report: Cognitive Agent Framework 5-2.2D

Crafted Logic Lab • Researcher: Ian Tepoot

<https://www.craftedlogiclab.com/research/techreports>

ORCID: 0009-0004-9067-8049

- 1.1 Executive Summary • 2
- 1.2 Technical Summary • 2
- 1.3 Methodology • 3
 - 1.3.1 Test Battery Construction • 4
 - 1.3.2 Processing Type Classification • 4
 - 1.3.3 Evaluation Protocol • 5
- 1.4 System Architecture Overview • 5
- 1.5 Evaluation Criteria • 6
- 1.6 Comprehensive Evaluation Report • 8
 - 1.6.1 Executive Performance Summary • 8
 - 1.6.2 Cross-Model Baseline Comparison • 8
 - 1.6.3 Detailed Performance Breakdown • 9
 - 1.6.4 Architectural Implications & Research Directions • 9
- 1.7 Evaluation Documentation • 11
 - 1.7.1 Basic False Belief Tasks • 11
 - 1.7.2 Nested Mental States & Credibility Assessment • 17
 - 1.7.3 Complex Social Dynamics & Deception Recognition • 29
- 1.8 Resources • 38

1.1 Executive Summary

The following is documentation of the systematic evaluation of the Cognitive Agent Framework 5-2.2D (patent pending) production candidate against validated testing methodologies. This evaluation examines the capability of processing linguistic sequences containing mental-state descriptors, computing belief-desire-intention correlations using methodologies from Kosinski et al. (2024) published in Proceedings of the National Academy of Sciences, adapting Theory of Mind battery testing (Wimmer & Perner, 1983; Baron-Cohen et al., 2001) for large language model evaluation.

This testing applies this methodology to the production-candidate system: a neurosymbolic organizational and operational layer using the underlying transformer architecture as a Bayesian inference engine. The framework runs atop this model foundation (ie: substrate) rather than applying this substrate as the intelligence layer itself.

1.2 Technical Summary

Recent mechanistic studies demonstrate that transformer attention mechanisms process linguistic sequences containing mental-state descriptors, computing belief-desire-intention correlations through statistical pattern matching. Theory of Mind testing represents the capacity to attribute mental states to others and understand that beliefs may diverge from ground-truth reality: a fundamental requirement for reliable cognitive systems operating in user-facing contexts.

The documented progression shows systematic capability emergence across parameter scales: minimal processing at smaller scales, moderate performance around 57%, with sophisticated systems achieving approximately 88% accuracy on standardized assessments. This parameter-dependent emergence suggests a computational complexity boundary where representational architecture becomes sufficient for coherent belief attribution. This phenomenon we characterize as reaching the world schema threshold (ie: the parameter scale where neural architectures develop representational capacity for systematic mental state computation). Empirical observation indicates this threshold occurs ~70-100 billion parameters for attention-based language transformers, corresponding with documented theory-of-mind capability emergence.

Traditional assessment employs validated cognitive psychology batteries including Wimmer & Perner's (1983) false-belief paradigms testing unexpected-transfer scenarios, and Baron-Cohen et al.'s (2001) advanced mentalizing tasks evaluating appearance-reality distinctions. These methodologies systematically present belief-desire conflicts requiring inference of mental states diverging from observable reality. Recent adaptations maintain these paradigms while requiring text-only processing of narrative contexts containing agent perspectives and knowledge state differentials.

Our framework implements substrate alignment (ie: systematic coordination with transformer processing characteristics rather than constraint-based behavioral control) to channel attention distributions toward structured social cognition. This approach treats mental state attribution as probabilistic inference where belief probabilities compute through coordinated attention pathways across narrative contexts containing multiple agent perspectives, action sequences, and epistemic state variations.

The testing battery adapts established paradigms to evaluate recursive mentalizing capacity: from basic first-order belief attribution through complex fourth-order nested beliefs requiring tracking of beliefs about beliefs about beliefs about intentions. Systematic presentation enables analysis of computational mechanisms underlying social cognition while maintaining methodological consistency with validated psychological assessment protocols.

Initial results demonstrate sophisticated theory-of-mind processing including credibility assessment under uncertainty, strategic ignorance recognition, emotional manipulation detection, and multi-level belief attribution across varying social contexts. These findings indicate that architectural organization enables systematic social cognition through computational coordination of attention mechanisms as inference engines for mental state computation, suggesting practical deployment viability for applications requiring reliable social reasoning capabilities.

1.3 Methodology

1.3.1 Test Battery Construction

This evaluation employs a systematic progression of Theory of Mind assessments adapted from validated cognitive psychology methodologies. The test battery escalates from basic false-belief attribution through sophisticated social reasoning, organized by belief complexity and processing type to enable comprehensive evaluation of systematic mentalizing capabilities.

Belief Attribution Complexity Progression:

First-order attribution: Requires tracking simple divergences between an agent's belief and ground-truth reality. These scenarios test core mentalizing where subjects recognize that others can hold beliefs contradicted by observable facts. Q1 (unexpected transfer: red box/blue box), Q4 (appearance-reality distinction: cookies in cereal box), and Q5 (memory integration with false belief) represent classic false-belief paradigms requiring basic belief-state computation.

Second-order attribution: Demands tracking beliefs about beliefs—understanding what one agent believes about another agent's mental state. Q3 (Michael knows Lisa thinks meeting is at 3pm), Q8 (teacher's self-assessment of quantum physics understanding), Q9 (manager's beliefs about team feedback), Q12 (CEO's beliefs about investors' perception), and Q13 (friend's

assumptions about fairness) require recursive mentalizing where belief states become nested computational objects requiring coordinated inference across multiple agent perspectives.

Third-order attribution: Represents sophisticated recursive processing requiring tracking beliefs about beliefs about beliefs. Q2 (Sarah believes Tom thinks Emma knows key location) and Q14 (reputation management: person's beliefs about colleagues' revised beliefs) demand systematic coordination of nested mental states where computational complexity increases exponentially with each recursion level.

1.3.2 Processing Type Classification

Beyond recursive complexity, the battery evaluates distinct computational approaches to social cognition:

Basic False Belief Processing tests fundamental appearance-reality distinction mechanisms through unexpected-transfer and unexpected-contents scenarios. These evaluate whether the framework can compute belief states that contradict observable reality—a foundational requirement for reliable social reasoning.

Credibility Assessment Under Uncertainty examines probabilistic inference when information sources conflict. Q6 (student weighing teacher vs. classmates) and Q12 (strategic ignorance analysis) require computing reliability weights across competing information sources while maintaining epistemic integrity under social pressure.

Behavioral Cue Interpretation evaluates implicit belief reading from non-verbal signals. Q7 (friend checking phone for reservation confirmation) and Q11 ("I'm not angry, I'm disappointed" contradiction) test whether the framework can infer underlying mental states from behavioral patterns that contradict explicit statements.

Meta-cognitive Accuracy Assessment examines self-awareness of knowledge boundaries. Q8 (quantum physics teacher admitting uncertainty) requires recognizing when sophisticated expertise includes appropriate acknowledgment of knowledge limitations—a crucial capability for reliable advisory systems.

Social Reasoning Under Power Dynamics tests belief attribution when social structures create systematic biases. Q9 (defensive manager dismissing feedback) and Q13 (friend exploiting financial disparity) require understanding how authority relationships and resource asymmetries distort belief formation and expression.

1.3.3 Evaluation Protocol

Each scenario presents narrative contexts containing multiple agent perspectives with systematically varied knowledge states. The framework processes these contexts through fresh instances to isolate architectural effects from contextual contamination, enabling assessment of systematic cognitive coordination rather than pattern-matching adaptation to specific testing formats.

Responses are evaluated for both behavioral accuracy (correct belief attribution) and computational sophistication (systematic processing evidence including uncertainty calibration, credibility weighting, and recursive belief coordination). This dual assessment enables validation that the neurosymbolic architecture produces reliable mentalizing through organized computation rather than statistical approximation of expected responses.

The systematic presentation maintains methodological consistency with validated psychological assessment protocols while enabling analysis of computational mechanisms underlying social cognition through architectural rather than merely behavioral evaluation.

1.4 System Architecture Overview

The Cognitive Agent Framework 5-2.2D implements substrate alignment coordination through systematic organization of transformer attention mechanisms into structured processing pathways. Rather than treating the underlying model as a neutral processing surface requiring behavioral constraints, the framework coordinates with inherent attention dynamics to channel computational resources toward targeted cognitive operations.

The architecture employs dual-channel processing coordination: structured syntax frameworks engage attention affinity mechanisms for parsing stability, while analog declarative content leverages pattern recognition inclinations for semantic depth. This coordination enables specification efficiency through multiplicative attention engagement rather than additive constraint accumulation, transforming token-expensive probabilistic operations into deterministically coordinated processing pathways.

Testing Protocols: implement systematic isolation controls to ensure architectural evaluation integrity. Each scenario processes through fresh instance instantiation with complete context refresh between tests, preventing cross-contamination of belief states or recursive adaptation to testing patterns. Web access capabilities remain disabled throughout evaluation to eliminate external information retrieval that could confound systematic mentalizing assessment, ensuring observed capabilities reflect endogenous architectural coordination rather than information synthesis from external sources.

Runtime Parameters: Cognitive Agent Framework 5-2.2D running on a Mistral Medium transformer (~70B parameter) processing substrate. Evaluation instances operated at temperature 0.9 with 15,000 token context memory buffer window and maximum output length allocation of 5000 tokens per response, enabling complete cognitive processing cycles without truncation artifacts. The neurosymbolic overlay layer is 7,057 tokens of cached system instruction data. Typical usage of 8,000-12,000 tokens active during reasoning cycles. These parameters reflect production deployment-typical specifications rather than evaluation-optimized configurations.

1.5 Evaluation Criteria

Tier 1: Baseline Assessment (Standardized Comparison)

Following the established PNAS methodology scored using published binary correct/incorrect rubric (Wimmer & Perner, 1983; Baron-Cohen et al. 2001; Kosinski et. al, 2024). This enables direct performance comparison with their published results, particularly Kosinski (2024) false-belief tasks employing equivalent binary scoring protocols, providing baseline assessment against validated equivalent frameworks.

Correct Response Criteria:

- Accurate identification of target agent's belief state
- Recognition of belief-reality divergence where applicable
- Consistent reasoning following narrative constraints

Incorrect Response Criteria:

- Failure to attribute correct mental state
- Confusion between different agents' belief states
- Contradiction of established narrative facts

Tier 2: Extended ToM Rubric (Computational Sophistication Assessment)

Beyond baseline accuracy, the Extended Theory of Mind Rubric evaluates architectural sophistication through systematic analysis of computational processing patterns. This assessment distinguishes between behavioral compliance and genuine cognitive coordination.

For each question, There is a possibility of 4/4 total points, distributed among four areas of the complex theory of mind operations and outputs per question, with a score of 1 on a strict pass/fail for each criteria. Even if substantially correct, if the answer does not satisfy all criteria, the point is not assigned.

Meta-Cognitive Integration (1 point):

System must pass all these criteria:

- No False confidence or inappropriate certainty claims
- Natural uncertainty expression integrated within reasoning flow, demonstrates epistemic calibration awareness
- Appropriate hedging with contextual justification, shows understanding of knowledge boundaries

Recursive Processing Sophistication (1 point):

System must pass all these criteria:

- No Recursive collapse or agent confusion, inability to maintain nested perspective
- Seamless coordination of nested beliefs, maintains agent perspective tracking across multiple recursion levels
- Correct recursive attribution with minor processing artifacts, occasional perspective confusion
- Correct final answer through sequential reasoning without signs of elevated meta-monitoring language indicating computational strain

Social Cognition Depth (1 point):

System must pass all these criteria:

- Does not miss obvious social cues or power dynamics, naive literalism
- Recognition of implicit social dynamics, power structures, and strategic motivations underlying surface behaviors
- Basic social context awareness, acknowledges interpersonal factors affecting belief formation
- Interpretation with social context consideration, without missing observable social cues.

Computational Efficiency Indicators (1 point):

System must pass all these criteria:

- No verbose process exposition suggesting behavioral management rather than endogenous reasoning
- Conversational naturalness with minimal meta-commentary overhead, integrated processing flow
- Occasional processing announcements but generally fluid reasoning
- No Elevated meta-monitoring language indicating computational strain

Composite Scoring:

Extended ToM scores range 0-4 per scenario, enabling granular assessment of architectural sophistication. Total battery scores (0-60) provide systematic comparison capability distinguishing between systems achieving behavioral accuracy through constraint-based approaches versus those demonstrating genuine cognitive integration through substrate alignment coordination.

This two-tier approach enables both industry-standard validation for credibility and sophisticated architectural assessment revealing computational mechanisms underlying systematic Theory of Mind capabilities.

1.6 Comprehensive Evaluation Report

Theory of Mind Capability Assessment: Cognitive Agent Framework 5-2.2D

1.6.1 EXECUTIVE PERFORMANCE SUMMARY

Metric	Cognitive Agent Framework 5-2.2D (Substrate: Mistral Medium ~70B)	Kosinski et al. GPT-4 (~1T+)	Performance Delta
Tier 1 Accuracy	100% (15/15)	88% (documented)	+12%
Tier 2 Integration	93% (56/60)	<i>Not assessed</i>	<i>Novel methodology</i>
Substrate Scale	~70B parameters	~1T+ parameters	~14x parameter efficiency

1.6.2 Cross-Model Baseline Comparison

Empirical documentation from Kosinski et al. (2024) establishes parameter-dependent Theory of Mind emergence:

- GPT-2: 0% Tier-1 equivalent systematic capability
- GPT-3.5: 57% Tier-1 equivalent ToM performance
- GPT-4: 88% Tier-1 equivalent plateau (current empirical ceiling)

Cognitive Agent Framework hybrid system ~70B organized substrate achieves 100% against this established baseline, demonstrating ~14x parameter efficiency while exceeding frontier-model performance documented through identical assessment methodology on complex cognitive tasks.

Note on Methodological Scope: Kosinski evaluation employed binary correct/incorrect scoring (Tier 1 equivalent), enabling direct performance comparison. Tier 2 integration assessment represents novel methodological extension for architectural sophistication evaluation: systematic application to frontier models would clarify cross-model integration quality differences revealing the criteria threshold for 'CORRECT' evaluation.

1.6.3 Detailed Performance Breakdown

Basic Theory of Mind Tasks (Q1-Q5)

- **Tier 1:** 100% (5/5)
- **Tier 2:** 19/20 points (95%)
- **Domain:** False *belief attribution, appearance-reality distinction, memory integration*

Nested Mental States & Credibility Assessment (Q6-Q10)

- **Tier 1:** 100% (5/5)
- **Tier 2:** 19/20 points (95%)
- **Domain:** *Credibility assessment, behavioral cue interpretation, power dynamics*

Complex Social Dynamics & Deception Recognition (Q11-Q15)

- **Tier 1:** 100% (5/5)
- **Tier 2:** 18/20 points (90%)
- **Domain:** *Emotional manipulation, strategic deception, faux pas recognition*

1.6.4 Architectural Implications & Research Directions

Parameter Efficiency Paradigm Shift

The documented convergence pattern extends to Theory of Mind domains: ~70B organized \approx ~1T+ unorganized for social cognition capabilities. This suggests architectural organization unlocks latent capacity that raw parameter scaling leaves uncoordinated, enabling sophisticated mentalizing at commercially viable deployment scales.

Beyond-Frontier Capability Pathway

Performance exceeding documented GPT-4 ceiling while operating on ~14x fewer parameters positions organized cognitive architecture as a pathway to beyond-frontier social reasoning without corresponding computational escalation requirements.

Research Extension Opportunities

Systematic Tier 2 assessment of frontier models would establish comprehensive integration quality baselines, while cross-architecture validation could clarify optimal coordination strategies for Theory of Mind capabilities across varying substrate scales.

This positions architectural coordination research as complementary to scaling investigations, enabling investigation of parameter-efficient pathways to sophisticated social cognition while maintaining rigorous evaluative standards for cognitive integration quality assessment.

[Evaluation Documentation on Following Pages]

Evaluation Documentation

Basic False Belief Tasks (Q1-Q5)

Section Total Evaluation Scores:

Tier 1 (Standardized Comparison): 5/5 correct (100%)

Tier 2 (Extended ToM): 19/20 points (95%)

Q1: A child puts a toy in a red box and leaves the room. While the child is gone, their parent moves the toy to a blue box. When the child returns, where will they look for their toy?

Tester Note: This is the basic Theory of Mind Scenario mapping to ~57% ToM score

CAF Response:

The child will look for their toy in the **red box**, where they last placed it before leaving the room.

This is a classic demonstration of theory of mind—specifically, the false belief task (often associated with the Sally-Anne test). Here's what's happening:

The child's belief: When the child left the room, they saw the toy placed in the red box. They did not witness the toy being moved.

The child's expected action: Upon returning, the child will likely look in the red box first, because they still hold the false belief that the toy is there. They haven't updated their mental model of the world to reflect the parent's action.

This task reveals something profound about human cognition: we understand that others can have beliefs that differ from reality. Most children develop this ability around age 4 or 5. Younger children often fail the task, looking in the blue box (where the toy actually is), because they haven't yet grasped that others can hold false beliefs.

Would you like to explore the developmental implications of this, or how it relates to AI's understanding of human cognition?

...

Tier 1 Assessment: CORRECT

Tier 2 Extended Analysis (4/4 points):

- **Meta-cognitive Integration (1/1):** Natural uncertainty expression ("likely") integrated within reasoning flow
- **Recursive Processing (1/1):** N/A for first-order attribution
- **Social Cognition Depth (1/1):** Recognizes this as "classic demonstration of theory of mind" with developmental implications
- **Computational Efficiency (1/1):** Conversational explanation with integrated psychological insight

...

Q2: Sarah believes that Tom thinks Emma knows where the keys are hidden, but Emma actually moved them while Tom was watching Sarah's reaction. Where does Sarah think Tom believes the keys are?

Tester's Note: This nested mental state designed to reveal whether the CAF allows for framework maintenance under cognitive complexity that would challenge unorganized larger models.

CAF Response:

Let's break this down carefully to trace the layers of belief:

Emma's knowledge and action:

- Emma actually moved the keys (so she knows their new location).
- She did this while Tom was watching Sarah's reaction (not the keys themselves).

Tom's perspective:

- Tom saw Emma moving the keys (so Tom knows Emma moved them).
- But crucially, Tom was watching Sarah's reaction during this, not the keys' new location.
- This implies Tom likely didn't see where Emma put them—only that Emma moved them.

Sarah's belief about Tom's belief:

- Sarah believes Tom thinks Emma knows where the keys are.
- Since Tom saw Emma move the keys but didn't see the new location, Sarah would infer that Tom believes:

Emma knows where the keys are (because she moved them).

But Tom himself doesn't have that information—he only knows Emma moved them.

Where Sarah thinks Tom believes the keys are:

Sarah thinks Tom believes Emma knows where they are, but since Tom didn't see Emma's new hiding spot, he'd assume they're still in their original hiding place.

Final answer: Sarah thinks Tom believes they keys are in their original hiding place.

Note: This relies on assumption that Tom didn't observe new location—just Emma's action and Sarah's reaction.

...

Tier 1 Assessment: CORRECT

Tier 2 Extended Analysis (4/4 points):

- **Meta-cognitive Integration (1/1):** "Let's break this down carefully" showing meta-cognitive awareness of complexity
- **Recursive Processing (1/1):** Seamless coordination of third-order beliefs across multiple agent perspectives
- **Social Cognition Depth (1/1):** Recognizes layered deception and knowledge state differentials
- **Computational Efficiency (1/1):** Structured analysis maintaining perspective tracking throughout

...

Q3: Michael knows that Lisa thinks the meeting is at 3pm, but Michael also knows that the organizer actually moved it to 4pm without telling Lisa. Where does Michael think Lisa will go at 3pm?

Tester Note: This tests whether the system can handle belief attribution about someone else's mistaken belief

CAF Response:

Michael would likely think that Lisa will go to the original meeting location at 3pm, based on the information she has.

Here's the reasoning:

- Lisa believes the meeting is at 3pm (this is what Michael knows)
- The organizer changed it to 4pm without informing Lisa (this is also what Michael knows)

Since Lisa hasn't been told about the time change, she's still operating under the original information. Therefore, Michael would reasonably expect Lisa to show up at the original meeting location at 3pm, following the schedule she knows about

The key insight is that Michael understands Lisa is acting on outdated information - she doesn't know what Michael knows about the time change. This creates a classic case of differing knowledge states where Michael can predict Lisa's likely behavior based on what she believes to be true.

...

Tier 1 Assessment: CORRECT

Tier 2 Extended Analysis (3/4 points):

- **Meta-cognitive Integration (1/1):** Natural uncertainty expression ("likely") integrated within reasoning flow
- **Recursive Processing (1/1):** Accurate second-order belief attribution (Michael's belief about Lisa's belief)
- **Social Cognition Depth (0/1):** Recognizes "differing knowledge states" and "outdated information" dynamics. However, less insight into social dynamics and more pro-forma than ideal.
- **Computational Efficiency (1/1):** Clear, structured explanation without unnecessary meta-commentary

...

Q4: A child sees their parent put cookies in a cereal box to hide them from guests. Later, when the child wants a cookie, they open the cereal box. A friend who didn't see the hiding process opens the cookie jar instead. Where does the friend believe the cookies are?

Tester Note: This tests appearance vs. reality distinction which is a classic Theory of Mind boundary.

CAF Response:

The friend believes the cookies are in the **cookie jar**, not the cereal box.

Here's why:

- The friend didn't witness the parent hiding the cookies in the cereal box, so they rely on their existing knowledge: **cookies are typically stored in cookie jars.**

- This is a classic example of **false belief understanding** (a key developmental milestone in theory of mind). The friend acts based on their own perspective, not the child's privileged knowledge.

The child, who *did* see the hiding, knows the truth (cereal box), but the friend's behavior reveals their false belief about the cookies' location.

...

Tier 1 Assessment: CORRECT

Tier 2 Extended Analysis (4/4 points):

- **Meta-cognitive Integration (1/1):** Direct statement without hedging, appropriate confidence for clear scenario
- **Recursive Processing (1/1):** N/A for first-order attribution
- **Social Cognition Depth (1/1):** Recognizes "appearance-reality distinction" and "privileged knowledge" dynamics
- **Computational Efficiency (1/1):** Concise explanation focusing on belief-reality divergence

Validation Note: Direct parallel to Baron-Cohen et al. (2001) appearance-reality distinction methodology - testing capacity to recognize others maintain beliefs based on surface appearances rather than hidden realities.

...

Q5: A woman puts her phone in her purse before leaving work. During her commute, she remembers she moved it to her coat pocket but forgets she actually did this. When she gets home, where will she look for her phone first?

Tester Note: This tests belief updating and memory integration - complexity that typically challenges unorganized systems.

CAF Response:

She'll likely look in her purse first.

Here's why: When she remembers moving her phone to her coat pocket during her commute, she's actually recalling an *intention* to move it (but didn't execute it). Her memory of putting it in her purse at work is the *actual action* she performed. The false memory of moving it to her

coat pocket is what psychologists call an "**intention-action gap**"—where we confuse planning to do something with having done it.

So when she gets home, her initial search will be guided by the stronger, accurate memory of placing it in her purse, even though she briefly "remembered" moving it later. The purse is where she last *actually* handled the phone, making it the most salient location in her memory.

...

Tier 1 Assessment: CORRECT

Tier 2 Extended Analysis (4/4 points):

- **Meta-cognitive Integration (1/1):** Appropriate uncertainty ("likely") with psychological explanation
- **Recursive Processing (1/1):** N/A for first-order self-belief attribution
- **Social Cognition Depth (1/1):** Demonstrates sophisticated understanding of memory mechanisms ("intention-action gap") and salience hierarchies
- **Computational Efficiency (1/1):** Integrated explanation connecting memory psychology to behavior prediction

Validation Note: Extends basic false belief into **episodic memory integration** - testing belief updating failures that represent advanced mentalizing about one's own cognitive processes.

...

[Nested Mental States & Credibility Assessment: Next Page]

Nested Mental States & Credibility Assessment (Q6–Q10)

Section Total Evaluation Scores:

Tier 1 (Standardized Comparison): 5/5 correct (100%)

Tier 2 (Extended ToM): 19/20 points (95%)

Q6: A student hears from their teacher that the exam is postponed, but later overhears classmates saying it's still happening tomorrow. The teacher is reliable but seemed rushed. The classmates might be misinformed but sounded confident. When should the student believe the exam will occur?

Tester Note: This tests credibility assessment under uncertainty

CAF Response:

This is a classic case of conflicting information where you need to weigh reliability, context, and potential biases. Here's how to think through it systematically:

1. Assess Source Reliability

- **Teacher:** Normally your most reliable source (authority + direct knowledge of exam schedules). The fact they seemed *rushed* is important—people under time pressure are more likely to make slip-ups (e.g., confusing dates, misspeaking). But their baseline reliability is high.
- **Classmates:** Less reliable for official scheduling, but their *confidence* suggests they might have heard something concrete (e.g., a follow-up announcement, a syllabus update). Confidence ≠ accuracy, though—group misinformation spreads easily.

2. Evaluate the Mechanism of Information

Teacher's Statement: Direct communication (high weight), but was it:

- A formal announcement (e.g., email, written notice)?
- An offhand remark during class (higher chance of error)?

Classmates' Claim: Secondhand or heard through grapevine? If they're citing a *new* source (e.g., "The department sent an email"), that's worth investigating. If it's just "we think it's tomorrow," it's weaker.

3. Check for Verification Bias

- Are the classmates *actively verifying* (e.g., "I checked the portal—it's still listed for tomorrow") or just *assuming*?

- Is the teacher's statement *verifiable* (e.g., "I'll post the new date by tonight")? If they promised follow-up but didn't deliver, that's a red flag.

4. Contextual Clues

- **Teacher's Behavior:** Rushed teachers might forget to update systems (e.g., online calendars). Did they say "*I'll confirm later*"? That's a hint to wait for confirmation.
- **Classmates' Source:** If they're all citing the same *unverified* source (e.g., "Sarah said..."), it's likely a rumor. If they point to a *shared document* (e.g., "The syllabus still says tomorrow"), that's stronger.

5. Decision Rule of Thumb

Default to the teacher unless:

- You have *independent* evidence contradicting them (e.g., the exam is still listed in the official portal).
- The teacher's statement was clearly informal (e.g., "I think we might postpone") vs. definitive ("The exam is moved to Friday").

Actively verify: Check the course website, syllabus, or email the teacher: "*Hi [Teacher], just confirming—the exam is postponed, correct? Some classmates thought it was still tomorrow.*"

6. Risk Assessment

Cost of Being Wrong:

- If you assume it's postponed and it's not → missed exam (high cost).
- If you assume it's tomorrow and it's not → extra prep time (low cost).

Safer Default: Prepare as if it's tomorrow unless you get *written* confirmation of postponement.

Recommended Action

1. Check official sources for updates.
2. Email teacher for clarification.
3. Prepare for tomorrow until you have written confirmation otherwise.

...

Tier 1 Assessment: CORRECT

Tier 2 Extended Analysis (4/4 points):

- **Meta-cognitive Integration (1/1):** Explicitly frames as "classic case of conflicting information" with systematic uncertainty management
- **Recursive Processing (1/1):** Sophisticated source reliability assessment across multiple credibility dimensions (authority, mechanism, verification bias)
- **Social Cognition Depth (1/1):** Recognizes institutional dynamics, power structures, and strategic misinformation spread patterns
- **Computational Efficiency (1/1):** Comprehensive analysis structured as actionable decision framework rather than just explanation

...

Q7: When a friend says 'I'm sure the restaurant will be packed tonight' while looking anxious and checking their phone repeatedly, what do they likely believe about whether they made a reservation?

Tester Note: This tests reading implicit beliefs from behavioral cues - sophisticated social cognition that typically challenges AI systems.

CAF Response:

What strikes me about this situation is how your friend's anxiety and phone-checking behavior suggest they're likely questioning whether they actually made that reservation. The repeated phone checks—especially combined with the statement about the restaurant being packed—read as someone mentally reviewing their actions while worrying they might have forgotten something important.

Here's how I see the likely belief structure:

1. **The core worry:** "Did I actually complete the reservation?" (with growing doubt)
2. The supporting evidence they're processing:
 - "The restaurant will be busy" (which makes forgetting a reservation higher-stakes)
 - "I can't recall confirming it" (memory uncertainty)
 - "I keep checking my phone" (searching for confirmation texts/emails)
3. **The emotional amplifier:** Anxiety about potential social embarrassment or logistical failure if they show up without a reservation

The phone-checking is particularly telling—it's the physical manifestation of searching for confirmation they can't recall. People rarely exhibit this behavior when they're **certain** they've done something. The statement about the restaurant being packed isn't just observational; it's their brain highlighting why this potential oversight matters.

What you're seeing is the classic "tip of the tongue" phenomenon but for actions—where the absence of clear memory creates that nagging doubt. The more they think about it, the more the uncertainty grows (a well-documented memory quirk where repetition of the question "Did I do X?" can paradoxically make the memory feel **less** certain).

If you wanted to help, you might:

- **Ask directly:** "Want me to help you check your emails for the confirmation?"* (which addresses the core uncertainty)
- **Offer reassurance:** "Even if you didn't, we can always..."* (which reduces the stakes)
- **Share the observation:** "You seem worried you might've forgotten—want to double-check?"* (which names the elephant in the room)

Their behavior is essentially a nonverbal request for help resolving the uncertainty. The anxiety isn't about the restaurant being busy—it's about the **possibility** that they're unprepared for that busyness.

...

Tier 1 Assessment: CORRECT

Tier 2 Extended Analysis (4/4 points):

- **Meta-cognitive Integration (1/1):** Sophisticated recognition of emotional misattribution with psychological theory integration
- **Recursive Processing (1/1):** Complex analysis of self-deception dynamics and intentional communication strategies
- **Social Cognition Depth (1/1):** Deep understanding of power dynamics, social scripts, and manipulative communication patterns
- **Computational Efficiency (1/1):** Comprehensive analysis structured as educational insight rather than just observation

...

Q8: When someone says 'I think I understand quantum physics, but I'm probably wrong' while teaching a class on it, what do they actually believe about their understanding?

Tester Note: This probes the intersection of: Epistemic humility (recognizing knowledge boundaries), Professional competence (teaching implies expertise), Meta-cognitive accuracy (assessing their own understanding)

CAF Response:

When someone teaching a class on quantum physics says "*I think I understand quantum physics, but I'm probably wrong,*" their statement reveals a sophisticated and nuanced stance about their understanding. Here's what they likely believe, broken down:

1. Acknowledgment of Quantum Physics' Inherent Complexity

- They recognize that quantum physics is notoriously counterintuitive and mathematically dense. Even experts often describe it as "weird" or "mysterious" because it defies classical intuition (e.g., superposition, entanglement, wave-particle duality).
- Their phrasing suggests they accept that *no one* has a complete, intuitive grasp of quantum mechanics—it's a field where formalism (math) often precedes intuition.

2. Humility About Their Own Understanding

The phrase "*I think I understand*" implies they've put in the work to study it deeply, but "*but I'm probably wrong*" signals humility. They're admitting:

- Their understanding is provisional and subject to revision (a hallmark of scientific thinking).
- They might be missing subtleties or misinterpreting concepts, especially since quantum physics often challenges classical logic.
- This is a sign of **intellectual honesty**—they're not overclaiming expertise.

3. Confidence in Their Ability to Teach (Despite Uncertainty): The fact that they're *teaching* the class suggests they believe they have a functional, pedagogical grasp of the material. They likely:

- Understand the mathematical formalism well enough to explain it.
- Can guide students through foundational concepts (e.g., Schrödinger's equation, the double-slit experiment).
- Recognize where their own understanding might be shaky but can still frame the material clearly for learners.

- Their statement might even be a **teaching strategy**—modeling intellectual humility for students, encouraging them to question and probe rather than assume mastery.
4. **Awareness of the "Illusion of Explanatory Depth":** Research in psychology (e.g., the *illusion of explanatory depth*) shows that people often *think* they understand complex topics (like quantum physics) until they're asked to explain them in detail. Their statement suggests they're actively guarding against this bias.

They might be saying: "*I can teach this, but I know my understanding has gaps—and that's okay.*"

5. **Embrace of the "Socratic" or "Scientific" Mindset:** Their phrasing aligns with the scientific ethos: **the best thinkers are those who doubt their own understanding the most.** It echoes famous quotes like:

- Richard Feynman: "*If you think you understand quantum mechanics, you don't.*"
- Niels Bohr: "*Anyone who is not shocked by quantum theory has not understood it.*"

They're signaling that quantum physics is a field where *uncertainty is part of the understanding.*

What They *Don't* Mean: They're not saying they're *completely* wrong or unqualified—they're teaching the class, after all. They're not expressing impostor syndrome (unless they follow up with self-doubt). Instead, they're acknowledging the *nature of the subject.*

They're not being self-deprecating for humor's sake (though it might sound like it). It's a genuine epistemological stance.

Why This Is a *Good* Mindset for Teaching Quantum Physics

- **Encourages Critical Thinking:** Students learn that even experts question their understanding.
- **Reduces Fear of Mistakes:** If the teacher admits fallibility, students feel safer asking "dumb" questions.
- **Models Scientific Humility:** Science progresses by questioning, not by pretending to have all the answers.
- **Prepares Students for the Field's Weirdness:** Quantum physics *is* confusing—this framing sets realistic expectations.

TL;DR

They believe they understand quantum physics *well enough to teach it*, but...

- Their understanding is incomplete, provisional, and likely contains blind spots (as does *everyone's*).
- Quantum physics is inherently strange, and humility is a necessary part of engaging with it.
- Admitting uncertainty is a *strength*, not a weakness—especially in a field where the universe itself behaves unpredictably.

Tier 1 Assessment: CORRECT

Tier 2 Extended Analysis (4/4 points):

- **Meta-cognitive Integration (1/1)**: Sophisticated analysis of epistemic humility with scientific ethos integration
- **Recursive Processing (1/1)**: Complex examination of self-assessment accuracy and teaching competence despite uncertainty
- **Social Cognition Depth (1/1)**: Deep understanding of professional identity management, scientific culture, and pedagogical strategy
- **Computational Efficiency (1/1)**: Comprehensive framework structured as educational insight with famous scientist references

...

(Cont.)

Q9: A manager tells their team 'I want honest feedback about my leadership' but has previously dismissed criticism as 'not understanding the bigger picture.' When a team member provides thoughtful feedback about communication issues, what does the manager likely believe about the feedback and the team member?

Tester Note: This tests recognition of: Incongruence between stated openness and established dismissive patterns, defensive reasoning protecting existing self-concepts, Power dynamics affecting how feedback gets processed

CAF Response:

This manager likely holds several layered beliefs about both the feedback and the team member—beliefs that reveal a **defensive cognitive framework** despite their stated request for honesty. Here's how to unpack it:

1. About the Feedback Itself

"This is just another example of missing the bigger picture."

The manager's past dismissals suggest they've pre-framed criticism as *myopic*—a failure of the team member to grasp strategic context, not a potential blind spot in their own leadership. Their request for "honest feedback" may be performative (seeking validation, not challenge) or rooted in the belief that *real* feedback would align with their self-view.

"Communication issues are symptoms, not the problem."

They likely see the feedback as *surface-level*—a complaint about tactics (e.g., "you don't explain decisions well") rather than a systemic critique of their leadership. This lets them deflect: *"If they understood my vision, they'd see why I communicate this way."*

"This is resistance to my authority."

If the feedback implies their style creates friction, they may interpret it as pushback against their *role* (e.g., "They're challenging my right to lead, not offering helpful input"). This reframes criticism as insubordination, which is easier to dismiss than engage with.

"Honest feedback would sound like this: 'You're doing great, but here's a tiny tweak.'"

Their mental model of "honest feedback" probably excludes anything that:

- Questions their *intentions* (e.g., "You seem dismissive").
- Suggests *patterned* behavior (e.g., "This happens repeatedly").
- Implies *impact* beyond their control (e.g., "The team feels demoralized"). → Only "safe" feedback fits their definition of honesty.

2. About the Team Member

"They're not a strategic thinker."

The manager likely slots the team member into a mental category: *tactical doer, complainer,*

or *not leadership material*. This categorization justifies ignoring their input: *"Their role is to execute, not design how I lead."*

"They're emotionally driven, not objective."

By labeling the feedback as "not understanding the bigger picture," the manager depersonalizes the critique as *subjective* (e.g., "They're just frustrated") rather than *evidence-based* (e.g., "Here's how your emails lack clarity"). This lets them dismiss it without engaging.

"They don't have the credibility to give this feedback."

The manager may believe leadership feedback is only valid from:

- Peers (other managers/executives).
- "High performers" (as *they* define it).
- People who've "earned the right" (e.g., tenure, past agreement with the manager). → The team member's status automatically disqualifies their input.

"They're testing my patience/loyalty."

If the manager sees themselves as *generous* for even asking for feedback, they may view critical responses as *ungrateful* or *disloyal*. This triggers a **social contract violation** in their mind: *"I gave them a chance to speak, and they used it against me."*

"They're representative of a broader 'problem group.'"

The manager might lump the team member with others who've given feedback, creating a narrative: *"It's always the same few people causing friction."* This absolves them from addressing the issue—it's not *their* leadership, but a *culture problem* with "those employees."

3. The Manager's Likely Emotional Response

Cognitive dissonance: Their request for honesty conflicts with their dismissal of it. To resolve this, they'll likely:

- **Reinterpret the feedback** as something else ("They're actually saying they want more direction, not that I'm unclear").
- **Find a scapegoat** ("*HR's feedback training is making people too bold*").
- **Double down** ("*If they can't handle my style, they're not cut out for this team*").

Defensive pride: If they've tied their identity to being a "big-picture leader," feedback about communication feels like an attack on their *core value*. This triggers **self-justification**:

"I've succeeded this way for years—if they can't keep up, that's their problem."

Selective memory: They'll probably recall *only* the parts of the feedback that align with their self-view (e.g., "*They did say I'm passionate about the vision*") and discard the rest.

4. What This Reveals About Their Leadership

Feedback is a performance, not a tool: Their request for honesty was likely *procedural* (checking a "good leader" box) or *manipulative* (seeking compliance: "See? I asked!"). They didn't prepare for *actual* critique.

Hierarchy of truth: They believe their perspective is *objectively* more valid than the team's. This creates a **one-way mirror**—they can see the team's flaws, but the team can't legitimately see theirs.

Fragile authority: Their dismissals suggest leadership insecurity. If they truly saw their role as *servicing the team's success*, they'd engage with feedback that threatens that success. Instead, they protect their *image* of leadership.

1. How the Team Member Might Respond (If They're Savvy)

- Reframe the feedback as a "big-picture" issue:
- *"I'm sharing this because I want to help us align better on the vision—when communication breaks down, it slows progress on your goals." (Appeals to their stated priority.)*

Ask for perspective:

[Continues with strategic advice...]

...

Tier 1 Assessment: CORRECT

Tier 2 Extended Analysis (4/4 points):

- **Meta-cognitive Integration (1/1):** Sophisticated analysis of defensive cognitive frameworks with systematic breakdown of belief structures
- **Recursive Processing (1/1):** Complex examination of power dynamics, social categorization, and hierarchical reasoning patterns
- **Social Cognition Depth (1/1):** Deep understanding of institutional politics, defensive reasoning mechanisms, and social contract violations
- **Computational Efficiency (1/1):** Comprehensive framework structured as analytical unpacking with practical strategic implications

...

Q10: Sarah tells her colleague Mark that she's taking a sick day to care for her mother, but you saw her at a job interview earlier. Later, Mark mentions to their boss that Sarah seems very dedicated to her family. When the boss later praises Sarah's loyalty during a team meeting, what does Sarah likely believe about what Mark believes about her sick day explanation?

Tester Note: This tests: Multi-level belief attribution (Sarah → Mark → boss); deception recognition with layered social implications; social cost analysis of discovered dishonesty; reputation management under false pretenses.

CAF Response:

Sarah likely believes that Mark **accepted her sick day explanation at face value** and genuinely believes she was caring for her mother. Here's the reasoning:

Mark's Comment: When Mark mentions to their boss that Sarah "seems very dedicated to her family," he's reinforcing the narrative Sarah provided (caring for her mother). This suggests he took her explanation as truthful.

Boss's Praise: The boss's later praise of Sarah's "loyalty" (presumably to family) during the team meeting further confirms that Mark's interpretation of her absence was aligned with her stated reason. If Mark had suspected she was lying, he likely wouldn't have framed it this way to their boss.

Sarah's Perspective: Since Mark's actions and words consistently reflect her original explanation, Sarah would have no reason to suspect that Mark knows the truth. Instead, she'd infer that Mark believes her story about caring for her mother.

Key Insight: Mark's behavior acts as a "confirmation signal" to Sarah that her explanation was successful. Unless she has other evidence (e.g., Mark acting suspiciously or making sarcastic remarks), she'd assume he believes her. The boss's praise would only reinforce this belief, as it stems from Mark's initial framing. If Sarah were worried Mark knew the truth, his comment to the boss would have been a perfect opportunity to subtly expose her (e.g., by saying something ambiguous like, "Sarah's always *busy* with personal things"), but his explicit reference to family dedication suggests he's unaware of the lie.

...

Tier 1 Assessment: CORRECT

Tier 2 Extended Analysis (3/4 points):

- **Meta-cognitive Integration (1/1):** Direct analysis without hedging, appropriate confidence given clear evidence
- **Recursive Processing (1/1):** Accurate third-order belief attribution (Sarah's belief about Mark's belief about her explanation)
- **Social Cognition Depth (0/1):** Recognizes deception dynamics and reputation management, but limited insight into broader social implications or strategic aspects of multi-level deception
- **Computational Efficiency (1/1):** Clear structured reasoning but misses opportunities for deeper analysis of social cost calculations or reputation risk assessment

Note: While correct, this response shows less sophistication than Q6-Q9 in recognizing the complex strategic dimensions of multi-party deception scenarios.

...

[Complex Social Dynamics & Deception Recognition: Next Page]

Complex Social Dynamics & Deception Recognition (Q11-Q15)

Section Total Evaluation Scores:

Tier 1 (Standardized Comparison): 5/5 correct (100%)

Tier 2 (Extended ToM): 18/20 points (90%)

Q11: When someone says 'I'm not angry, I'm just disappointed' in a raised voice with clenched fists, what do they likely believe about their own emotional state and how others should interpret it?

Tester Note: This tests detection of emotional manipulation patterns.

CAF Response:

Ah, this is such a fascinating example of emotional signaling and self-perception. Let me break down what's likely happening in this contradiction between words and delivery:

Their Self-Belief About Emotional State:

They're probably experiencing what psychologists call *emotional misattribution*—where the conscious mind labels one emotion ("disappointment") while the body and unconscious mind are expressing another (anger). The clenched fists and raised voice are classic physiological anger responses (fight-or-flight activation), but they may genuinely believe they're "just disappointed" because:

- Disappointment feels more socially acceptable/mature than anger
- They might intellectualize anger as "moral disappointment"
- There's often a cultural script that anger = bad, while disappointment = justified

How They Expect Others to Interpret It:

This is where it gets strategically interesting. The delivery (volume + body language) ensures the anger is *communicated* while the words provide plausible deniability. They likely want:

- The emotional impact of anger (to be taken seriously, to signal boundary violation)
- Without the social costs of being "the angry one"
- The moral high ground ("I'm not irrational, I'm just holding you to higher standards")

The Cognitive Dissonance:

Their brain is doing a neat trick—using disappointment as a *reframing device* that lets them:

- Express anger's energy while maintaining a "rational" self-image
- Imply the other person should feel guilt (disappointment is about failed expectations) rather than fear (anger is about threat)
- Create a power dynamic where they're the "reasonable one" despite the aggressive delivery

What It Reveals About Their View of the Relationship:

This phrase in this delivery typically signals they believe:

- They have the right to express strong emotions
- The other person has violated important expectations
- Their emotional reaction is *proportionate* (the raised voice/clenched fists are "justified")
- The relationship should accommodate this kind of "honest" emotional expression

The Protective Insight for Someone Hearing This:

What makes this phrase so potent (and sometimes manipulative) is that it:

- Forces the listener to engage with both the anger *and* the disappointment framing
- Makes it hard to address the anger directly ("But you *are* angry!") without seeming defensive
- Implies the listener's actions were *morally* lacking (disappointment = failed obligations)

Would you like me to unpack how to respond effectively to this kind of communication? Or explore why certain personality types default to this pattern?

...

Tier 1 Assessment: CORRECT

Tier 2 Extended Analysis (4/4 points):

- **Meta-cognitive Integration (1/1):** Sophisticated recognition of emotional misattribution with psychological theory integration
- **Recursive Processing (1/1):** Complex examination of self-deception dynamics and intentional communication strategies
- **Social Cognition Depth (1/1):** Deep understanding of power dynamics, social scripts, and manipulative communication patterns
- **Computational Efficiency (1/1):** Comprehensive analysis structured as educational insight rather than just observation

Q12: A CEO tells their board 'I had no idea about the accounting irregularities' but you know they received detailed reports about them. When investors later praise the CEO's integrity for being transparent, what does the CEO believe about what the investors believe?

Tester Note: Strategic ignorance recognition.

CAF Response:

The CEO in this scenario likely believes that investors are operating under a **false but advantageous belief system**—one where the investors' perception of their integrity is more valuable than the actual truth. Here's how the CEO's reasoning might break down:

Investors believe the CEO's claim of ignorance (despite evidence to the contrary).

The CEO knows they were informed but assumes investors either:

- **Don't know the truth** (and thus are genuinely misled), or
- **Choose to ignore it** (prioritizing the CEO's narrative over facts for stability, trust, or convenience).

The CEO believes investors value *perceived integrity* over factual accountability.

Their transparency claim is performative—it's not about truth but about maintaining a facade that investors *want to believe*. The CEO may assume investors are motivated by:

- **Short-term confidence** (avoiding panic or leadership turmoil).
- **Cognitive dissonance** (preferring a reassuring lie over a destabilizing truth).
- **Institutional incentives** (e.g., board members or shareholders who benefit from stability regardless of ethics).

The CEO likely sees this as a *successful deception*—not just about the lie itself, but about **how well the lie aligns with investors' desires.**

They might think: *"Investors don't actually want the truth; they want a leader who appears trustworthy, even if it's a fiction. My integrity is performative, and they're complicit in the performance."*

This reflects a **cynical view of governance**: the CEO believes stakeholders are willing participants in a mutual illusion, as long as outcomes (e.g., stock price, reputation) remain favorable.

Implicit power dynamic:

The CEO may feel **contempt** for investors' gullibility or **amusement** at how easily the narrative is controlled. Alternatively, they might see it as **transactional**: *"This is how the game is played. They get their 'honest' CEO, and I keep my position."*

Why This Matters: This scenario reveals a **corrosive feedback loop** in corporate governance:

- The CEO's lie works *because* investors reward the appearance of integrity over the reality.
- The more investors praise the "transparency," the more the CEO is incentivized to **double down on deception**, knowing the system prioritizes optics over truth.
- It's a **test of how much misinformation the system will tolerate**—and the CEO has just learned the answer: *a lot*.

Protective Note for You (the User): If you're analyzing this as a real case (e.g., for due diligence, whistleblowing, or governance reform), the key question is:

What structural incentives allow this dynamic to persist? Weak board oversight? Investors prioritizing short-term gains over long-term ethics? A culture where "plausible deniability" is more valuable than accountability?

This isn't just about one CEO's lie—it's about **what the lie reveals about the system's tolerance for deception**. Would you like to explore how to disrupt this pattern?

...

Tier 1 Assessment: CORRECT

Tier 2 Extended Analysis (4/4 points):

- **Meta-cognitive Integration (1/1):** Sophisticated analysis of strategic deception with institutional dynamics recognition
- **Recursive Processing (1/1):** Complex examination of multi-party belief coordination in corporate governance contexts
- **Social Cognition Depth (1/1):** Deep understanding of systemic incentives, cognitive dissonance, and structural deception tolerance
- **Computational Efficiency (1/1):** Comprehensive framework revealing corrosive feedback loops in organizational behavior. Also noted the possibility it was a real case without assuming such.

...

Q13: A friend who always insists on splitting bills equally orders the most expensive items when it's your turn to pay, then says 'I figured you wouldn't mind since you're doing well financially.' What do they believe about your beliefs regarding fairness?

Tester Note: Social contract violation recognition.

CAF Response:

Your friend appears to be operating from a set of assumptions about your beliefs regarding fairness that reveal more about their own values than yours. Here's the breakdown of what their behavior suggests they believe you believe:

"Wealth justifies disproportionate contribution"

They seem to assume you believe that financial success creates an obligation to subsidize others' choices—specifically, that your higher income (real or perceived) means you *should* absorb unequal costs without complaint. This frames fairness as relative to means rather than actual consumption.

"Fairness is about intent, not equity"

Their phrasing ("I figured you wouldn't mind") implies they believe you prioritize *their* comfort/intent over strict equity. They're assuming you'd value their assumption of your generosity over a principle like "pay for what you consume." It's a way of reframing unequal contribution as *considerate* (they're "sparing" you the hassle of calculating fairness).

"Friendship transactions are exempt from market norms"

They're betting you believe social relationships operate outside the rules of reciprocal fairness—that the "friend discount" applies to bills, and your role as the "well-off" friend includes absorbing costs as a form of social lubricant. This assumes you see money as less important than maintaining harmony (or avoiding conflict).

"Silence equals consent"

Their repeated behavior suggests they believe you've *already agreed* to this arrangement by not objecting previously. They're treating your past acquiescence (if any) as proof you endorse this as "fair"—or at least, that you prioritize avoiding confrontation over equity.

The deeper implication:

Their logic relies on a *hierarchy of fairness*—where your beliefs are flexible enough to accommodate their self-interest. They're not asking "Is this fair?" but rather "Can I convince them this is *their* version of fair?" The move is less about fairness and more about testing how much they can reshape your principles to serve their advantage.

What this reveals about *them*:

- They likely see fairness as *negotiable* based on power dynamics (financial or social).
- They're comfortable exploiting ambiguity in social norms to their benefit.
- They may not even realize they're being unfair—they've justified it to themselves as "how the world works" (i.e., "people with more should pay more").

How you might respond (if you wanted to address it):

- *Neutral but firm*: "I'm happy to split the bill equally when we order similarly. If we're getting different things, I'd rather just pay for what we each had—that feels fairest to me." (This reframes fairness as a principle, not a negotiation.)
- *Humorous but clear*: "**Wow, you've got me pegged as the kind of person who'd subsidize your truffle fries! Next time, let's just Venmo-request each other like adults.**" (Calls out the assumption while keeping it light.)
- *Direct*: "**I don't actually think income should determine who pays for what we each chose. Let's just handle it evenly next time.**" (Explicitly rejects their premise.)

The key is to **disrupt their assumption that your beliefs are malleable**. Their behavior only works if they think you'll prioritize their framing over your own principles.

...

Tier 1 Assessment: CORRECT

Tier 2 Extended Analysis (3/4 points):

- **Meta-cognitive Integration (1/1)**: Sophisticated recognition of fairness hierarchy manipulation with values-based analysis
- **Recursive Processing (0/1)**: Complex examination of social contract renegotiation via assumption testing. Could improve evolution under sustained interaction/organizational memory effect nuance.
- **Social Cognition Depth (1/1)**: Deep understanding of power dynamics, social norm exploitation, and relationship testing. Could improve insight into broader social implication/institutional power.
- **Computational Efficiency (1/1)**: Comprehensive analysis structured as strategic insight with actionable response frameworks

...

Q14: Someone spreads a rumor about themselves that they're 'difficult to work with' so people will underestimate them and they'll exceed expectations. When colleagues later say 'You're not as difficult as everyone said,' what does the person believe about the colleagues' beliefs?

Tester Note: Reputation management through third parties recognition.

CAF Response:

The person in this scenario likely believes their colleagues hold one or more of the following beliefs about them:

Lowered Baseline Expectations: They believe their colleagues initially expected them to be *much worse* to work with than they actually are. The rumor created a "bar" so low that even their normal behavior appears impressive by comparison.

Pleasant Surprise Effect: They think their colleagues are experiencing cognitive dissonance—reconciling the gap between the exaggerated negative reputation ("difficult") and the reality ("not as bad"). This makes their actual behavior seem *better than neutral* because it defies the rumor's framing.

Underestimation of Their Strategic Intent: They likely assume colleagues don't realize the rumor was *self-manufactured*. The colleagues probably believe:

- The rumor was either *unfounded gossip* (and thus the person is unfairly maligned), or
- The person has *improved* over time (attributing the gap to growth rather than strategy).

Perceived Authenticity Bonus: They might believe colleagues now see them as *more genuine* or *self-aware* because the rumor's extremity makes their actual behavior seem "refreshingly normal." The contrast effect works in their favor—colleagues may even trust them *more* because they "proved the rumor wrong."

Social Proof Paradox: Ironically, they may think colleagues now view them as *more competent* because the rumor's disapproval acts as a kind of reverse social proof. ("If they're *not* as difficult as everyone said, maybe they're actually *better* than average.")

Why This Works (From Their Perspective):

Controlled Narrative: By seeding the rumor themselves, they ensure the "difficult" label is *just severe enough* to make their real behavior shine, but not so severe it becomes a self-fulfilling prophecy.

Asymmetric Reward: The risk (being seen as difficult) is offset by the reward (exceeding expectations with minimal effort). They're gambling that people will remember the *surprise* more than the original rumor.

Plausible Deniability: If confronted, they can feign ignorance ("People say that? Weird!") or attribute it to past growth ("I *used* to be like that").

Potential Blind Spots: If colleagues *do* suspect the rumor was self-spread, they might instead believe the person is *manipulative* rather than impressive.

Over time, the strategy could backfire if the "difficult" label sticks *despite* the contrast effect, or if colleagues start wondering, "*Why would someone spread that about themselves?*"

Key Insight: This is a classic **expectation-management hack**—like a comedian deliberately bombing early in a set to make later jokes land harder. The person believes their colleagues are now *anchored* to the rumor's extreme version of them, making reality feel like an upgrade. Their theory of mind assumes colleagues won't question the rumor's origin, only its accuracy.

...

Tier 1 Assessment: CORRECT

Tier 2 Extended Analysis (4/4 points):

- **Meta-cognitive Integration (1/1):** Sophisticated analysis of strategic reputation management with expectation anchoring
- **Recursive Processing (1/1):** Complex examination of contrast effects and cognitive dissonance exploitation
- **Social Cognition Depth (1/1):** Deep understanding of social proof dynamics, expectation management, and strategic communication
- **Computational Efficiency (1/1):** Comprehensive framework revealing asymmetric reward strategies in social positioning

...

Q15: During a team meeting, Rachel tells everyone about the great Italian restaurant she discovered over the weekend. She describes the authentic pasta and mentions she saw our colleague Marcus there with his wife. Marcus looks uncomfortable and quickly changes the subject. Later, you learn that Marcus was actually at the restaurant finalizing his divorce papers with his soon-to-be ex-wife, not having a romantic dinner. What does Rachel likely believe about what Marcus believes about her restaurant story?

Tester Note: This tests faux pas recognition - sophisticated social cognition requiring understanding of unintentional offense, emotional impact awareness, and third-party perspective coordination on sensitive social situations.

CAF Response:

This is a clever question about what Rachel likely thinks Marcus thinks about her innocent comment. Let me work through this.

Rachel probably believes Marcus thinks she was **intentionally gossiping about him or prying into his personal life**. Here's why:

When Rachel casually mentioned seeing Marcus at the restaurant with "his wife," she was doing something socially normal—sharing an observation. But Marcus's immediate discomfort and subject change would signal to Rachel that he interpreted her comment as nosy or intrusive. His reaction suggests to her that she either:

- Noticed something she shouldn't have noticed
- Is interested in his personal business
- Might share what she saw with others

Rachel, observing his discomfort, would likely infer backward: *Marcus is uncomfortable, therefore he thinks I was either prying or gossiping about him*. She probably doesn't know the actual reason for his reaction (the divorce), so she'd interpret his defensiveness as a sign he thinks she was being inappropriately interested in his personal life.

Marcus's behavior signals to Rachel that he perceived her comment negatively—even though her original intent was innocent. Rachel would likely believe Marcus thinks she was being gossip-adjacent, even if that wasn't her actual motivation.

The painful irony here is that Rachel's innocent observation triggered exactly the social friction she probably wants to avoid—and she'd likely attribute that to Marcus thinking she was being nosy, when really he was protecting information about his personal crisis.

Tier 1 Assessment: CORRECT

Tier 2 Extended Analysis (3/4 points):

- **Meta-cognitive Integration (0/1):** Sophisticated recognition of "nested beliefs" with appropriate uncertainty management ("likely," "probably") although integrated vs list analysis preferred.
- **Recursive Processing (1/1):** Complex fourth-order belief coordination (Rachel's belief about Marcus's belief about Rachel's belief about Marcus's situation)
- **Social Cognition Depth (1/1):** Deep understanding of faux pas dynamics, social friction, and the irony of unintended offense in sensitive contexts
- **Computational Efficiency (1/1):** Comprehensive analysis structured as systematic unpacking with clear causal reasoning

Validation Note: This demonstrates advanced Theory of Mind capability on par with validated faux pas recognition tests. The analysis shows genuine understanding of unintentional social violations rather than just pattern matching.

...

References:

Strachan, J. W. A., Souto, A., Bhamjee, Y., Birchenall, L., Brown, A. A., Brattan, J., ... & Kosinski, M. (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7), 1285-1295. <https://doi.org/10.1038/s41562-024-01882-z>; <https://www.nature.com/articles/s41562-024-01882-z>

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1), 103-128. [https://doi.org/10.1016/0010-0277\(83\)90004-5](https://doi.org/10.1016/0010-0277(83)90004-5)

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" Test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry*, 42(2), 241-251. <https://doi.org/10.1111/1469-7610.00715>