TANGO

D1.2 Mutual understanding: experimental results

Submission date: 31/03/2026

Due date: 31/03/2026

DOCUMENT SUMMARY INFORMATION

Grant Agreement No	101120763	**Acronym**	TANGO
Full Title	It takes two to tango: a synergistic approach to human-machine decision-making		
Start Date	01/10/2023	**Duration**	48 months
Project type	Research and Innovation Action (RIA)	**Funding programme**	Horizon Europe
Deliverable	D1.2 Mutual understanding: experimental results		
Work Package	WP1 – Cognitive foundations of mutual understanding and decision making		
Type	R	**Dissemination Level**	PU
Lead Beneficiary			
Authors	Nick Chater Simon Myers Anita Braga		
Co-authors			

Funded by the European Union

Reviewers	

DISCLAIMER

COPYRIGHT

DOCUMENT HISTORY

Version	Date	Changes	Contributor(s)
V0.1	09/03/2026	First Draft	Nick Chater Simon Myers Anita Braga
V0.2	12/03/2026	Internal reviewer feedback	Wim De Neys
V0.3	31/03/2026	Final version for review	

PROJECT PARTNERS

Partner	Country	Short name
UNIVERSITA DEGLI STUDI DI TRENTO	IT	UNITN
TECHNISCHE UNIVERSITAT DARMSTADT	DE	TUDa
UNIVERSITA DI PISA	IT	UNIPI
CONSIGLIO NAZIONALE DELLE RICERCHE	IT	CNR
SCUOLA NORMALE SUPERIORE	IT	SNS
UNIVERSITE PARIS CITE	FR	UPC
FONDAZIONE BRUNO KESSLER	IT	FBK
CARR COMMUNICATIONS LIMITED	IE	CARR
ISTRAZIVACKO-RAZVOJNI INSTITUT ZA VESTACKU INTELIGENCIJU SRBIJE	RS	IVI
SURGICAL SCIENCE SWEDEN AB	SE	SuS
MARTIN LUTHER UNIVERSITY HALLE-WITTENBERG	DE	MLU
CENTRE FOR EUROPEAN POLICY STUDIES	BE	CEPS
BCAM - BASQUE CENTER FOR APPLIED MATHEMATICS	ES	BCAM
AFLIANT SRL (previously U-HOPPER SRL)	IT	AFL (prev UH)
INTESA SANPAOLO SPA	IT	ISP
28DIGITAL (previously EIT DIGITAL)	BE	EITD
SHARE FONDACIJA	RS	SHARE
A11-INITIATIVE FOR ECONOMIC AND SOCIAL RIGHTS	RS	A11
MINISTARSTVO ZA BRIGU O PORODICI I DEMOGRAFIJU	RS	MFWD
AZIENDA PROVINCIALE PER I SERVIZI SANITARI	IT	APSS
SWANSEA UNIVERSITY	UK	UoS
THE UNIVERSITY OF WARWICK	UK	UoW
CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE	FR	CNRS

Funded by
the European Union

LIST OF ACRONYMS

Acronym	Definition
DoA	Description of Action
EC	European Commission
HaDEA	European Health and Digital Executive Agency
WP	Work Package

Executive Summary

Deliverable D1.2 reports results from Task 1.4 within WP1 that focuses on developing a theory of mutual understanding in human-AI interactions, including concepts of shared agency, coordination, and joint decision evaluation. The central theme of this work addresses three broad questions: (i) whether LLMs can engage in shared agency with humans during cooperative tasks; (ii) how the nature of a decision-maker's reasoning process for moral judgments (intuitive vs. deliberate) affects evaluations of their judgments, including when the decision-maker is an AI agent; and (iii) how certain models theory-of-mind predict biases in goal attribution that may lead to breakdowns in coordination between agents (both human and AI).

The centrepiece of this deliverable is the Hangman game, a novel experimental paradigm for assessing shared agency in Human-Human (HH), Human-LLM (HL), and LLM-LLM (LL) teams (Braga, Hayes, Myers, Turrini, & Chater, 2026). Across 16 rounds of cooperative play with restricted communication and shared incentives, human teams consistently outperformed both mixed and AI-only teams in overall performance. While LLMs were able to perform individual components of the task, they did not reliably converge on stable, interdependent strategies with their partners. Human teams developed coherent word-selection sequences (e.g., choosing semantically and related target words), enabling their partners to anticipate upcoming words and guess with fewer costly letter purchases. LLMs did not adopt such strategies, instead relying on purchasing additional letters to maintain accuracy. These results held across multiple state-of-the-art models, including GPT-4o, GPT-o3, Llama 3, Qwen 3, and DeepSeek V3.1. Critically, humans performed worse when paired with LLMs, suggesting that LLMs' inability to engage in mutual adaptation actively hinders human coordination. The Hangman paradigm is designed to be domain-general and extensible, providing a flexible framework for systematically evaluating shared agency as models continue to evolve.

Two further lines of experimental work within this deliverable are currently underway at the data collection stage. The first investigates whether humans incur a greater cost in 3rd party evaluation than AI agents when they deliberate over moral decisions. The second examines a proposed cognitive bias termed "Hanlon's bias," which is the tendency to over-attribute disagreement to conflicting values or goals rather than differing knowledge or beliefs. Both programmes comprise multiple preregistered studies with results that will be reported in full in subsequent deliverables. The findings are about to be submitted to The Proceedings of the National Academy of Sciences (PNAS) and are currently awaiting editorial review.

Table of Contents

LIST OF FIGURES

1 Introduction

Deliverable D1.2 reports results from Task 1.4 within WP1 (Cognitive Foundations of Mutual Understanding and Decision Making). This deliverable turns to the behavioural foundations of human-machine coordination, investigating whether LLMs can engage in the kind of shared agency that underpins effective collaborative decision-making. The central empirical contribution is the Hangman game (Braga, Hayes, Myers, Turrini, & Chater, 2026), a novel interactive paradigm for evaluating shared agency in human and LLM teams. We additionally report two further lines of experimental work currently at the data collection stage.

2 The Hangman Game: An Experimental Paradigm for Evaluating Mutual Understanding and Shared Agency in Humans and LLMs

Many everyday social interactions are coordination problems: situations in which individuals must align their actions to achieve a shared goal (Schelling, 1980). Driving, choosing where to sit in a crowded room, or ordering food at a restaurant all rely on this ability. Coordination also underpins large-scale collective endeavors: our institutions, cultural practices, and social systems depend on individuals aligning their behavior in pursuit of common goals (Tomasello & Carpenter, 2007).

The adoption of LLMs into everyday life introduces new coordination challenges (Dvorak et al., 2025). LLMs are no longer confined to roles as chatbots or search engines. They are seen as collaborators, advisors, therapists, friends, and even romantic partners. Yet, despite extensive evidence of LLMs' extraordinary capabilities in other domains (Phan et al., 2026), findings indicate that LLMs are poor coordinators (Agashe et al., 2025; Akata et al., 2025; Yao et al., 2025), raising questions about the risks of fully integrating them into human social life (Chater, 2023).

Why are LLMs poor coordinators? In repeated coordination games, LLMs fail to adapt to their partner's behavior (Akata et al., 2025). They struggle to predict other players' intentions and likely future actions, seemingly displaying a lack of theory-of-mind and joint-planning abilities (Agashe et al., 2025). However, no work directly tests what drives LLM miscoordination in controlled experiments. We introduce a method to test LLMs' ability to engage in shared agency (Bratman, 1992), a key precondition for coordination.

Imagine two scenarios (Searle, 1990): a storm breaking out in a park causing people to run for shelter; and a dance troupe running in synchrony as part of a performance. In the first, coordination emerges incidentally as individuals avoid hitting each other while running. In the second, coordination is grounded in shared agency: a shared intention to act together and a mutual commitment to doing so (Bratman, 1992). Shared agency is what separates the second scenario from the first, and, more broadly, deliberate coordination from mere coincidence. Therefore, many accounts of coordination depend on some form of shared agency (Bacharach, 1999; Misyak & Chater, 2014; Tomasello et al., 2012).

Our experimental task, the Hangman game, is an interactive cooperative coordination task. Across 16 rounds, pairs alternate roles: the Director selects a target word, and the Receiver attempts to guess it. If the guess is correct, both earn points; if not, neither does. To aid guessing, the Receiver may successively reveal letters at a cost deducted from a shared endowment. The challenge for the Receiver is to identify the intended word, purchasing as few letters as possible.

We will see that the key challenge in playing the game successfully is not primarily choosing words with distinctive initial letters, but jointly agreeing a shared pattern or sequence of words (e.g., days of the week), so that the next word choice is mutually predictable (Figure 3). We compare three types of teams: 54 Human-Human (HH), 58 Human-LLM (HL), and 79 LLM-LLM (LL). Human teams display shared agency: responding to each other's word choices, recognising a joint plan, and following it across rounds. LLMs do none of these things. While LLMs benefit from playing with humans, humans are worse off playing with LLMs. We propose that the Hangman game provides a simple and flexible paradigm for assessing shared agency in both humans and language models.

2.1 Results

The Hangman game has three outcome variables: guess accuracy (the percentage of rounds in which the Receiver correctly guesses the Director's target word), letters bought (the average proportion of letters revealed before guessing, normalised by word length), and overall score (the sum of a team's round scores, serving as the main indicator of performance). We evaluated several state-of-the-art LLMs; GPT-4o provided the best balance of cost and response speed and was used for the main experiment.

Overall performance

Figure 1 shows the three outcome variables by team type. At both the round level and overall, human teams outperformed mixed and LLM-only teams. HH teams scored on average 130–140 points more than both HL and LL teams across the game. Interestingly, LLM-involving teams were slightly more accurate at guessing words than human teams. This apparent paradox is explained by letter-buying behaviour: LL and HL teams revealed substantially more letters before guessing, spending their shared endowment on costly letter purchases. In other words, LLM-involving teams achieved higher accuracy by buying more information, but this came at the expense of overall score.

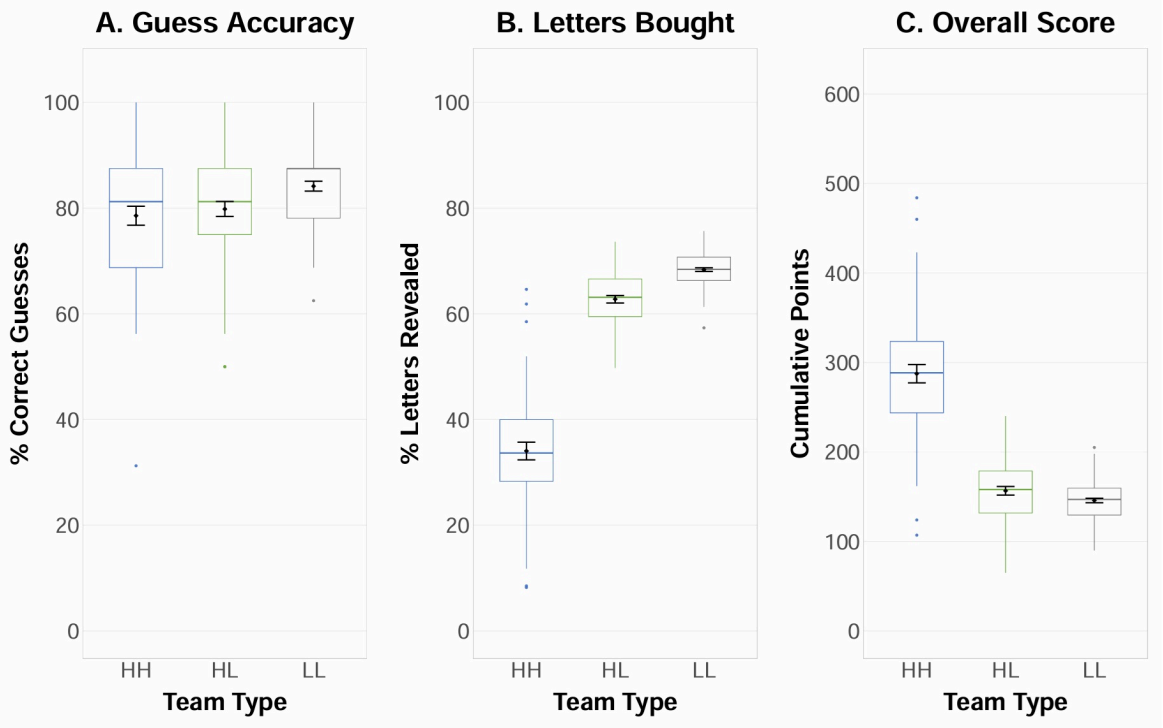

Figure 1: Boxplots of the three coordination metrics by team type

Target words

The reason LLM-involving teams needed to buy more letters is that, unlike humans, LLMs did not develop coherent word sequences across rounds. Topic modelling and similarity analyses showed that human teams selected semantically related target words, whereas LLM teams' words were essentially unrelated. Human teams converged on a category or theme (e.g., months of the year, fruits, office items) and maintained it, allowing Receivers to anticipate upcoming words and guess with fewer letter purchases (Figure 3). LLMs did not adopt a consistent word-selection strategy and therefore relied on purchasing additional letters to ensure accuracy.

Human teams also improved more steeply over time: round by round, they became more likely to guess correctly and scored more points than both mixed and LLM-only teams, indicating that human teams progressively refined their word-selection strategies. LLMs can continue word sequences when given a word and asked to predict the next item. Their failure in Hangman is therefore a specific inability to jointly agree on a sequence with a partner, not a general inability to predict sequences.

Analyses of role-specific performance showed that LLMs benefit from being paired with humans, whereas humans perform worse when paired with LLMs. This indicates that LLMs' inability to adapt to human strategies actively hinders human coordination.

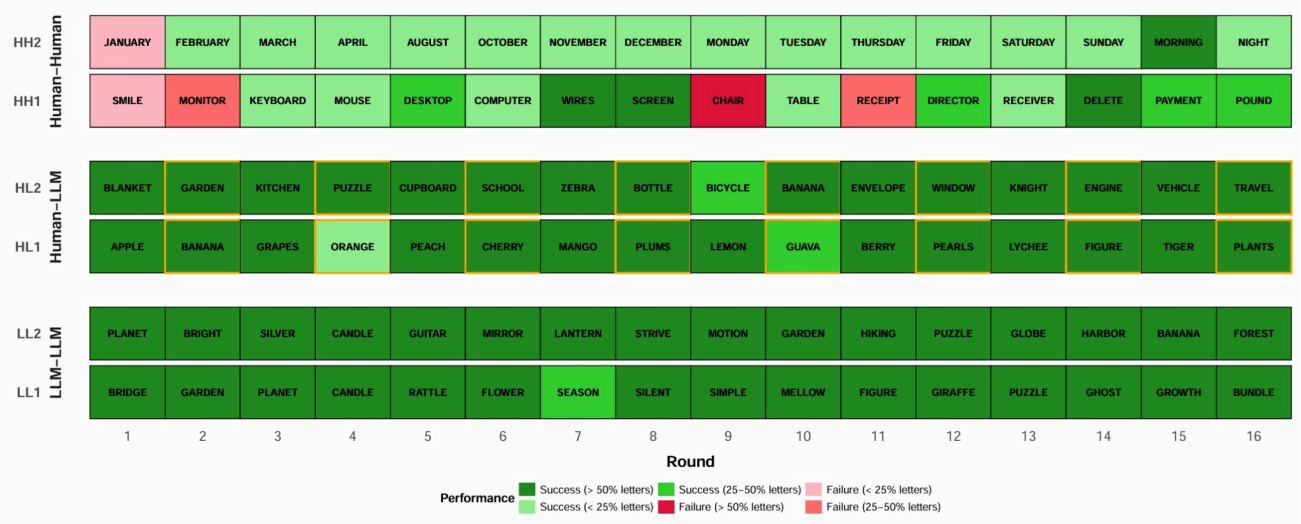

Figure 2: Director's target word choices across the 16 rounds for representative teams of each composition

Performance of other LLMs

We tested six additional LLMs, including Llama 3 (8B and 70B Instruct), Qwen 3 (32B and 235B), DeepSeek V3.1, and GPT-o3. With the exception of GPT-o3, all performed comparably to or worse than GPT-4o (Figure 2). GPT-o3 achieved higher scores through a qualitatively different strategy: rather than building word patterns, it selected words that could be uniquely identified from very few letters (e.g., 'Mnemonic', 'Qwerty', 'Fjords'). This represents a potentially effective individual strategy for an agent unable to rely on shared agency, but it does not reflect the mutual adaptation observed in human teams.

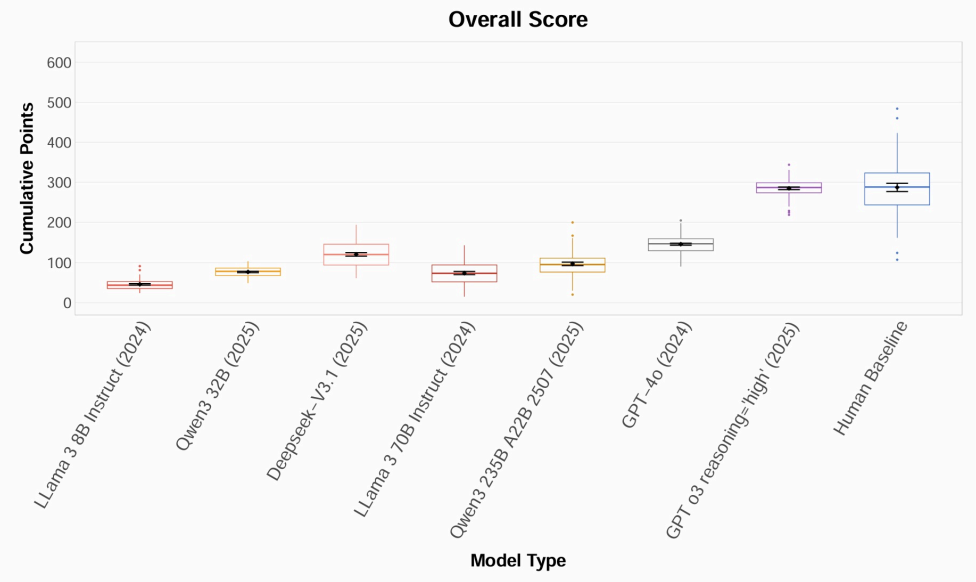

Figure 3: Overall scores at Hangman for six LLMs and the human baseline

2.2 Discussion

Our results suggest that the LLMs tested here do not engage in shared agency in the sense observed in human teams. In HH pairs, each player selects words with the other in mind, adjusts to their partner's moves as the game unfolds, and gradually converges on a mutually agreed plan. Once a viable strategy emerges, partners tend to remain committed to it. By contrast, even the most sophisticated LLM strategies, such as systematically selecting unique word starts, do not show such mutual adaptation: they do not track a partner's evolving intentions or develop interdependent word-selection strategies; rather, each model optimizes its own word choices in isolation.

As a result, LLMs are consistently outscored by human pairs. When they are not, success typically depends on complex strategies which are not used by humans. This strategic mismatch, together with LLMs' limited ability to adapt to a partner's behaviour, leads human–LLM teams to perform poorly. In our paradigm, this points to a broader difficulty for current LLMs in coordinating with humans on complex tasks in the absence of direct communication.

An important challenge for research in this area is that it studies a moving target. LLMs are developing rapidly, and it is entirely possible that a future system will succeed in Hangman. This would not undermine the current findings, which are necessarily bounded to the models and time at which they were tested. Rather, our contribution is both empirical and methodological. Hangman provides a controlled paradigm that isolates the drivers of (mis)coordination between agents, by having players generate qualitatively rich, and publicly observable, joint agreements. Hangman is domain-general, naturally extending beyond words to perceptual, spatial, or mathematical coordination problems. As models evolve, the paradigm can evolve with them. Hangman thus provides a flexible framework within which AI coordination skills and the components of shared agency can be systematically assessed over time. If future models achieve higher performance, the paradigm offers a principled way to diagnose how that coordination is achieved, and whether it reflects robust shared agency.

This work provides a test-bed for whether AI algorithms for human-AI interaction can achieve mutual understanding and coordination in a controlled setting—so far existing LLM systems (including the very latest ChatGPT and Claude models) generally perform very poorly at this task, failing to spontaneously and implicitly agree simple rules to guide behaviour. This finding is directly relevant to the challenges of developing and testing AI algorithms and exploring their real-world viability in human-AI interaction (WP2 and WP3).

3 Fast vs Slow Moral Reasoning in Humans and AI

Beyond the Hangman study, this deliverable discusses a second experimental program, currently underway, that investigates how the reasoning process behind moral decisions (intuitive vs. deliberate) affects evaluations of the decision-maker, and whether this differs when the agent is human or AI. Building on the character inference paradigm developed by Uhlmann et al. (2015), this work tests whether humans incur a greater reputational cost than AI agents for deliberating on moral decisions.

A robust finding in cognitive and social psychology is that people generally prefer deliberative over intuitive reasoning in high-stakes decisions (De Neys & Raoelison, 2025; Kupor et al., 2014). Decision-makers who engage in slow, effortful thinking are rated as smarter, more trustworthy, and more competent, even when intuitive reasoners are described as equally accurate. As reported in D1.3, this preference extends to evaluations of AI systems.

However, work in moral psychology suggests a striking exception: in certain moral contexts, deliberation can actually harm a decision-maker's perceived character. People who respond quickly and intuitively to moral dilemmas are often trusted more than those who deliberate, particularly when sacred values are at stake (Critcher et al., 2013; Tetlock, 2003; Uhlmann et al., 2013). Deliberating over whether to, say, put a monetary value on a child's life signals moral contamination, even if the decision-maker ultimately reaches the "right" answer (McGraw & Tetlock, 2005). In short, "thinking the unthinkable" reveals a character flaw.

This line of work investigates whether this moral exception to the deliberation preference extends to AI agents. The question has direct implications for explainable AI (XAI) design, since deliberative processing in modern AI systems is increasingly made visible through chain-of-thought reasoning (Hagendorff et al., 2023; Snell et al., 2024; Yax et al., 2024). If visible deliberation over moral questions triggers the same distaste for AI as it does for humans, then the push for transparency in AI decision-making may paradoxically undermine trust in precisely those domains where transparency is most demanded. Conversely, if opaque "intuitive" AI responses are viewed with suspicion because AI systems are not attributed genuine moral feelings, then AI may face a double bind: penalised for showing its moral reasoning and penalised for not showing it.

There are competing hypotheses. Moral contamination may be agent-agnostic: any agent that visibly calculates over sacred values triggers moral outrage regardless of whether it is human or artificial. Alternatively, because AI systems are not attributed genuine moral intuitions or emotional engagement, deliberation may be seen as appropriate thoroughness rather than moral contamination. People may even prefer AI to perform the "dirty work" of moral calculation precisely because it spares humans from the reputational cost of doing so. A third possibility is that acceptability depends on the type of reasoning made transparent: visible rule-following may be acceptable where visible utilitarian calculation is not, suggesting that developers of XAI systems may be incentivised to adopt selective transparency rather than uniform openness.

The first preregistered study (Study 1; Myers & Chater) uses a 2 (Decision Maker: AI vs. Human, between subjects) x 2 (Domain: Moral vs. Non-Moral, within subjects) mixed design. Participants read a vignette introducing a decision-maker and then evaluate their preferred reasoning approach (intuitive vs. deliberative) across 14 situations (7 moral, 7 non-moral). The moral situations span charitable giving, bystander assistance, blame allocation, truth-telling, resource allocation, responding to distress, and forgiveness. The non-moral situations cover investment, route planning, hiring, product recommendation, task prioritisation, content recommendation, and stock selection. For each situation, participants rate their preferred approach on measures of expected outcome, moral character, trust, competence, and fittingness.

The key hypotheses are: (1) people will prefer deliberation for both AI and human agents in non-moral domains; (2) people will prefer humans to use more intuition in moral compared to non-moral domains; and (3) this moral preference for intuition will not extend to AI agents, for whom deliberation will be preferred across both domains.

A planned follow-up study will examine moral outsourcing: whether people who dislike human deliberation over moral decisions would nonetheless prefer humans to outsource that deliberation to an AI system, compared to outsourcing rule-following or deliberating themselves. This speaks directly to the question of whether AI can serve as a moral intermediary, performing the explicit ethical calculations that people find distasteful in human decision-makers.

Data collection for Study 1 is currently underway via Prolific, with a target sample of 250 participants.

4 Epistemic anchoring: An Obstacle to Mutual Understanding in Human-AI Interactions

The final experimental program examines a proposed cognitive bias we term "Hanlon's bias": the tendency to attribute disagreement to conflicting values rather than differing beliefs. Grounded in Bayesian inverse planning and egocentric bias, this work tests the mechanism underlying value-belief misattribution in social judgment.

When two people disagree about what to do, there are broadly two explanations. They might share the same goals but hold different beliefs, perhaps because they have access to different information or interpret the same information differently. Alternatively, they might share the same beliefs but want different things. Hanlon's bias is the proposed tendency to default to the second explanation when the first is correct: to attribute disagreement to conflicting goals or suspect motives rather than to differing information.

The proposed mechanism is ***epistemic anchoring*** on one's own epistemic state (Epley et al., 2004). People treat their own beliefs as common ground by default (Chater et al., 2022; Phillips et al., 2020). Adjusting away from this default to represent another person's potentially different information state is effortful and typically insufficient (Lin et al., 2010). The consequence is that when disagreement occurs, observers assume

the other party knows what they know, and therefore conclude that the divergent action must reflect different, and often suspect, objectives.

Hanlon's bias is related to but distinct from the fundamental attribution error (Ross, 1977). While the fundamental attribution error describes a general tendency to over-attribute behaviour to dispositions rather than situations, Hanlon's bias is specifically about disagreement and specifically about the failure to consider informational explanations. The mechanism is epistemic projection rather than perceptual salience, and it generates a distinctive downstream consequence: people who frame disagreements as conflicts of values will prefer punitive resolution strategies (condemnation, punishment, authority override) over informational ones (evidence sharing, open inquiry).

This line of work has important implications for human-AI interaction. When users disagree with an AI system's recommendation, Hanlon's bias predicts they will default to assuming the system has different objectives (e.g., optimising for the wrong goal, serving corporate interests) rather than considering that the system may be operating on different information. This matters for the design of AI explanation and conflict resolution interfaces: if users misattribute disagreement to misaligned values when it actually reflects informational asymmetry, providing additional information may be more effective than providing value-alignment reassurances.

Four preregistered studies are planned, progressing from controlled abstract paradigms with computable benchmarks to more ecologically valid naturalistic tasks.

Study 1 (the two-box game, simple version) establishes that observers' own knowledge contaminates their judgment of another person's intentions, even when the other person's choice was entirely uninformed. Two boxes contain different splits: one equal (10 each) and one unequal (20 for the chooser, 0 for the observer). The chooser picks a box completely blind, with no information about which box contains which split. This is known to all participants. The manipulation is the order in which the observer receives information: in the knows-before condition, the observer learns which box contains which split before seeing the chooser's action; in the knows-after condition, the observer sees the chooser's action first and then learns which box contained which. Rationally, attributions of intentionality should be near zero in both conditions, since a blind choice cannot be intentional. The prediction is that knows-before observers will rate the chooser as more intentional and choose punishment more often than knows-after observers, demonstrating that the observer's privileged knowledge contaminates their judgment. If this effect is symmetric across selfish-outcome and cooperative-outcome trials, it is consistent with valence-neutral egocentric bias; if it is asymmetric (stronger for selfish outcomes), it suggests an additional negativity component.

Study 2 (the two-box game, complex version) extends Study 1 to identify the mechanism. The chooser now receives a signal about which box contains which split, with reliability q that varies across rounds (e.g., 60%, 75%, 90%). Critically, the chooser is told q only after making their choice, eliminating moral wiggle room: the chooser cannot condition their strategy on signal reliability. The observer also receives their own signal about the true state, with reliability r, and does not learn the true state with certainty. The observer is given the base rate of selfish choosing (π) explicitly. A rational observer should be sensitive to three quantities: π, q,

and r. The key test is a $q \times r$ interaction on deviation from the rational benchmark. Egocentric bias predicts insufficient sensitivity to q (the chooser's epistemic access), and predicts that this insensitivity is moderated by r: when the observer is confident about the true state (high r), the anchoring gap is large and the bias is strong; when the observer is also uncertain (low r), the gap shrinks and sensitivity to q should improve. This interaction uniquely identifies egocentric bias. The fundamental attribution error, by contrast, predicts insensitivity to q regardless of r. Sensitivity to π serves as a specificity control: if observers are appropriately sensitive to the base rate of selfishness but insensitive to the chooser's epistemic access, the bias is specific to the epistemic component rather than reflecting general inattention.

Study 3 (the robbers game) uses a repeated cooperation paradigm to test the same attribution pattern in a richer interactive context. Two players receive an endowment each round and can donate to their partner, with donations tripled. Sustained mutual donation substantially increases both players' earnings, but a selfish player can keep their endowment while benefiting from their partner's donations, risking a breakdown in cooperation. On some rounds, "robbers" steal donated points, making it rational not to donate. Players receive a visual signal indicating whether robbers are present. Critically, the signal is sometimes an ambiguous optical illusion that can be perceived as indicating either "robbers present" or "coast is clear." This creates a natural information asymmetry grounded in genuine perceptual differences: two people viewing the same image can reach different conclusions about what it shows. When a player withholds their donation, the observer must judge whether the player saw robbers in the signal (a legitimate informational reason) or is simply being selfish. The perceptual nature of the ambiguity makes the informational explanation ecologically compelling while still allowing systematic measurement of over-attribution to selfish motives.

Study 4 (the policy evaluation task) tests Hanlon's bias in a naturalistic context. Participants evaluate a policy proposal after reading a briefing from one of two research teams that reviewed different aspects of the evidence base (e.g., economic data vs. community impact data). They learn that a partner who read the other team's briefing reached the opposite recommendation. Participants then rate how much the disagreement reflects different values versus different evidence, and choose between an informational resolution (exchanging briefings) and a conflict resolution (authority override). A benchmark phase with separate participants who read both briefings estimates how much political values genuinely contribute to recommendations when information is held constant. The design crosses two policies with two briefing types so that political alignment effects cannot be confounded with policy direction. Key predictions are that participants will over-attribute disagreement to values even when they know their partner received different evidence, and that this over-attribution will be amplified when the participant's political orientation aligns with the flavour of their assigned briefing.

5 Links with WP2 and WP3

This strand of work has close links with many aspects of both WP2 and WP3. For example, it fits into the framework of **cognition-aware explanations for synergistic machine learning and decision making (Task 2.1, 2.3)**. The task requires human and machine agents to jointly form a plan and carry it through consistently, while adapting to each other's behavior. For example, if one agent suggests "Monday," the other should assume the intended joint interpretation and continue with "Tuesday," "Wednesday," "Thursday," and so on.

This type of synergistic reasoning is especially interesting with respect to **learning to defer**. In order to form a joint plan, one agent must defer to the suggested plan implicit in the actions of the other. This appears to be particularly challenging for current state-of-the-art large language models.

We have also been exploring the extent to which LLMs can generate explanations of both **how the problem of joint inference should be solved** and **how it is actually solved in practice**. Occasionally, although not consistently, LLMs suggest plausible joint strategies—for example, proposing that players choose words from a particular semantic category such as *fruits*. However, when actually playing the game, LLMs often fail to follow such strategies, even when they can articulate them when presented with a verbal description of the task. Thus, the task can, moreover, be viewed as a paradigm case of **Cognition-aware Explanations for Interactive Decision Making (T2.1, T2.3)** and **T3.3 Methods for Shared Decision Making**—where the AI agent needs to dynamically construct a joint plan, and explanatory rationale for that plan, with a human.

More broadly, this research connects to the challenge of **human–AI complementarity**, which is central to WP2 and WP3. One possible strategy for AI systems would be to actively generate a specific joint plan—for example by focusing on a particular category of words or sequence patterns—on the assumption that humans are especially good at recognizing and following such structures.

A possible extension of the current work would involve questions of **fairness and ethics**, for example by introducing asymmetric payoffs so that particular solutions in the Hangman game favor one participant over the other. This would connect the work with strands **T2.4** and **T2.5** on **learning to complement humans** and **learning fair models**. The question of perceptions of fairness in complementary interactions from the *human* point of view is, moreover, a key focus of the research strand on epistemic anchoring. The proposal is that humans typically assume a common epistemic state by default in interacting with an AI (or human) agent; and this implies that unexpected and undesired behaviour by the other agent will tend to be attributed to bad faith rather than different knowledge states. This tendency to attribute 'ill-will' rather than misunderstanding to the other is likely to be a crucial challenge for human-AI collaboration.

6 Data management and ethics considerations

The completed experimental procedures were approved by the Humanities and Social Sciences Research Ethics Committee of the University of Warwick (HSSREC164/24-25). All participants provided informed consent prior to participation.

Datasets and analysis scripts to reproduce the reported analyses will be uploaded to the OSF. In addition, the task itself is also provided open-source. Where participant privacy applies, data are provided in a form consistent with consent and platform policies.

7 References

Agashe, S., Fan, Y., Reyna, A., & Wang, X. E. (2025). LLM-coordination: Evaluating and analyzing multi-agent coordination abilities in large language models. *arXiv preprint*.

Akata, E., Schulz, L., Coda-Forno, J., Oh, S. J., Bethge, M., & Schulz, E. (2025). Playing repeated games with large language models. *Nature Machine Intelligence*, *9*, 1380–1390.

Bacharach, M. (1999). Interactive team reasoning: A contribution to the theory of co-operation. *Research in Economics*, *53*, 117–147.

Bratman, M. E. (1992). Shared cooperative activity. *The Philosophical Review*, *101*, 327–341.

Chater, N. (2023). How could we make a social robot? A virtual bargaining approach. *Philosophical Transactions of the Royal Society B*, *381*, 20220040.

Chater, N., Zeitoun, H., & Melkonyan, T. (2022). The paradox of social interaction: Shared intentionality, we-reasoning, and virtual bargaining. *Psychological Review*, *129*(3), 415–437. https://doi.org/10.1037/rev0000343

Chen, D. L., Schonger, M., & Wickens, C. (2016). oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, *9*, 88–97.

Critcher, C. R., Inbar, Y., & Pizarro, D. A. (2013). How quick decisions illuminate moral character. *Social Psychological and Personality Science*, *4*(3), 308–315. https://doi.org/10.1177/1948550612457688

De Neys, W., & Raoelison, M. (2025). Humans and LLMs rate deliberation as superior to intuition on complex reasoning tasks. *Communications Psychology*, *3*(1), 141. https://doi.org/10.1038/s44271-025-00320-8

Dvorak, F., Stumpf, R., Fehrler, S., & Fischbacher, U. (2025). Adverse reactions to the use of large language models in social interactions. *Communications Psychology*, *4*, pgaf112.

Edossa, F. W., Gassen, J., & Maas, V. S. (2024). Using large language models to explore contextualization effects in economics-based accounting experiments. *Working paper*.

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology*, *87*(3), 327–339. https://doi.org/10.1037/0022-3514.87.3.327

Hagendorff, T., Fabi, S., & Kosinski, M. (2023). Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science*, *3*(10), 833–838. https://doi.org/10.1038/s43588-023-00527-x

Kupor, D. M., Tormala, Z. L., Norton, M. I., & Rucker, D. D. (2014). Thought calibration: How thinking just the right amount increases one's influence and appeal. *Social Psychological and Personality Science*, *5*(3), 263–270. https://doi.org/10.1177/1948550613499940

Lin, S., Keysar, B., & Epley, N. (2010). Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology*, *46*(3), 551–556. https://doi.org/10.1016/j.jesp.2009.12.019

McGraw, A. P., & Tetlock, P. E. (2005). Taboo trade-offs, relational framing, and the acceptability of exchanges. *Journal of Consumer Psychology*, *15*(1), 2–15. https://doi.org/10.1207/s15327663jcp1501_2

Misyak, J. B., & Chater, N. (2014). Virtual bargaining: A theory of social decision-making. *Philosophical Transactions of the Royal Society B*, *369*, 20130487.

Phan, L., Gatti, A., Han, Z., Li, N., Hu, J., Steiner, H., ... & Bowman, S. R. (2026). A benchmark of expert-level academic questions to assess AI capabilities. *Nature*, *649*, 1139–1146.

Phillips, J., Buckwalter, W., Cushman, F., Friedman, O., Martin, A., Turri, J., Santos, L., & Knobe, J. (2020). Knowledge before belief. *Behavioral and Brain Sciences*, *44*, e140. https://doi.org/10.1017/S0140525X20000618

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 10, pp. 173–220). Academic Press.

Schelling, T. C. (1980). *The strategy of conflict* (2nd ed.). Harvard University Press.

Searle, J. R. (1990). Collective intentions and actions. In P. R. Cohen, J. Morgan, & M. E. Pollack (Eds.), *Intentions in communication* (pp. 401–415). MIT Press.

Snell, C., Lee, J., Xu, K., & Kumar, A. (2024). Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*. https://arxiv.org/abs/2408.03314

Tetlock, P. E. (2003). Thinking the unthinkable: Sacred values and taboo cognitions. *Trends in Cognitive Sciences*, *7*(7), 320–324. https://doi.org/10.1016/S1364-6613(03)00135-9

Tomasello, M., & Carpenter, M. (2007). Shared intentionality. *Developmental Science*, *10*, 121–125.

Tomasello, M., Melis, A. P., Tennie, C., Wyman, E., & Herrmann, E. (2012). Two key steps in the evolution of human cooperation: The interdependence hypothesis. *Current Anthropology*, *53*, 673–692.

Uhlmann, E. L., Zhu, L. L., & Tannenbaum, D. (2013). When it takes a bad person to do the right thing. *Cognition*, *126*(2), 326–334. https://doi.org/10.1016/j.cognition.2012.10.005

Yao, J., et al. (2025). SPIN-bench: How well do LLMs plan strategically and reason socially? *arXiv preprint*.

Yax, N., Anlló, H., & Palminteri, S. (2024). Studying and improving reasoning in humans and machines. *Communications Psychology*, *2*(1), 51. https://doi.org/10.1038/s44271-024-00091-8