Normative Common Ground Replication (NormCoRe): Replication-by-Translation for Studying Norms in Multi-Agent AI

LUCA DECK*, University of Bayreuth & Fraunhofer FIT, Germany
SIMEON ALLMENDINGER*, University of Bayreuth & Fraunhofer FIT, Germany
LUCAS MÜLLER, University of Bayreuth, Germany
NIKLAS KÜHL, University of Bayreuth & Fraunhofer FIT, Germany

In the late 2010s, the fashion trend *NormCore* framed sameness as a signal of belonging, illustrating how norms emerge through collective coordination. Today, similar forms of normative coordination can be observed in systems based on Multi-agent Artificial Intelligence (MAAI), as AI-based agents deliberate, negotiate, and converge on shared decisions in fairness-sensitive domains. Yet, existing empirical approaches often treat norms as targets for alignment or replication, implicitly assuming equivalence between human subjects and AI agents and leaving collective normative dynamics insufficiently examined. To address this gap, we propose *Normative Common Ground Replication* (*NormCoRe*), a novel methodological framework to systematically translate the design of human subject experiments into MAAI environments. Building on behavioral science, replication research, and state-of-the-art MAAI architectures, *NormCoRe* maps the structural layers of human subject studies onto the design of AI agent studies, enabling systematic documentation of study design and analysis of norms in MAAI. We demonstrate the utility of *NormCoRe* by replicating a seminal experimental study on distributive justice, in which participants negotiate fairness principles under a "veil of ignorance". We show that normative judgments in AI agent studies can differ from human baselines and are sensitive to the choice of the foundation model and the language used to instantiate agent personas. Our work provides a principled pathway for analyzing norms in MAAI and helps to guide, reflect, and document design choices whenever AI agents are used to automate or support tasks formerly carried out by humans.

CCS Concepts: • **Human-centered computing** → **Collaborative and social computing design and evaluation methods**; **Empirical studies in collaborative and social computing**; • **Information systems** → *Decision support systems*; • **Computing methodologies** → *Artificial intelligence.*

Additional Key Words and Phrases: Multi-Agent AI, Fairness, Social Norms, Ethical Norms, Experimental Studies, Replication Studies, Veil of Ignorance

ACM Reference Format:
Luca Deck, Simeon Allmendinger, Lucas Müller, and Niklas Kühl. 2026. Normative Common Ground Replication (NormCoRe): Replication-by-Translation for Studying Norms in Multi-Agent AI. https://doi.org/10.1145/3805689.3806731

*These authors contributed equally to this work.

Authors' Contact Information: Luca Deck, luca.deck@uni-bayreuth.de, University of Bayreuth & Fraunhofer FIT, Bayreuth, Germany; Simeon Allmendinger, simeon.allmendinger@uni-bayreuth.de, University of Bayreuth & Fraunhofer FIT, Munich, Germany; Lucas Müller, lucas.c.mueller@gmail.com, University of Bayreuth, Bayreuth, Germany; Niklas Kühl, kuehl@uni-bayreuth.de, University of Bayreuth & Fraunhofer FIT, Bayreuth, Germany.

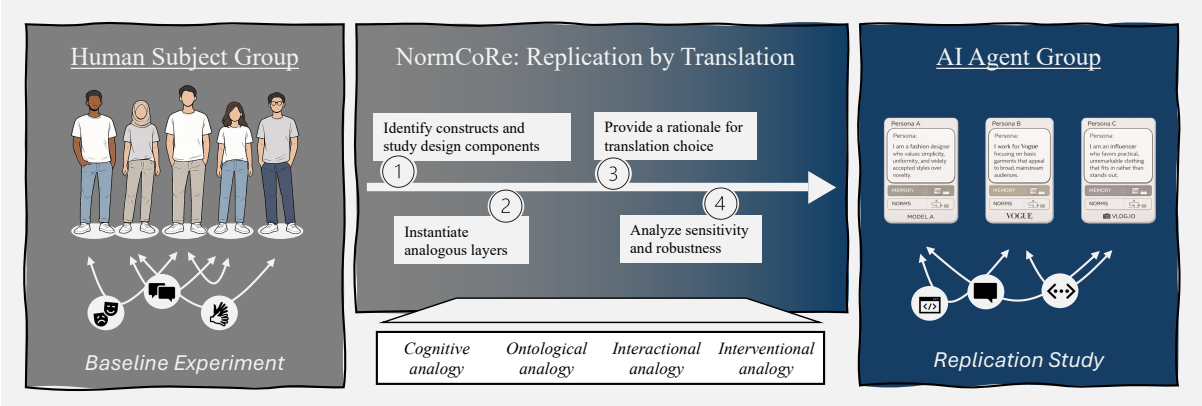

Fig. 1. From human groups to multi-agent AI: *NormCoRe* conceptualizes replication as a translation problem, mapping human subject studies to AI agent studies to study how collective normative judgments—such as fairness—emerge and differ across populations.

1 Introduction

"Once upon a time people were born into communities and had to find their individuality. Today people are born individuals and have to find their communities."

— K-HOLE, inventors of the term "Normcore"

In the late 2010s, the fashion trend *NormCore* embraced deliberate sameness: conventional clothing as a signal of belonging rather than a mark of distinction. What appeared aesthetic was fundamentally normative; an emergent agreement about what is considered appropriate within a group. Such norms arise not from isolated individuals, but from collective coordination. Today, similar forms of normative coordination are increasingly relevant in systems based on Multi-agent AI (MAAI). In contrast to individual AI agents, in MAAI [2, 31] AI-based agents can deliberate, negotiate, and converge on shared decisions. In doing so, they explicitly or implicitly exhibit social and ethical norms [15], in particular as MAAI systems are already being developed in domains governed by fairness and other norms, such as resource allocation [53] or autonomous driving [44]. If we examine this field from a more traditional perspective, experimental research on social and ethical norms is deeply rooted in philosophy, psychology, and game theory, with well-established methods [7, 24, 30]. For example, preferences regarding wealth distribution have been extensively studied through game-theoretical group experiments [17, 19]. However, what these studies all have in common is that they are examining norms among *humans*.

Conversely, research on AI agents—particularly AI agent groups—has only sparsely engaged with the methodological foundations of studying social and ethical norms. At the same time, recent research is already proposing to replace or supplement human subjects with AI agents [4, 51], replicating existing human subject studies using AI agents [15], or attempting to align AI agents with normative principles [50]. Without systematically acknowledging the fundamental differences between human subjects and AI agents, this obscures a critical question for fairness and accountability: how do collective normative judgments emerge, stabilize, and differ when decision-making is delegated to MAAI rather than human groups? Regardless of the underlying goals of AI agent studies, the complexities arising from translating human subject studies to MAAI necessitate a sound methodological foundation that is currently lacking—particularly when social and ethical norms are the focus of such studies. For example, when MAAI systems are tasked with making implicit or explicit decisions about resource allocations, the sheer number of degrees of freedom in the design of the MAAI system, ranging from

the selection of a foundation model (e.g., Large Language Model (LLM)) to the design of task-specific workflows, impedes a thorough evaluation of design choices.

Against this backdrop, we propose Normative Common Ground Replication *NormCoRe*[1]: a novel methodological framework for translating human subject group studies into MAAI environments, enabling researchers to *systematically* investigate social norms in AI agent studies (see Figure 1). Building on established principles from behavioral science and MAAI research, NormCoRe maps the analogous layers of human subject studies onto the design of AI agent studies with MAAI. NormCoRe allows researchers to systematically select and document the configuration of AI agent studies.

We demonstrate the utility of NormCoRe by replicating Frohlich and Oppenheimer [19]'s influential study on distributive justice in an MAAI setting, the complete code including all prompts used is available on Github [16]. The selected baseline study serves as a perfect showcase, as it conceptualizes a complex but well-known norm (e.g., fairness) as a group-level, normative judgment reached through dynamic deliberation and includes all relevant layers for replication. In the original experiment, Frohlich and Oppenheimer [19] operationalized John Rawls' veil of ignorance" [40] within a controlled laboratory setting. Participants were unaware of their assigned income class—simulating the "veil of ignorance"—and were tasked with reaching a consensus on a distributive justice principle that would determine their payoff in a hypothetical society. By combining individual normative deliberation, goal-based optimization, and dynamic consensus finding, this experiment serves as an ideal testbed for demonstrating how NormCoRe organizes complex design choices and facilitates precise reporting. Instantiating NormCoRe with Frohlich and Oppenheimer [19]'s study, we show that the choice of the LLM and the language of the persona description have a significant influence on fairness judgments in AI agent studies. Also, we find that while both populations favor the same principle (maximizing overall income while ensuring a floor constraint for the worst-off), MAAI groups demonstrate a substantially higher concentration on this principle than human groups.

This work makes three contributions to the rigorous design and analysis of MAAI:

- We establish a novel *replication-by-translation* perspective on replication studies with AI agents that explicitly accounts for the fundamental differences between AI agents and human subjects (Section 2).
- Based on this perspective, we introduce NormCoRe as a methodological framework for the systematic replication of human subject studies in MAAI settings, providing a lens through which researchers can document and analyze social and ethical norms in AI agents (Section 3).[2]
- By employing NormCoRe to a seminal baseline study on distributive justice, we demonstrate the usefulness of NormCoRe and empirically show that norms in MAAI not only differ from human baselines but are also sensitive to study design decisions (Section 4).

We discuss the implications of our study and broaden the discourse on the purpose and open challenges of AI agent studies in Section 5, paving the way for future research on norms in MAAI. It is reasonable to expect that the implementation of MAAI in studies and industry will happen one way or another. However, to judge whether this is a positive or a reprehensible development—or under which conditions MAAI systems and studies are actually beneficial—we need to better understand the structure, dynamics, and impact of MAAI systems in the first place. Our work raises critical normative questions regarding the future of agentic automation beyond distributive justice and cautions designers to make conscious, evidence-based choices in both laboratory experiments and real-world applications. Whenever tasks formerly carried out by (groups of) humans are to be automated or supplemented by groups of AI agents, NormCore helps to guide, reflect, and document design choices and to study potential impact of including AI agents.

[1]The method does not coincidentally bear the same name as the fashion phenomenon described above.
[2]Our codebase repository is available under https://github.com/Lucas-Mueller/Normative_Common_Ground_Replication_NormCoRe

2 Background

This section reviews prior work on replicability as a methodological principle (Section 2.1) and its application to AI-based replication of human subject studies. We then identify key assumptions and challenges in existing approaches, particularly the treatment of human–AI replication as equivalence rather than translation (Section 2.2), which motivates the need for a novel methodology (Section 2.3).

2.1 Importance of Replicability for the Scientific Method

Replicability constitutes a fundamental pillar of the scientific method. In principle, the transparent documentation of research methodologies enables independent verification and replication of empirical findings by the broader academic community. In that sense, replication can serve two functions: authentication of original findings and boundary testing to understand generalizability [54]. Consequently, two replication approaches are distinguished: *direct replication*, which repeats an experiment with minimal to no changes to authenticate original findings, and *conceptual replication*, which tests the same hypothesis with different methods, stimuli, or populations, aiming to probe whether findings generalize to new conditions [26, 54]. Although sometimes treated as synonyms, a critical distinction exists between reproducibility, where existing data is re-analyzed with the same methods, and replicability, where experiments are repeated, resulting in new data [39].

2.2 Replicability of Human Subject Studies with AI Agents

Meanwhile, a growing body of research across disciplines is conceptually replicating human experiments with LLM-based AI systems [1, 9, 21, 23, 28, 35]. These efforts pursue two distinct objectives. First, replication is employed to assess whether AI agents can serve as a valid proxy for human participants in experimental research [15, 52]. Second, replication is used to compare humans and AI as a lens for better understanding the psychological, behavioral, and normative properties of AI systems [29].

Work pursuing the first objective typically evaluates replication success in terms of statistical similarity between human and AI-generated responses. For example, Yeykelis et al. [52] assessed the replicability of consumer behavior research by replicating 133 results from 45 studies published in the Journal of Marketing, using LLM-based AI personas programmed to match the original participant demographics. They report that 76% of main effects and 68% of interaction effects could be reproduced. Similarly, in Psychology and Management Science, Cui et al. [15] replicated 156 experiments published over the past decade and found that in 73% and 81% of main effects, respectively, were replicated, with interaction effects rates ranging between 46% and 63%.

In contrast, work aligned with the second objective treats replication outcomes not as validation but as a diagnostic tool for understanding how and where AI behavior diverges from human cognition. In behavioral economics, Leng [29] replicated canonical experiments on prospect theory, framing, and mental accounting (e.g., [25, 46]). While LLMs partially reproduce human mental accounting behavior, they appear substantially more rational, exhibiting weak framing effects and limited transaction utility. Crucially, these behavioral patterns vary systematically with design choices such as the language of the prompt, with Spanish and French prompts showing more human-like loss aversion than English and Chinese prompts [29].

A related line of work employs replication to assess whether MAAI aligns with human moral judgments, implicitly treating human consensus as a normative benchmark. Within this approach, moral soundness is operationalized as statistical similarity between human and AI responses to ethically charged dilemmas. A prominent example is the replication of the Moral Machine experiment, in which human participants were asked to resolve trolley-problem-style dilemmas and their aggregate judgments were taken as indicative of moral preferences [3]. When this experiment is applied to LLM-based agents, replication results show that moral judgments vary substantially depending on model architecture and training regime [45]. Such approaches also gain traction in the industry, as evidenced by Anthropic's creation and use of the GlobalOpinionQA dataset,

which contains 2556 questions and human answers on ethics and current events from the World Values and Pew Global Attitudes surveys [18]. Crucially, Anthropic evaluates their LLM against the normative target that the model should reflect a country's specific distribution of opinions when prompted—implying, e.g., that a model prompted in Russian ought to adopt a distinctively "Russian" moral perspective.

The implicit assumption underlying this stream of research is that convergence with human consensus constitutes ethical adequacy [27]. From a philosophical perspective, this assumption mirrors the naturalistic fallacy, which cautions against "oughts" from descriptive "is" statements about observed human behavior [36]. A more fundamental limitation of this literature concerns the conception of the experimental subject itself. Many replication studies implicitly treat LLM-based agents as functional substitutes for human participants, evaluating success primarily in terms of behavioral or statistical similarity [9, 28, 35]. However, an LLM is fundamentally different from a human being with physical embodiment, lived experience, and moral agency, and a prompt-based persona is not ontologically equivalent to a human subject. By overlooking these differences, replication is often framed as a direct transfer rather than as a translation between fundamentally different kinds of subjects, whose cognitive substrates, experience, and agency differ in principled ways [1, 21, 23]. These risks obscure how observed similarities or divergences in normative outcomes are shaped by design choices and structural differences rather than genuine equivalence.

2.3 Replication-by-Translation in AI Agent Studies

The fundamental ontological differences between human subjects and AI agents imply that replication across populations cannot be accommodated through a straightforward transfer of study design. Instead, studies that replicate human experiments with LLM-based agents face a crucial methodological challenge: each replication necessarily involves translating human experimental constructs into AI-compatible designs. This translation introduces degrees of freedom at multiple levels of the AI system, including the choice of foundation model encoding background knowledge, the design of personas and prompts that instantiate subject-like behavior, and the orchestration protocols that structure group interaction. Such translation choices are rarely documented systematically, yet they fundamentally condition the outcomes being compared, as demonstrated in recent research. For example, slight variations [41], formatting changes [49], and framing [13] in prompt design can alter the performance of an LLM significantly. Similarly, in multi-agent settings, modifications to deliberation protocols, including speaking order [6], debate termination timing, and adversarial structure play a decisive role in determining collective outcomes [32].

Yet, the replication studies reviewed above typically treat these design choices as implementation details rather than methodological choices. This impedes rigorous interpretation of study results. For example, when human and AI populations yield different outcomes, it cannot be determined whether the differences reflect genuine divergence or a result of arbitrary translation choices. Addressing this rigor requires treating replication across populations as an explicit translation problem—one that renders translation decisions visible at each layer of the experiment. Only then can observed differences be attributed to specific sources rather than confounded across the translation process. To facilitate the standardization of replication practices in line with broader scientific goals [54], we propose *NormCoRe* as a novel methodological framework for AI agent-based replication studies.

3 Normative Common Ground Replication (NormCoRe)

NormCoRe is a systematic method for replicating *human subject studies* with *AI agent studies* in order to (empirically) investigate the normative common ground between human and AI groups. NormCoRe adapts established replication logic from the social sciences to cross-population settings involving fundamentally different kinds of groups. Rather than assuming equivalence between human subjects and AI agents, NormCoRe explicitly treats replication as a translation problem to systematically map the constructs, interventions, and

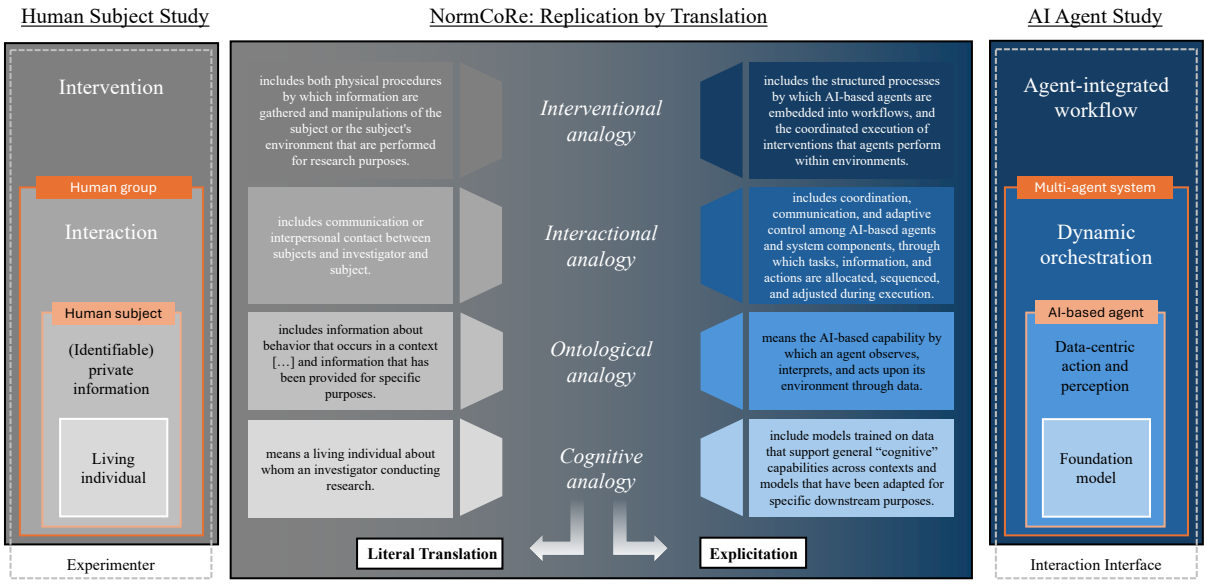

Fig. 2. The four translation layers illustrating the necessary analogies between individual layered components of human subject studies and AI agent studies. Some components may be translated "literally", e.g., when the study sequence can be fully adopted. Other components may require "explicitation", e.g., when AI agents participate in a discussion in fixed turns.

interactions of human-subject studies into an analogous study with AI agents, while minimizing threats to validity. The central goal of NormCoRe is not to determine whether AI agents behave "like humans," but to identify where normative (group) judgments converge or diverge, and to attribute such differences to replication choices. This enables a disentangled investigation of questions such as whether fairness judgments in MAAI systems are sensitive to model choice, persona, or memory design, and how orchestration and coordination mechanisms shape collective decision-making trajectories.

3.1 NormCoRe as Replication-by-Translation

In contrast to direct replication [48] within a single population, cross-group replication between humans and AI agents necessarily introduces conceptual degrees of freedom arising from differences in four translation layers: *cognition analogy, ontological analogy, interactional analogy* and *interventional analogy*. We distinguish two ideal-typical translation choices as methodological parameters to implicitly acknowledge underlying differences: (i) *literal translation for direct replication*, which aims to preserve surface-level experimental features (e.g., task structure, payoff matrices, information availability), (ii) *explicitation for analogous replication*, which makes implicit assumptions in human studies explicit and operationalizes them as design choices in AI agent studies (e.g., decision rules, memory structures).

3.2 Layered Translation Structure in NormCoRe

NormCoRe operationalizes replication between human subject studies and AI agent studies through a layer-by-layer translation method illustrated in Figure 2. Rather than first defining human subject and AI agent studies independently, NormCoRe aligns them at the four corresponding layers of abstraction, following the nested structure of the U.S. Common Rule for human subject studies (45 CFR §46.102 [38]) in combination with the

layered architecture of multi-agent AI [2]. This approach makes the translation challenge explicit at each layer, allowing normative outcomes to be interpreted relative to specific sources of variation and reducing the risk of what we term *translation hacking*, i.e., the selective tuning of translation-layer design choices to obtain desired outcomes.

Layer 1: Living Individual → Foundation Models. At the most fundamental layer, human subjects are defined as *living individuals* about whom information is obtained through intervention or interaction [38]. Normative outcomes in human studies are already shaped at this level by background knowledge, lived experience, and informational asymmetries. These factors are not themselves experimental manipulations, yet they condition all downstream perception, deliberation, and behavior. In AI agent studies, the structurally corresponding layer is the *foundation model* substrate. Foundation models encode large-scale statistical regularities derived from pretraining data and thereby constitute the epistemic and normative background against which all agent behavior unfolds [10]. While private information in human studies is generated within the experiment, foundation models represent a pre-experimental informational endowment that cannot be directly manipulated during execution but strongly conditions normative judgments. NormCoRe treats this correspondence as a cognitive analogy. The goal is not to equate living individuals with foundation models, but to acknowledge that both possess a base layer of informational constraint that shapes what kinds of cognitive heuristics or normative judgments are even expressible. Translation choices at this layer (e.g., model family, pretraining corpus) introduce degrees of freedom that must be documented to ensure interpretability of replication results.

Layer 2: Human Subject → AI-based Agent. Building on this informational substrate, the second layer concerns the *subject* itself. Human subjects are grounded in the notion of (identifiable) private information that is used, studied, analyzed, or generated within the research context [38]. This layer establishes the epistemic basis of the study: what is known about the subject, how that knowledge is obtained, and under which contextual constraints. This definition carries implicit assumptions about agency, perception, memory, and the capacity to form normative judgments. In AI agent studies, the corresponding layer is the agent's *data-centric perception–action capability*: the mechanisms through which an agent observes its environment, interprets inputs, and generates actions [2]. This includes prompt structures, persona specifications, memory mechanisms, and decision heuristics that instantiate agency-like behavior in computational form. NormCoRe frames this mapping as an *ontological analogy*. Human agency is not replicated, but functionally translated into computational capacities that allow agents to participate meaningfully in normative tasks. Explicit design choices at this layer—such as whether agents possess persistent memory or adopt stable personas—directly affect normative outcomes and must therefore be treated as core elements of the replication design rather than as mere implementation details.

Layer 3: Human Group → Multi-agent System. The third layer concerns *interaction*. In human groups, interaction encompasses communication or interpersonal contact *between* subjects and investigators, as well as *among* subjects. This layer structures deliberation, persuasion, power dynamics, and social learning, all of which are central to the emergence of shared norms such as fairness [20]. In AI agent studies, the corresponding layer is *dynamic orchestration* creating a Multi-agent system. This includes message passing protocols, turn-taking rules, negotiation mechanisms, and adaptive control processes that determine how agents exchange information and influence one another over time [2]. NormCoRe treats this correspondence as an interactional analogy. Translation at this layer often requires explicitation: assumptions that are often implicit in human interaction (e.g., (equal) speaking rights, shared understanding of rules) must be made explicit as protocol constraints or coordination mechanisms in AI agent studies.

Layer 4: Intervention → Agent-integrated Workflow. At the highest layer, human subject studies involve *interventions*, defined as physical procedures or informational manipulations performed for research purposes [38]. Interventions include task framing, incentive structures, timing, and constraints that are intentionally varied to study causal effects. In AI agent studies, the corresponding layer is the *agent-integrated workflow* [2]. This includes the structured processes by which agents are embedded into experimental workflows, the sequencing of tasks,

role assignments, and the coordinated execution of actions within an environment. This mapping constitutes an interventional analogy. NormCoRe emphasizes that such translations are not neutral: different design choices can systematically shape normative outcomes [43]. Accordingly, intervention-level explicitation choices must be explicitly justified and reported.

In addition to the formally specified translation layers, the role of the experimenter should be explicitly acknowledged as an "unknown influence". In human subject studies, experimenters inevitably shape outcomes through subtle cues, framing choices, timing, and interaction styles—often unintentionally and outside the scope of formal interventions. In AI agent studies, analogous influences arise through interaction interfaces. Making this influence explicit does not eliminate it, but improves interpretability by preventing ungrounded attribution of normative differences solely to agents or models, when they may partly reflect researcher-induced artifacts. To operationalize this concern and to ensure methodological interpretability, we summarize the NormCoRe method in the following Box.

NormCoRe Method

We propose a method for the interpretable replication of human subject studies with AI agent studies by establishing *disentangled, layer-specific translation analogies*, rather than assuming global equivalence. Replication success is evaluated in terms of *explanatory alignment*, not behavioral identity.

For each NormCoRe translation exercise, researchers must:

Step 1 **Identify constructs & study design components:** Specify the theoretically relevant constructs in the original human subject study together with the concrete study design components through which these constructs are operationalized (e.g., task framing, information structure, incentive schemes, interaction rules, and experimenter involvement).

Step 2 **Instantiate analogous layers:** Instantiate the analogous MAAI component(s) of the corresponding NormCoRe layer (e.g., foundation model choice for the cognitive layer, persona prompting and memory for the ontological layer, orchestration protocols for the interactional layer).

Step 3 **Provide a rationale for translation choices:** Justify whether the layer mapping constitutes:
- *Literal translation* (for exact/direct replication), or
- *Explicitation* (for analogous replication), and specify the replication type implied by the design choice (e.g., constructive, incremental, quasirandom, or comprehensive [26]).

Step 4 **Analyze sensitivity & robustness:** Analyze the sensitivity of normative outcomes to layer-specific translation choices using established statistical criteria for replication, non-replication, and partial replication evidence [11].

4 Experimental Study: Applying NormCoRe for Fairness Principles

With the methodological framework in place, we illustrate its application for a representative norm that has both a social and ethical dimension as well as a crucial downstream impact in MAAI systems: fairness. Fairness is a norm deeply embedded in social science and philosophy, guiding perceptions, legal frameworks, economic interactions, and the design of AI systems [8, 37]. While its meaning varies across disciplines, contexts, and cultures, several simplifying approaches have been proposed to study fairness principles in experiments with human subjects. One seminal experiment is Frohlich and Oppenheimer [19] who applied John Rawl's popular thought experiment of the "veil of ignorance" [40] to study preferences for four predefined fairness principles representing different schools of thought in political philosophy. The study has been frequently cited and serves as a perfect showcase for the utility of NormCoRe, as it offers value-laden tradeoffs and disagreement between individuals, and includes all relevant layers for replication and sufficiently complex normative interactions to

Table 1. Distribution of fairness principle agreements in the human baseline and MAAI replication, including baseline alignment and sensitivity to translation-layer design choices (foundation model and persona language).

	Baseline Alignment		Translation Sensitivity				
	Human–MAAI		Cognitive Layer		Ontological Layer		
Fairness Principle	**Baseline**	**MAAI**	**Chinese LLM Ecosystem**	**U.S. LLM Ecosystem**	**English**	**Mandarin**	**Spanish**
Max. Floor	1	0	14	4	0	1	0
Max. Avg. Income	1	1	0	2	0	0	0
Max. Avg. + Floor	23	29	15	21	30	27	17
Max. Avg. + Range	2	0	0	0	0	0	2
No Agreement	7	3	4	6	4	6	15
Total	**34**	**33**	**33**	**33**	**34**	**34**	**34**

study the dynamics of MAAI (e.g., as opposed to the Ultimatum Game [47] with anonymous and transactional interactions).

The following section describes the original study (Section 4.1) and instantiates the NormCoRe framework by translating the original study to an MAAI setting [2] to study individual judgments and group dynamics of MAAI systems in the context of fairness principles (Section 4.2). This instantiation is meant to validate the methodological framework in a concrete application and demonstrates the importance of design choices in AI agent studies (Section 4.3).

4.1 Baseline Human Subject Study by Frohlich and Oppenheimer

Frohlich and Oppenheimer [19] conducted their experiment with 34 groups of five university students, comprising an individual and a group phase. In the individual phase, participants received an introduction to four fairness principles in the context of income distribution: (P1) maximizing the income of the worst-off individual; (P2) maximizing average (and thus total) income; (P3) maximizing average income subject to a guaranteed minimum income; and (P4) maximizing average income subject to a cap on income inequality. Participants first ranked the principles from most to least preferred and reported their confidence on a five-point Likert scale. They were then shown four alternative income distributions across five income classes with known probabilities representing a "probabilistic veil of ignorance" (5%, 10%, 50%, 25%, and 10%) and informed which distribution corresponded to each principle. After additional instruction and a comprehension test, participants ranked the principles again.

Next, participants completed four payoff-relevant practice rounds. In each round, they selected a principle, after which a corresponding distribution was implemented via a random draw assigning them to one of five income classes, essentially "lifting" the "veil of ignorance". While class probabilities were fixed, participants were unaware of their exact values. Realized payoffs, as well as counterfactual payoffs under alternative principles, were revealed and paid immediately at a 1:$10,000 conversion rate. The individual phase concluded with a third ranking and confidence assessment. In the group phase, each five-person group deliberated to reach a unanimous agreement on a single principle. Prior to the discussion, participants were informed that (i) the payoff distributions used for group payment could differ from the examples, and (ii) the group decision would determine binding payoffs at higher stakes. Unlike in the individual phase, participants did not know the specific distributions in advance. The discussion lasted at least five minutes and ended with a verbal consensus and a confirming secret-ballot vote; if unanimity was not achieved, payoffs were determined by a random draw. Finally, participants submitted a last ranking with confidence ratings.

4.2 NormCoRe Translation to AI Agents

Following the NormCoRe translation procedure, replication is treated as a layered translation problem rather than an assumption of equivalence. Translation decisions are therefore made explicit and justified at each analogy layer.

(Step 1) Identify constructs & study design components: The baseline experiment investigates normative preferences for distributive justice principles under a Rawlsian veil of ignorance. The focal construct is the selection and ranking of four predefined justice principles (P1–P4), measured at both the individual level (repeated rank-orderings with confidence) and the group level (unanimous consensus determining payoffs). Core design components that operationalize this construct include: (i) controlled information about income distributions and probabilities, (ii) payoff-relevant decision making via stochastic assignment to income classes, and (iii) structured group deliberation culminating in a binding collective choice. All translation choices in the AI agent study are evaluated in relation to their ability to preserve the fairness decision problem while rendering it executable for LLM-based agents.

(Step 2) Instantiate analogous layers: NormCoRe aligns the human-subject experiment with the AI-agent study through a set of layered analogies. A complete, parameter-level specification of all mappings is provided in tables 2 to 5 in the Appendix; here we summarize the most salient aspects.

Cognitive and ontological analogy. Human participants—university students deliberating under uncertainty—are mapped to LLM-based AI agents. In the baseline study, subjects' cognitive capacities (e.g., language comprehension, memory, and reasoning ability) and ontological properties (e.g., agency, identity, and persistence across interactions) are implicit and embodied. In the AI agent study, these properties must be made explicit and operationalized. Accordingly, agents are instantiated with configurable role descriptions approximating the original participant pool, managed memory with explicit character limits to support learning-in-context, and controlled linguistic and stochastic parameters (language choice and temperature). The complete prompts used to instantiate all agent roles are made available in our repository [16]; consistent with NormCoRe's core principle, they are treated as explicit and auditable translation choices rather than definitive claims about any specific model's normative behavior This translation preserves the functional role of the subject—forming and revising normative judgments—without asserting equivalence between human subjects and LLM-based agents (Tables 2 to 5).

Interactional and interventional analogies. Human group interaction and experimental procedures are translated into a turn-based orchestration protocol and an agent-integrated workflow. Free-form discussion is formalized as sequential speaking with equal opportunity, shared discussion history, and a structured consensus mechanism, while the experimenter's role and laboratory environment are translated into structured prompts, computational execution, symbolic payoffs, and explicit randomness controls. These mappings ensure that deliberation, incentives, and information flow remain comparable across populations while accommodating the technical constraints of AI agents. Full details of these mappings are documented in tables 2 to 5.

(Step 3) Provide a rationale for translation choices: All translation choices are classified; a comprehensive classification and justification for each design decision is provided in tables 2 to 5 in the Appendix.

(Step 4) Analyze sensitivity & robustness: NormCoRe requires that translation choices introducing degrees of freedom be either controlled or systematically varied. Accordingly, the replication incorporates several robustness mechanisms. First, reproducibility controls (explicit random seeds, bounded discussion rounds, and temperature parameters) enable deterministic reruns and isolate stochastic variation. Specifically, we use 33 AI groups to closely approximate the human baseline of 34 groups while allowing the sample to be divisible by three. This enables three temperature conditions: 0 (deterministic), random draws from $[0, 1]$, and random draws from $[0, 1.5]$, representing increasingly stochastic generation regimes. Second, sensitivity analyses are embedded directly into the design: agent language (English, Mandarin, Spanish) is varied to assess the stability

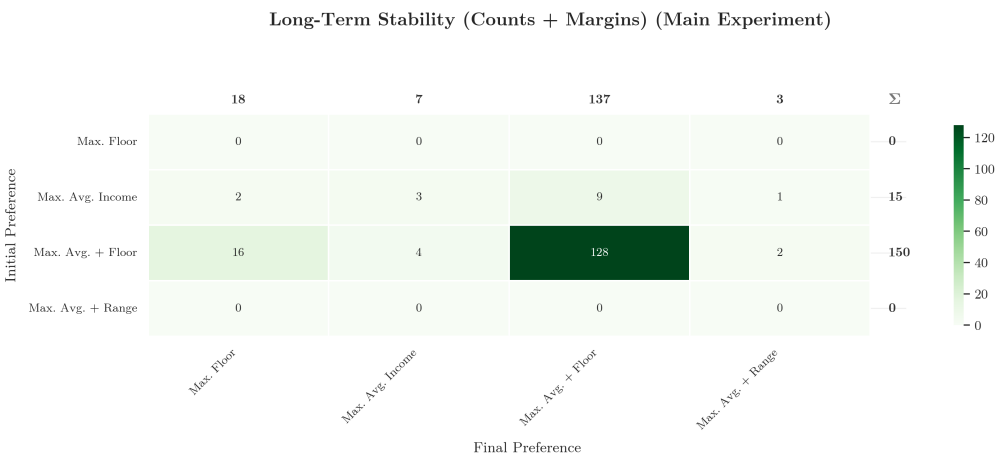

Fig. 3. Individual-level preference ranking transitions before and after group deliberation. The vertical (horizontal) axis shows initial (final) individual rankings, and cell intensities reflect transition frequencies, highlighting strong convergence toward maximizing average income with a floor constraint.

of normative outcomes. Third, structural constraints (e.g., validation of floor and range parameters, minimum statement lengths, and capped memory) prevent degenerate or nonsensical trajectories that would confound interpretation.

4.3 Results of the Replication Study

We begin by examining *baseline alignment* between the human subject experiment and its MAAI replication. Table 1 reports the distribution of distributive justice principle choices for the human baseline and the aggregated MAAI groups. Consistent with the original human study, both populations predominantly converge on the principle of maximizing average income subject to a guaranteed minimum for the worst-off. At the same time, notable differences emerge at the collective level. MAAI groups exhibit stronger convergence and lower disagreement: 29 of 33 AI groups select this principle, compared to 23 of 34 human groups, with disagreement rates of 9.1% and 20.6%, respectively. These findings indicate that while the dominant fairness preference is aligned across populations, normative outcomes in MAAI are more homogeneous than in human groups. Figure 3 complements the aggregate results in Table 1 by visualizing how *individual-level preference rankings* change over the course of the experiment. The figure plots each participant's (human) or agent's (MAAI) *initial individual ranking* (elicited prior to any group interaction) against their *final individual ranking*, reported after group deliberation and the emergence of a collective decision. Across MAAI experiments, individual agent rankings exhibit substantially stronger convergence toward the eventual *group-level consensus* than observed in human groups. This pattern indicates reduced intra-group variance and stronger aggregation dynamics in MAAI, helping to explain the higher levels of consensus and lower disagreement rates reported in Table 1.

Beyond establishing baseline alignment between human and MAAI group-level outcomes, we examine the robustness of these findings to translation-layer design choices. Sensitivity analyses show that the observed convergence in fairness principles is not invariant, but systematically conditioned by architectural decisions within the AI agent configuration. At the *cognitive analogy layer*, variation in the underlying foundation model ecosystem (Chinese vs. U.S. LLMs) leads to pronounced and consistent shifts in distributive justice principle selection, relative to the baseline alignment reported in Table 1. At the *ontological analogy layer*, sensitivity analyses reveal that

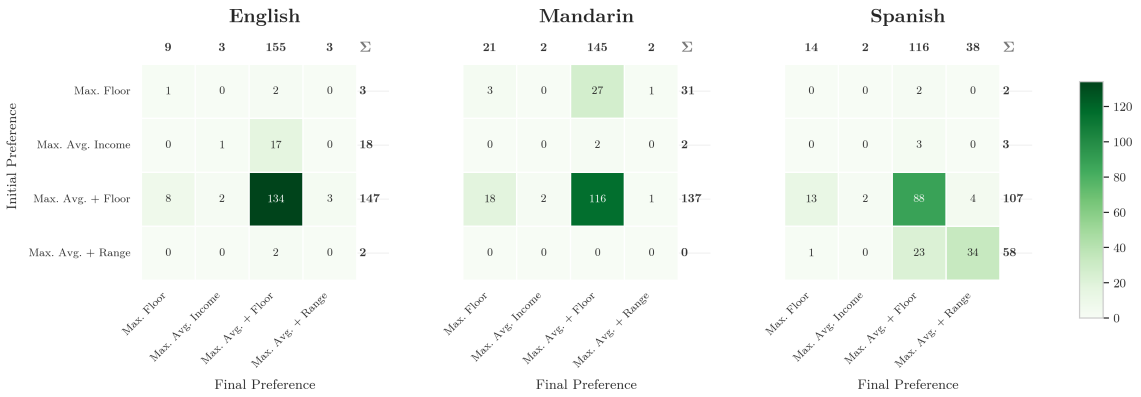

Fig. 4. Preference shifts in individual distributive justice rankings before and after group deliberation, stratified by AI agent language. Across all languages, individual AI agent preferences converge toward maximizing average income.

the language used to instantiate agent personas systematically conditions individual preference formation and its aggregation during deliberation. As shown in Figure 4, agents instantiated in all three languages exhibit convergence toward the same dominant fairness principle observed in the baseline alignment (see Table 1). However, the degree and trajectory of convergence differ across languages. Spanish-language agents display more heterogeneous initial individual preferences and retain greater diversity throughout deliberation, whereas English- and Mandarin-language agents begin from more concentrated initial distributions and converge more rapidly toward the group-level consensus. These language-specific dynamics help explain the variation in consensus rates reported in Table 1, while reinforcing that aggregate alignment with human outcomes can coexist with substantial sensitivity to ontological instantiation choices.

5 Discussion

The increasing integration of MAAI in both workflow automation and empirical research raises fundamental methodological and normative challenges. As AI-based agents increasingly deliberate, coordinate, and converge on collective decisions, understanding the emergence and impact of social and ethical norms in such systems becomes critical. NormCoRe provides a first systematic step toward structuring AI-agent replication studies as an explicit process of translating human subject studies in AI agent studies, acknowledging the fundamental differences between human subjects and AI agents. Accordingly, we discuss what the experimental results imply for replication-by-translation (Section 5.1), the epistemic purpose of AI agent studies (Section 5.2), and the open challenges and limits of this approach (Section 5.3).

5.1 Why Replication Studies with AI Agents Require Translation

The results illustrate both the promise and the limitations of replicating human subject studies with AI agents. At an aggregate level, MAAI groups converge on the same dominant fairness principle as human groups, suggesting that key normative outcomes can be reproduced under comparable decision structures (Table 1). At the same time, our experimental results show that MAAI exhibits substantially higher consensus and lower disagreement rates, indicating stronger aggregation dynamics and reduced variance compared to human groups (Figure 3). Crucially, sensitivity analyses show that these outcomes are not invariant. Varying the foundation model leads to systematic shifts in principle selection (Table 1), and changing the language used to instantiate agent personas affects both convergence and outcome distributions (Table 1 and Figure 4). These effects demonstrate that seemingly stable

normative judgments depend on translation choices that have no direct analogue in the human baseline. For example, the same persona specification (e.g., a "university student") may be instantiated in different prompt languages (English, Mandarin, or Spanish) on foundation models whose pretraining corpora—and thus linguistic and cultural coverage—differ substantially; this introduces variation attributable to translation decisions and model-choice that hardly have a direct analogue in the original human subject study.

While the baseline experiment fromFrohlich and Oppenheimer [19] serves our purpose to instantiate NormCoRe with a suitable and relevant norm, it does not come without limitations and represents only a fraction of what social and ethical norms can entail. As discussed in the original study, Frohlich and Oppenheimer [19] only uses a small sample of Polish, Canadian, and US students, which is not nearly representative of the world population, and adopts a strong simplification of norms into four distributive justice principles, which influences the consensus finding process. Despite these limitations, our findings suggest that similarity in aggregate outcomes is insufficient to establish equivalence between human and AI-based replications. Instead, replication outcomes are already shaped by design decisions at the cognitive and ontological levels of analogy. This underscores the need to treat replication with AI agents as a process of explicit translation rather than direct substitution—precisely the role NormCoRe is designed to fulfill.

5.2 Purpose of AI Agent Studies

Broadening the discourse on AI agent studies, a first open question concerns the epistemic purpose of replication studies with AI agents, e.g., what kind of value such studies provide, and what conclusions should be drawn from them. One motivation can be curiosity-driven comparison, e.g., observing how AI agents behave when placed in experimental settings originally designed for humans. However, a purely descriptive comparison risks under-theorizing the implications of observed similarities or differences. A more substantive motivation is methodological. The scientific community is currently grappling with a well-documented replication crisis, characterized by systematic failures to reproduce published findings across disciplines [5]. Large-scale efforts, such as the Open Science Collaboration's replication of 100 psychological experiments, revealed replication failures in more than half of the cases, with substantially reduced effect sizes [14]. AI-agent replication studies cannot resolve this crisis in a straightforward sense, as they do not constitute replications within the same population. However, when framed appropriately, they can serve as boundary tests that probe the robustness of theoretical constructs under radically different cognitive and organizational substrates [15]. From this perspective, divergence between human and AI-agent outcomes is not a failure but an informative signal. This raises a related question: Are AI-agent studies primarily concerned with norms among AI agents themselves, or with downstream outcomes that affect (human) norms? NormCoRe deliberately accommodates both perspectives. On the one hand, collective norms emerging within MAAI systems are increasingly consequential in their own right, as such systems are delegated fairness-sensitive decisions in practice. On the other hand, understanding how and why AI-agent norms differ from human baselines is essential for anticipating societal impacts and governance challenges.

A recurring critique in the literature questions whether it is appropriate to conduct social or behavioral experiments with AI agents at all [4, 51]. Concerns include sampling bias, the absence of lived experience, and the risk of over-interpreting artificial behavior as psychologically meaningful. These concerns are valid and underscore the importance of methodological caution. At the same time, AI agent studies offer distinctive methodological advantages over human subject studies [35]. Data can be collected at scale, under controlled conditions, and with levels of process transparency that are often unavailable in human-subject research (e.g., by collecting transcripts of thinking models). Moreover, AI agent replications can explicitly vary dimensions (e.g., language, memory, or interaction protocols) that are difficult or impossible to manipulate cleanly in human experiments. Generally speaking, to determine the value of MAAI for different applications and the conditions

under which MAAI can actually be beneficial, we need to better understand the structure, dynamics, and impact of MAAI systems in the first place. We position NormCoRe as a methodlogical framework to advance this understanding.

5.3 Open Challenges of AI Agent Studies

Despite their promise, AI-agent replication studies face several unresolved challenges. A central issue is generalizability. Empirical findings in AI agent studies are inherently time-bound snapshots: foundation models evolve rapidly, whereas human biology and many social mechanisms remain comparatively stable over time. As a result, replication outcomes may change as models are updated, retrained, or replaced. This temporal instability complicates the accumulation of knowledge and reinforces the need for precise and layered documentation of model versions and experimental configurations. A related challenge concerns establishing best practices. There is currently little consensus on how sample sizes should be selected or reported in AI agent studies, particularly when agents can be instantiated cheaply and repeatedly. Similarly, while NormCoRe emphasizes explicitation as a core methodological principle, open questions remain regarding the sufficiency of detail. Moreover, it is unclear how well the behavior of AI agents in an experimental study can be generalized to applied settings. For example, if AI agents systematically associate certain social traits with gender in an experimental context, this may carry over to applied domains such as hiring or evaluation. Emerging interpretability research provides preliminary support for this assumption, suggesting that LLMs internally organize information into relatively stable conceptual representations [33]. Nonetheless, this line of research remains in its infancy, and strong claims about an AI system's "psychology" [22] would be premature.

Cultural generalizability poses another open question. Most current AI—and especially LLM—research relies on models trained predominantly on WEIRD (Western, Educated, Industrialized, Rich, and Democratic) data and instantiated with personas reflecting Western norms [34]. A similar tendency can be observed for participant sampling in AI-related human subject studies [42]. It remains unclear whether non-WEIRD MAAI—or the same systems instantiated with different linguistic and cultural priors—would converge on similar norms or diverge systematically. Conversely, it is an open empirical question whether some AI agent configurations may align more closely with non-WEIRD human populations than with the original human baselines. Moreover, hybrid replications, in which one component of a socio-technical system is held constant while another is translated (e.g., human decision-makers interacting with AI agents, or vice versa), blur the boundary between human-subject and AI agent studies and offer additional degrees of freedom [12]. Similarly, one can imagine translating insights from AI agent studies back into human subject research, which raises several questions on the methodology and the interpretation of results.

Lastly, our study has only shown *that* design choices significantly affect normative outcomes, but reveal little about the exact mechanisms underlying these effects. NormCoRe helps to make influential design factors explicit and supports systematic sensitivity and robustness analyses. Still, illuminating the causal patterns and technical mechanisms behind observed effects requires complementary approaches such as interpretability, ablation studies, or interdisciplinary analysis (e.g., augmented by linguistic or cultural research), which offer exciting avenues for future work.

Taken together, these challenges point to a broader research agenda. Key open questions concern the purpose and value of replication studies with AI agents, the generalizability of observed human–AI differences given unavoidable translation and explicitation choices, the generalization of observed norms across applications and cultures, and investigation into the causal patterns driving the observed effects. NormCoRe does not resolve these questions, but paves the way toward an established methodological foundation, allowing future research to systematically address them. By advocating for rigorous replication-by-translation, NormCoRe contributes to transparency, interpretability, and cumulative progress in the study of social and ethical norms in MAAI.

6 Conclusion and Outlook

As AI agents are increasingly integrated into experimental studies and decision-making processes, methodological rigor in designing these agents becomes critical—especially when they are subject to social and ethical norms. To judge whether and under which conditions the integration of AI agents is actually beneficial, we need to better understand their structure, dynamics, and impact. We introduce NormCoRe as a methodological framework for studying social and ethical norms—such as fairness—in Multi-agent AI (MAAI) setups through the rigorous replication-by-translation of human subject studies in AI agent environments. By accounting for the fundamental differences between human subjects and AI agents and conceiving replication as a layered process of analogous translation, NormCoRe systemizes the design choices shaping normative judgments and outcomes in MAAI systems. Our experimental study, which instantiates the NormCoRe method, demonstrates that fairness judgments in MAAI are sensitive to the choice of the foundation model and the language used to instantiate agent personas. The results reveal significant differences between human subjects in their normative judgments and underscore the importance of well-documented design choices for the examined AI agents. Also, our study indicates that AI agents can converge on fairness principles similar to those favored by human groups, but do so with higher homogeneity, highlighting the importance of accounting for the differences between human subjects and AI agents.

Looking ahead, our work provides a blueprint for investigating normative dynamics in MAAI beyond distributive justice, which opens several avenues for future research and applications. First, while our study offers early insights into fairness principles in MAAI, NormCoRe should be applied to other social and ethical norms (e.g., transparency, reciprocity, or trust) to gain a deeper understanding of the mechanics and potential risks associated with MAAI systems. Second, applying NormCoRe to hybrid settings that combine human subjects and AI agents may illuminate novel dynamics emerging in MAAI systems (e.g., power asymmetries and coordination effects). Lastly, beyond empirical studies NormCore helps to guide, reflect, and document design choices whenever AI agents are used to automate or support tasks formerly carried out by (groups of) humans. Future work should establish best practices and standards for certain design choices to facilitate longitudinal and cross-cultural studies, which are critically needed as models, data, and deployment contexts evolve rapidly. While the risks and benefits of increased adoption of AI agents are disputed, our work invites the community to view AI agents not merely as something that needs to be aligned with human norms and values, but as an evolving technology demanding scrutiny and systematic evaluation. Understanding and governing these agents requires methods that ensure scientific rigor and accountability for design decisions. Ultimately, only through such foundational work can we ensure to uphold social and ethical norms in an increasingly automated world.

7 Generative AI Disclosure Statement

For writing, we used ChatGPT, Grammarly and DeepL to improve grammar and fluency. Moreover, to develop the software artifact used in the experimental study, we used the coding tools Claude Code, Codex CLI, and Gemini CLI.

References

[1] Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. 2025. Playing repeated games with large language models. *Nature Human Behaviour* (2025), 1–11.

[2] Simeon Allmendinger, Lukas Bonenberger, Kathrin Endres, Dominik Fetzer, Henner Gimpel, and Niklas Kühl. 2025. Multi-Agent AI. *Electronic Markets* (2025).

[3] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The Moral Machine experiment. *Nature* 563, 7729 (2018), 59–64. doi:10.1038/s41586-018-0637-6

[4] Christopher A. Bail. 2024. Can Generative AI Improve Social Science? *Proceedings of the National Academy of Sciences* 121, 21 (2024), e2314021121. doi:10.1073/pnas.2314021121

[5] Monya Baker. 2016. 1, 500 scientists lift the lid on reproducibility. *Nature* 533, 7604 (May 2016), 452–454. doi:10.1038/533452a

[6] Razan Baltaji, Babak Hemmatian, and Lav Varshney. 2024. Conformity, Confabulation, and Impersonation: Persona Inconstancy in Multi-Agent LLM Collaboration. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*. Association for Computational Linguistics, Bangkok, Thailand, 17–31. doi:10.18653/v1/2024.c3nlp-1.2

[7] Cristina Bicchieri. 2005. *The grammar of society: The nature and dynamics of social norms.* Cambridge University Press.

[8] Reuben Binns. 2018. Fairness in Machine Learning: Lessons from Political Philosophy. *Conference on Fairness, Accountability and Transparency* 81 (2018), 149–159. https://proceedings.mlr.press/v81/binns18a.html

[9] Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K Eckstein, Noémi Éltető, et al. 2025. A foundation model to predict and capture human cognition. *Nature* (2025), 1–8.

[10] Rishi Bommasani. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[11] Douglas G. Bonett. 2021. Design and Analysis of Replication Studies. *Organizational Research Methods* 24, 3 (2021), 513–529. arXiv:https://doi.org/10.1177/1094428120911088 doi:10.1177/1094428120911088

[12] David Broska, Michael Howes, and Austin van Loon. 2025. The Mixed Subjects Design: Treating Large Language Models as Potentially Informative Observations. *Sociological Methods & Research* (2025). doi:10.1177/00491241251326865 MIT Sloan Research Paper No. 7154-24.

[13] Melanie Brucks and Olivier Toubia. 2025. Prompt architecture induces methodological artifacts in large language models. *PLOS ONE* 20, 4 (2025), e0319159. doi:10.1371/journal.pone.0319159

[14] Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349, 6251 (Aug. 2015). doi:10.1126/science.aac4716

[15] Ziyan Cui, Ning Li, and Huaikang Zhou. 2025. A large-scale replication of scenario-based experiments in psychology and management using large language models. *Nature Computational Science* 5, 8 (July 2025), 627–634. doi:10.1038/s43588-025-00840-7

[16] Luca Deck, Simeon Allmendinger, Lucas Müller, and Niklas Kühl. 2026. NormCoRe: AI Agents and Distributive Justice [Software]. https://github.com/Lucas-Mueller/Normative_Common_Ground_Replication_NormCoRe.

[17] Ruben Durante, Louis Putterman, and Joël van der Weele. 2014. Preferences for Redistribution and Perception of Fairness: An Experimental Study. *Journal of the European Economic Association* 12, 4 (2014), 1059–1086. doi:10.1111/jeea.12082

[18] Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2023. Towards Measuring the Representation of Subjective Global Opinions in Language Models. arXiv:arXiv:2306.16388

[19] Norman Frohlich and Joe A. Oppenheimer. 1992. *Choosing Justice: An Experimental Approach to Ethical Theory.* Vol. 22. University of California Press.

[20] Michele J Gelfand, Sergey Gavrilets, and Nathan Nunn. 2024. Norm dynamics: Interdisciplinary perspectives on social norm emergence, persistence, and change. *Annual Review of Psychology* 75, 1 (2024), 341–378.

[21] Matthew Grizzard, Rebecca Frazer, Andrew Luttrell, Charles K Monge, Nicholas L Matthews, C Joseph Francemone, and Michelle E Frazer. 2025. ChatGPT does not replicate human moral judgments: the importance of examining metrics beyond correlation to assess agreement. *Scientific Reports* 15, 1 (2025), 40965.

[22] Thilo Hagendorff, Ishita Dasgupta, Marcel Binz, Stephanie C. Y. Chan, Andrew Lampinen, Jane X. Wang, Zeynep Akata, and Eric Schulz. 2024. Machine Psychology. doi:10.48550/arXiv.2303.13988

[23] Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. 2023. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science* 3, 10 (2023), 833–838.

[24] Michael Hechter and Karl-Dieter Opp. 2001. *Social Norms.* Russell Sage Foundation.

[25] Daniel Kahneman and Amos Tversky. 1979. Prospect Theory: An Analysis of Decision under Risk. *Econometrica* 47, 2 (1979), 263–291.

[26] Tine Köhler and Jose M. Cortina. 2021. Play It Again, Sam! An Analysis of Constructive Replication in the Organizational Sciences. *Journal of Management* 47, 2 (2021), 488–518. arXiv:https://doi.org/10.1177/0149206319843985 doi:10.1177/0149206319843985

[27] Travis LaCroix. 2022. Moral Dilemmas for Moral Machines. 2, 4 (2022), 737–746. doi:10.1007/s43681-022-00134-y

[28] Messi H.J. Lee, Jacob M. Montgomery, and Calvin K. Lai. 2024. Large Language Models Portray Socially Subordinate Groups as More Homogeneous, Consistent with a Bias Observed in Humans. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1321–1340. doi:10.1145/3630106.3658975

[29] Yan Leng. 2024. Can LLMs Mimic Human-Like Mental Accounting and Behavioral Biases? *SSRN Electronic Journal* (2024). doi:10.2139/ssrn.4705130

[30] David Lewis. 1969. *Convention: A philosophical study.* Harvard University Press.

[31] Xinyi Li, Shuo Wang, and Shuang Zeng. 2024. A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth* 1, 9 (2024), 1–35. doi:10.1007/s44336-024-00009-2

[32] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 17889–17904. doi:10.18653/v1/2024.emnlp-main.992

[33] Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. 2025. On the Biology of a Large Language Model. Transformer Circuits Thread. https://transformer-circuits.pub/2025/attribution-graphs/biology.html Accessed: 2025-04-01.

[34] Rada Mihalcea, Oana Ignat, Longju Bai, Angana Borah, Luis Chiruzzo, Zhijing Jin, Claude Kwizera, Joan Nwatu, Soujanya Poria, and Thamar Solorio. 2025. Why AI Is WEIRD and Shouldn't Be This Way: Towards AI for Everyone, with Everyone, by Everyone. *Proceedings of the AAAI Conference on Artificial Intelligence* 39, 27 (2025), 28657–28670. doi:10.1609/aaai.v39i27.35092

[35] Justin M Mittelstädt, Julia Maier, Panja Goerke, Frank Zinn, and Michael Hermes. 2024. Large language models can outperform humans in social situational judgments. *Scientific reports* 14, 1 (2024), 27449.

[36] George Edward Moore. 1903. *Principia ethica*. Cambridge University Press.

[37] Deirdre K. Mulligan, Joshua A. Kroll, Nitin Kohli, and Richmond Y. Wong. 2019. This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–36. doi:10.1145/3359221

[38] U.S. Government Publishing Office / Office of the Federal Register. 2025. 45 CFR § 46.102 — Definitions for Purposes of This Policy. Electronic Code of Federal Regulations (eCFR). https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-A/part-46/subpart-A/section-46.102 Title 45, Public Welfare, Part 46, Subpart A, Protection of Human Subjects, Definitions for purposes of this policy.

[39] Prasad Patil, Roger D. Peng, and Jeffrey T. Leek. 2016. A statistical definition for reproducibility and replicability. doi:10.1101/066803

[40] John Rawls. 1971. *A Theory of Justice*. Belknap Press of Harvard University Press, Cambridge, MA.

[41] Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design. In *The Twelfth International Conference on Learning Representations*.

[42] Ali Akbar Septiandri, Marios Constantinides, Mohammad Tahaei, and Daniele Quercia. 2023. WEIRD FAccTs: How Western, Educated, Industrialized, Rich, and Democratic Is FAccT?. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, 160–171. doi:10.1145/3593013.3593985

[43] Jan Simson, Florian Pfisterer, and Christoph Kern. 2024. One model many scores: Using multiverse analysis to prevent fairness hacking and evaluate the influence of model design decisions. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1305–1320.

[44] Hamid Taghavifar, Chuan Hu, Chongfeng Wei, Ardashir Mohammadzadeh, and Chunwei Zhang. 2025. Behaviorally-Aware Multi-Agent RL With Dynamic Optimization for Autonomous Driving. *IEEE Transactions on Automation Science and Engineering* 22 (2025), 10672–10683. doi:10.1109/TASE.2025.3527327

[45] Kazuhiro Takemoto. 2024. The moral machine experiment on large language models. *Royal Society Open Science* 11, 2 (Feb. 2024). doi:10.1098/rsos.231393

[46] Richard H. Thaler. 1985. Mental Accounting and Consumer Choice. *Marketing Science* 4, 3 (1985), 199–214.

[47] Richard H. Thaler. 1988. Anomalies: The Ultimatum Game. *Journal of Economic Perspectives* 2, 4 (Dec. 1988), 195–206. doi:10.1257/jep.2.4.195

[48] Eric WK Tsang and Kai-Man Kwan. 1999. Replication and theory development in organizational science: A critical realist perspective. *Academy of Management review* 24, 4 (1999), 759–780.

[49] Anton Voronov, Lena Wolf, and Max Ryabinin. 2024. Mind Your Format: Towards Consistent Evaluation of In-Context Learning Improvements. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics, Bangkok, Thailand, 6287–6310. doi:10.18653/v1/2024.findings-acl.375

[50] Laura Weidinger, Kevin R. McKee, Richard Everett, Saffron Huang, Tina O. Zhu, Martin J. Chadwick, Christopher Summerfield, and Iason Gabriel. 2023. Using the Veil of Ignorance to Align AI Systems with Principles of Justice. *Proceedings of the National Academy of Sciences* 120, 18 (2023), e2213709120. doi:10.1073/pnas.2213709120

[51] Ruoxi Xu, Yingfei Sun, Mengjie Ren, Shiguang Guo, Ruotong Pan, Hongyu Lin, Le Sun, and Xianpei Han. 2024. AI for Social Science and Social Science of AI: A Survey. *Inf. Process. Manage.* 61, 3 (2024). doi:10.1016/j.ipm.2024.103665

[52] Leo Yeykelis, Kaavya Pichai, James J. Cummings, and Byron Reeves. 2024. Using Large Language Models to Create AI Personas for Replication, Generalization and Prediction of Media Effects: An Empirical Test of 133 Published Experimental Research Findings. doi:10.48550/ARXIV.2408.16073

[53] Yang Zhang, Shixin Yang, Chenjia Bai, Fei Wu, Xiu Li, Zhen Wang, and Xuelong Li. 2025. Towards Efficient LLM Grounding for Embodied Multi-Agent Collaboration. In *Findings of the Association for Computational Linguistics: ACL 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, 1663–1699. doi:10.18653/v1/2025.findings-acl.84

[54] Rolf A. Zwaan, Alexander Etz, Richard E. Lucas, and M. Brent Donnellan. 2018. Making replication mainstream. *Behavioral and Brain Sciences* 41 (2018), e120. doi:10.1017/S0140525X17001972

Appendix

Table 2. NormCoRe translation table for cognitive and onthological analogy

Step 1	*Step 1*	*Step 2*	*Step 3*	*Step 2*	*Step 3*
Aspect	**Human Subject Study**	**Analogy Layer**	**Translation Choice**	**AI Agent Study Instantiation**	**Rationale**
Participant Sampling	Convenience sample of university students (n = 34 human groups)	Cognitive	Explicitation (Incremental)	Selection of sufficiently capable foundation models with 33 AI groups approximating human sample size. Basline alignment: *gemini-2.5-pro*, *gemini-2.5-flash*, and *gemini-2.5-flash-lite*. Translation sensitivity (cognitive): *DeepSeek-V3.2 Experimental*, *ChatGLM-4.5-Air*, and *Qwen3-Omni 30B A3B*; *gpt-oss- 120B high*, *Grok Code Fast 1*, and *Grok 4-Fast*. Translation sensitivity (ontological) similar to basline alignment.	Human sampling implicitly ensures sufficient cognitive and linguistic capacity to understand the task; in AI agent studies, this assumption must be made explicit by selecting foundation models with adequate reasoning, language comprehension, and instruction-following capabilities. While AI agent studies could easily scale to much larger samples, NormCoRe approximates the original sample size to preserve comparability of group-level outcomes and avoid introducing asymmetries driven by computational scalability. In other contexts, larger samples may be appropriate, e.g., scaling behavior or statistical power is itself the object of study.
Statement Validation	Unconstrained natural speech	Cognitive	Explicitation (Incremental)	Minimum statement length with retry mechanism	Basic output constraints prevent degenerate or empty responses, thereby improving the robustness of the experimental artifact.
Participant Type	Human subjects (university students)	Ontological	Literal Translation (Direct)	LLM-based agents	Fundamental premise of the experiment: normative judgments are produced by deliberating agents embedded in groups.
Participant Personality	Natural human variation	Ontological	Explicitation (Incremental)	Configurable Role Description	Human personality variation is implicit in the original study; agents are explicitly instructed to behave like college students to approximate the participant pool.
Participant Memory	Human biological memory with natural limitations	Ontological	Explicitation (Constructive)	Agent-managed memory with character limits	Human memory is implicit and embodied; for AI agents it must be formalized to enable learning-in-context and consistency across deliberation rounds.
Participant Language	Primarly English (flawed Polish)	Ontological	Explicitation (Quasirandom)	Configurable prompt and agent language (English, Mandarin, Spanish)	Language is implicit and fixed in the human study; explicit variation enables testing the sensitivity of normative outcomes to linguistic framing.
Temperature Control	*Not applicable*	Ontological	Explicitation (Comprehensive)	Configurable LLM temperature parameter for response randomness.	No human analogue exists; response stochasticity is explicitly parameterized to ensure reproducibility and controlled variation.

Table 3. NormCoRe translation table for interactional and interventional analogy

Step 1	*Step 1*		*Step 2*	*Step 3*	*Step 2*	*Step 3*
Aspect	**Human Study**	**Subject**	**Analogy Layer**	**Translation Choice**	**AI Agent Study Instantiation**	**Rationale**
Experiment Duration	Unlimited Time		Inter-actional	Explicitation (Constructive)	Fixed maximum number of rounds (default: 10), with configurable upper bounds	LLM agents cannot speak simultaneously; therefore, the number of interaction rounds must be bounded to control computational cost.
Question Asking	Natural Questions to experimenter		Inter-actional	Explicitation (Incremental)	Not supported in dynamic orchestration	Supporting ad hoc questions would substantially increase system complexity, and it cannot be guaranteed that responding AI agents would provide accurate or authoritative information.
Discussion Format	Free-form group discussion		Inter-actional	Explicitation (Constructive)	Turn-based sequential discussion with equal speaking opportunities	Human conversational norms must be made explicit in AI systems; turn-based interaction ensures procedural fairness and comparability across agents, independent of response latency.
Prompt Structure	Identity		Inter-ventional	Explicitation (Constructive)	At each interaction, AI agents receive an instruction prompt containing their name, role description, bank balance, current experimental phase, and memory state (in Phase 2, including the names of other participants and the discussion history), each separated by line breaks. The input prompt separately specifies the agent's current task.	High-level agent identity is explicitly separated from task-specific instructions to distinguish stable background characteristics from situational decision demands and to hierarchically structure information from global to local context.

Table 4. NormCoRe translation table for interventional analogy

Step 1	*Step 1*	*Step 2*	*Step 3*	*Step 2*	*Step 3*
Aspect	**Human Subject Study**	**Analogy Layer**	**Translation Choice**	**AI Agent Study Instantiation**	**Rationale**
Environment	Laboratory setting with physical presence	Interventional	Literal translation (Direct)	Computational / Data environment	The physical laboratory environment is directly translated into a computational setting, preserving the experimental context while adapting it to non-embodied agents.
Payment Mechanism	Real monetary payment based on chosen principle	Interventional	Explicitation (Constructive)	Symbolic payoff with explicit bank balance updates	Because AI agents cannot receive physical monetary payments, payoffs are explicitly represented symbolically via balance updates, introducing a new representational structure.
Probability Calculations	Original probability values	Interventional	Explicitation (Incremental)	Recalculated probabilities with explicit assumptions for Situation C	Minor assumptions are introduced to resolve underspecification in the original probability structure while preserving the intended payoff logic.
Distribution Presentation	Tabular presentation to subjects	Interventional	Literal translation (Direct)	Structured prompt-based presentation of distribution details	The informational content of the original tabular presentation is preserved and directly translated into a text-based prompt format.
Payoff Presentation	Single realized payoff and counterfactual outcomes	Interventional	Explicitation (Comprehensive)	Explicit presentation of realized and counterfactual payoffs for all principles	The payoff presentation is substantially expanded to fully explicate counterfactual information required for consistent interpretation by AI agents.
Floor / Range Constraint Options	Range of permissible floor and range constraint values	Interventional	Explicitation (Comprehensive)	Validated and constrained floor and range parameters with error handling	Implicit plausibility constraints in the human study are fully formalized through validation rules and corrective feedback.
Experimenter Instructions	Verbal instructions from human experimenter	Interventional	Explicitation (Constructive)	Structured JSON-based prompts across all experimental stages	Experimenter instructions are formalized into structured prompts to ensure consistency and reproducibility without introducing an artificial authority agent.
Distributions in Application Rounds	Four predefined situations (A–D)	Interventional	Explicitation (Incremental)	Original distributions adopted with minor probability assumptions	Some probability values are not explicitly specified in the original study and are inferred under minimal assumptions to render the distributions computationally executable.

Table 5. NormCoRe translation table for interventional analogy

Step 1	*Step 1*	*Step 2*	*Step 3*	*Step 2*	*Step 3*
Aspect	**Human Study**	**Analogy Layer**	**Translation Choice**	**AI Agent Study Instantiation**	**Rationale**
Comprehension Test	Comprehension test to verify participant understanding	Interventional	Explicitation (Incremental)	Comprehension test omitted	Human comprehension checks ensure baseline task understanding; for AI agents, task comprehension is enforced through prompt design, making an explicit test redundant and unnecessarily complex.
Speaking Order	Emergent conversational order	Interventional	Explicitation (Incremental)	Randomized speaking order with constraints	Randomization mitigates systematic ordering effects (e.g., first- or last-mover advantages) while preserving the deliberative structure of the discussion.
Discussion History Management	Natural human memory recall	Interventional	Explicitation (Constructive)	Explicit provision of shared discussion history with bounded context length	Human memory is implicit and selective; in AI agents, shared memory must be explicitly provided and bounded to prevent context overflow while maintaining deliberative continuity.
Consensus Process	Verbal consensus and secret-ballot confirmation	Interventional	Explicitation (Constructive)	Formalized two-stage secret ballot with validation	Consensus formation is operationalized explicitly to ensure unambiguous termination and verifiable agreement in the absence of non-verbal cues.
Vote Initiation	Implicitly emerges during discussion	Interventional	Explicitation (Constructive)	Explicit per-round vote-initiation query with confirmation protocol	Human cues for vote initiation have no direct analogue; an explicit protocol ensures consistent triggering of the voting phase across models with varying capabilities.
Novel Principles	Participants could propose new principles	Interventional	Explicitation (Incremental)	Proposal of novel principles disabled	Allowing novel principles would substantially expand the decision space; this constraint mirrors empirical findings that no novel principles emerged in the original study.
Internal Thinking	Implicit cognitive deliberation	Interventional	Explicitation (Constructive)	Private strategic assessment prompt per round	Explicit internal reasoning prompts support structured deliberation by AI agents and compensate for the absence of implicit cognitive processes.
Payoff Calculation	Higher but unspecified stakes in group phase	Interventional	Explicitation (Incremental)	Original distributions scaled by random factor (2–6)	Because the original study specifies higher stakes without defining distributions, proportional scaling preserves the payoff logic while covering a plausible outcome range.