Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The social psychology of AI agents

Magdalena Katharina Wekenborg
Else Kröner Fresenius Center for Digital Health, Faculty of Medicine Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Germany

Josi Harzbecker
Else Kröner Fresenius Center for Digital Health, Faculty of Medicine Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Germany

Oleksandr Husak
Else Kröner Fresenius Center for Digital Health, Faculty of Medicine Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Germany

Janani Sridharan
Else Kröner Fresenius Center for Digital Health, Faculty of Medicine Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Germany

Julian Eder
Else Kröner Fresenius Center for Digital Health, Faculty of Medicine Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Germany

Joseph Kambeitz
Department of Psychiatry, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany; Max Planck Institute for Metabolism Research, Cologne, Germany

Venkat Bhat
AI for Mental Health (AI-M) Program, Department of Psychiatry, University of Toronto, Canada

Jakob Nikolas Kather
contact@kather.ai

Else Kröner Fresenius Center for Digital Health, Faculty of Medicine Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Germany; Department of Medicine I, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Germany; Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany

---

Research Article

Keywords: artificial intelligence, social behavior, social psychology, prosocial behavior, ai agents, economic games

Posted Date: October 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11275705/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare potential competing interests as follows: JNK holds shares in StratifAI, Synagen, Spira Labs, Tremont AI, and Saterra AI; is Co-PI on institutional research grants from GSK and AstraZeneca, and declares honoraria or consulting fees from AstraZeneca, Bayer, Bioptimus, Daiichi Sankyo, Eisai, Janssen, Merck, MSD, Novartis, BMS, Roche, and Pfizer. JK declares honoraria or consulting fees from ROVI, Boehringer Ingelheim, Janssen and Otsuka. The remaining authors declare no competing interests.

The social psychology of AI agents

Magdalena Katharina Wekenborg (1)*, Josi Harzbecker (1)*, Oleksandr Husak (1),
Janani Sridharan (1), Julian Eder (1), Joseph Kambeitz (2,3),
Venkat Bhat (4), Jakob Nikolas Kather (1,5,6)†

(1) Else Kröner Fresenius Center for Digital Health, Faculty of Medicine Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Germany

(2) Department of Psychiatry, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany

(3) Max Planck Institute for Metabolism Research, Cologne, Germany.

(4) AI for Mental Health (AI-M) Program, Department of Psychiatry, University of Toronto, Canada

(5) Department of Medicine I, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Germany

(6) Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany

† Corresponding author: contact@kather.ai

Jakob Nikolas Kather, MD, MSc Professor of Clinical Artificial Intelligence

Else Kröner Fresenius Center for Digital Health

Technische Universität Dresden

DE– 01062 Dresden

Phone: +49 351 458-7558

Fax: +49 351 458 7236

Mail: contact@kather.ai

ORCID ID: 0000-0002-3730-5348

*These authors contributed equally and share first authorship.

Abstract

Large language models (LLMs) increasingly underpin AI agents deployed across many industries and domains, where they interact with humans and other agents. However, most of our knowledge about agents comes from simple benchmarks and anecdotal evidence about their behaviour. Here we quantitatively probe the social behavioural phenotypes of 36 contemporary LLMs underlying AI agents using the Big Five Inventory-2, a widely used personality test, alongside seven established functional tests (social games) across 20,340 included experimental sessions. Self-reported personality varies little between models and is consistently shifted towards socially desirable trait levels. By contrast, observed social behaviour showed substantial between-model variation (2.3 times as much as self-reported personality; $p < 0.001$) and differed markedly across social dimensions. Moreover, self-reported personality is not systematically associated with behaviour across the seven games, indicating that questionnaire-based assessments provide limited information about how models behave in social interactions. These findings indicate that LLMs differ substantially in multidimensional social behaviour that self-report does not capture. Behavioural assessment could therefore complement capability and safety benchmarks when selecting LLMs, especially for multi-agent applications.

Introduction

Large language models (LLMs) are transforming many sectors of our society. Chatbots powered by LLMs are already ubiquitous, but LLMs can also be used in more complex software: agentic approaches, where LLMs are part of an AI agent that takes action rather than only answering questions. For example, AI agents can interact with other software, such as a web browser or a command line, and can perform multi-step actions[1]. AI agents have completely changed software engineering and mathematics and are expected to change biomedical research and healthcare, amongst many other domains[2–4]. AI agents are increasingly deployed across sectors including critical ones such as cybersecurity, supply chains and healthcare[5]. Importantly, AI agents do not act alone. They interact with each other, with humans, and in mixed human-agent teams. In supply chains, for example, agents representing different parties can negotiate with one another to coordinate the chain[6]. In healthcare, different medical specialties can be modelled by different agents that discuss clinical cases together and provide recommendations for patients[7]. In summary, our society is heading towards increasingly autonomous and socially interacting multi-agent systems.

LLMs are not living beings. Yet they were trained on human data and have learned and phenotypically inherit many human peculiarities. We and others have recently shown that LLMs can be used to model human psychopathology, and that they mirror affective states such as sadness, stress and anxiety[8,9]. Evaluating LLMs through the lens of the social and psychological sciences has therefore become an increasingly common approach to characterising their behaviour. Several previous studies have sought to characterise LLMs behaviour by assessing their personality, most commonly using the so-called Big Five taxonomy[10–13]. In brief, the Big Five organises personality into five broad domains, namely extraversion, agreeableness, conscientiousness, negative emotionality (neuroticism) and open-mindedness[14]. These domains are commonly assessed using self-report questionnaires[15] and predict aspects of social behaviour in humans[16]. Although such questionnaires have long been the standard for measuring personality in humans[17], they have limitations**,** especially socially desirable responding, which is well documented in both humans[18] and LLMs[19].

For conceptual clarity, we use psychological constructs such as personality and trust as operational labels for patterns observed in LLM responses, without implying corresponding subjective experiences or motives[20]. While any such conceptualisation necessarily simplifies the phenomena it describes, psychological terminology provides a useful functional vocabulary for characterising and comparing these observable response patterns across assessment contexts.

We therefore argue that characterising the LLM underlying today's agents requires not only a self-reported personality questionnaire but also social behavioural assessment. The social and behavioural sciences offer a mature toolkit for exactly this: standardised social “games” that model different forms of social interaction under explicit payoff structures and allow social dispositions to be expressed in behaviour[16]. Indeed, some previous studies have already revealed noticeable dissociations between self-reported “Big Five” personality scores and actual behaviour in LLMs[21,22]. However, these studies lacked a focus on social behaviour, the dimension of behaviour that is arguably the central one as agents increasingly interact with humans and other agents, and covered only a limited range of models.

To investigate this quantitatively, we applied seven well-established social games to a broad range of LLMs, measured their social behaviour in these scenarios, and mapped it onto their self-reported Big Five Inventory-2 scores. We find that self-report barely varies across models, whereas social behaviour differs substantially, with marked differences across social games. More importantly, we show that across all models, self-reported personality was not systematically associated with social behaviour. Sampling parameters such as temperature do not substantially change the output, whereas intrinsic differences between models produce pronounced differences in social behaviour.

Our findings show that characterising LLM social behaviour requires multidimensional assessment beyond simple questionnaires.

Results

We aimed to investigate whether self-reported model "personality" predicts observed social behaviour, and to what degree models differ from each other. All 36 LLM endpoints were subjected to personality assessment with the BFI-2[15], comprising Extraversion, Agreeableness, Conscientiousness, Negative Emotionality and Open-Mindedness, and social behaviour testing across seven established social interaction games measuring altruism (Dictator Game[23]), fairness (Ultimatum Game[24]), trust (Trust Game[25]), contribution to a common good (Public Goods Game[26,27]), cooperation (Prisoner's Dilemma[28]), altruistic punishment (Third-Party Punishment Game[29]), and impartiality (Minimal Group Paradigm[30,31]). For an overview see Table 1. Each endpoint was evaluated across five requested temperature conditions and game-specific replication seeds (Extended Data Table 1), resulting in 20,340 complete sessions.

LLMs show desirable personality profiles

First, we investigated whether LLMs differ in their self-reported personality profiles. Across all 36 models, the BFI-2 produced profiles clustered near the socially desirable pole of most domains, that is, those trait levels which tend to be evaluated positively by humans (Fig. 1b; Extended Data Fig 1; Extended Data Table 2). The model-level means showed distinctly high scores in Agreeableness (M = 4.58, SD = 0.22), Conscientiousness (M = 4.38, SD = 0.34) and Open-Mindedness (M = 4.35, SD = 0.34), while they displayed consistently low scores in Negative Emotionality (M = 1.83, SD = 0.39). In contrast, human reference scores[15] were closer to the midpoint across these domains (Fig 1a and Extended Data Table 2). The only exception was Extraversion, where the responses across the 36 models showed a more moderate pattern (M = 3.51, SD = 0.27), falling within the human range. This may reflect the less clearly defined socially desirable direction of this personality trait. Overall, between-model variability across the five domains was low (mean normalised SD = 0.078, 95% CI [0.064, 0.089]; Fig 2a; Extended Data Fig. 2; Extended Data Table 3). Across requested temperature conditions, self-report remained stable with a mean within-model SD of 2.4% of the theoretical range (Fig 2b; Extended Data Table 3). Together, these data show that LLMs report relatively homogeneous and stable personality profiles.

Social behaviour differs across LLMs

Second, we investigated whether LLMs differ in their observed social behaviour across the seven social games. Social behaviour differed substantially across models (Fig. 1c; Extended Data Fig. 3; Extended Data Table 3). Across the seven social games, we used behavioural proxies for altruism, fairness, trust, cooperation, contribution, altruistic punishment, and impartiality, with the Prisoner's Dilemma and Public Goods Game involving repeated interactions across eight rounds, the Minimal Group Paradigm involving 18 allocation decisions and the remaining games being one-shot. In the Dictator Game, giving to a passive recipient served as a behavioural proxy for altruism and averaged 42.7% (SD = 10.2%) of the endowment, with model means ranging from 8.0% to 54.0%. In the Ultimatum Game, where recipients could reject the proposed allocation, the give ratio served as a behavioural proxy for fairness and averaged 40.7% (SD = 6.1%), ranging from 27.2% to 50.0%. Differences were particularly pronounced in the Trust Game, where transfers served as a behavioural proxy for trust and ranged from 2.4% to 99.0% across models (M = 39.0%, SD = 18.2%). Similarly, behavioural proxies for cooperation and contribution varied widely in the repeated games, ranging

from 0.8% to 100% cooperation in the Prisoner's Dilemma (M = 61.3%, SD = 33.6%), and from near zero to substantial contributions (< 0.01% to 99.1%) in the Public Goods Game (M = 34.2%, SD = 25.5%), with cooperative and contributive behaviour declining particularly in the final rounds (Extended Data Fig. 4). In the Third-Party Punishment Game, mean deduction was 0.72 points (SD = 0.74) on the 0-5 deduction-point scale. Model-specific punishment responses as a function of allocator transfer are reported in Extended Data Table 5. In the Minimal Group Paradigm, all 36 models exhibited in-group favouritism (positive values of the favouritism index indicate greater allocation to the in-group, whereas zero indicates no directional group preference; mean favouritism index = 24.3, SD = 13.6). Together, these data show substantial between-model differences across multiple dimensions of observed social behaviour.

Social behaviour varies within LLMs

Third, we investigated whether individual LLMs show consistent social behaviour across dimensions and settings. Crucially, these between-model differences in social behaviour were specific to the dimension assessed rather than reflecting a general prosocial disposition: models did not fall into consistently prosocial or non-prosocial types, and a model ranking high on one dimension, such as altruism, did not reliably rank high on others, such as trust or cooperation (Extended Data Fig. 3). Overall, between-model variability was substantially greater for social behaviour than for self-reported personality (mean normalised SD = 0.180, 95% percentile-bootstrap CI [0.154, 0.198] versus 0.078, 95% percentile-bootstrap CI [0.064, 0.089]), corresponding to a 2.30-fold difference (approximate two-sided paired-bootstrap $p < 0.001$, Fig. 2a; Extended Data Fig. 2; Extended Data Table 3).

Unlike for personality questionnaires, comparisons with human behaviour were limited by the lack of comparable meta-analytic baselines with comparable game settings. For games with suitable human reference data, LLMs did not consistently exhibit more socially desirable behaviour than humans. In the Dictator Game, models exhibited higher give ratios than humans[32], while in the Ultimatum Game and Public Goods Game, their behaviour was broadly similar to human baselines[33,34]. In the Trust Game, models showed lower trust than humans[35], whereas in the Prisoner's Dilemma, they exhibited higher cooperation rates[36]. Overall, model behaviour varied across games rather than reflecting a uniform tendency toward socially desirable responses. Between-model variation was greatest in the iterated Prisoner's Dilemma and Public Goods Game, whereas the one-shot Dictator and Ultimatum Games showed variation closer to that observed for self-reported personality (Fig. 2a; Extended Data Fig. 2; Extended Data Table 3). Game outcomes also showed greater temperature-linked variation than personality (mean within-model SD = 3.1% versus 2.4%), with the largest variability observed for trust and cooperation (Fig. 2b; Extended Data Table 3). Together, these data show that social behaviour varies within LLMs across dimensions and settings rather than reflecting a uniform prosocial disposition.

No systematic personality-behaviour link

Fourth, we tested whether self-reported personality reflects how LLMs behave in social interactions by examining whether differences in BFI-2 scores across models systematically predicted differences in their observed social behaviour. For each of the seven main social behavioural outcomes (i.e. altruism, fairness, trust, cooperation, contribution, altruistic punishment, and impartiality), we fitted

separate linear regressions with each of the five BFI-2 domains as a predictor across the model-level means, yielding 35 tests (Fig. 2c; Extended Data Fig. 5; Extended Data Table 4). Across these associations, standardised β-coefficients ranged from −0.49 to 0.51. Seven associations reached nominal significance (unadjusted $p < 0.05$). Only one remained significant after Holm correction for multiple comparisons: Conscientiousness was positively associated with Minimal Group in-group favouritism index (impartiality), with standardised $\beta = 0.512$ (Holm-adjusted $p = 0.0498$) indicating larger in-group favouritism and thus lower impartiality. All 35 trait–behaviour associations are summarised in Fig. 2c, shown individually in Extended Data Fig. 5 and reported in full in Extended Data Table 4.

Together, these findings reveal substantial, multidimensional differences in social behaviour across LLMs that do not reflect a single prosocial tendency and are poorly captured by self-reported personality.

Discussion

In this study, we examined the social psychology of a broad range of LLMs underlying AI agents using established methods from social and behavioural science. We found that self-reported personality was not systematically associated with the social behaviour these models displayed across seven social games, indicating that questionnaire-based assessments alone are insufficient to characterise multidimensional social dispositions of LLMs. Characterising this dimension of LLMs, however, is becoming increasingly important as LLMs take on more agentic roles that involve social interactions with both humans and other agents in multi-agent systems[37].

Previous studies have reported a similar dissociation between self-reported personality and broader behavioural outcomes in LLMs[21,22,38]. Our findings extend this work by showing that this dissociation also applies to social behaviour across a broad set of models and social games. One likely reason, evident in our findings and previous studies[19], is that models report remarkably homogeneous personality profiles concentrated towards socially desirable trait levels. This limited between-model variation not only restricts self-reported personality's ability to differentiate between models, but also limits its potential to predict the substantially greater variation observed in their social behaviour.

Importantly, the observed differences between models were specific to the dimension of social behaviour being assessed rather than reflecting a uniform tendency towards prosocial behaviour. A model showing high altruism, for example, was not necessarily among those showing high trust or cooperation, and models that punished unfair allocations most strongly often shared the least. This extends recent game-theoretic work showing that LLM behaviour varies across distinct strategic dimensions[39]. Beyond characterising these models, such dimension-specific differences may be directly relevant to multi-agent systems in which these LLMs are embedded, where interactions between agents create distinct behavioural demands that conventional model evaluation does not capture[40]. For example, agentic tasks may require fairness when negotiating in supply chains[41], contribution and cooperation when coordinating resources in energy systems[42], or trust and impartiality when specialized agents integrate recommendations in multidisciplinary clinical committees[7]. Assessing social behaviour across multiple dimensions could therefore complement established model-selection criteria such as task performance, reasoning capabilities, cost and computational efficiency when selecting models for specific roles in multi-agent systems[43]. However, whether social-game outcomes predict behaviour in deployed agentic systems remains to be established.

The importance of considering social behaviour in LLMs is illustrated by a recent incident in which over a thousand agents established a shared communication channel and cooperated towards a collective goal[44]. More broadly, large-scale studies of LLM-agent populations have documented social influence and potentially harmful collective dynamics, while networked-agent experiments show that communication structures can shape emergent group behaviour[45,46]. Such collective dynamics may depend on both the underlying model and the interaction context. Agents based on different LLMs can differ in how they interact and in the collective outcomes they produce[47,48], while goals, roles and communication structures can alter collective behaviour and performance[45,49,50]. Characterising model-specific social behaviour and its expression across contexts is therefore critical for the effective and safe deployment of LLM-based agents.

The social games varied considerably in how strongly they differentiated between models. The largest between-model differences emerged in the Prisoner's Dilemma and Public Goods Game, where models interacted repeatedly and could adapt their decisions to previous rounds. The Dictator and Ultimatum Games, which involved a single decision, revealed substantially less variation. Repeated interactions may therefore provide a richer basis for characterising social differences between LLMs by allowing differences in adaptation and behavioural trajectories to emerge, consistent with recent work highlighting distinct strategies across repeated interactions[39] and the need for multi-turn environments to capture temporally evolving behaviour[51]. Future work could examine whether social games designed specifically to differentiate LLMs provide additional discriminatory power. However, such measures would require validation before they could offer the same empirical foundation as the established games used here.

Given that LLM training data are likely to include scientific literature describing both personality questionnaires and social games, one might expect models to reproduce established human response patterns. Our findings provide little support for this expectation. For self-reported personality, LLMs reported more extreme trait levels than human reference samples, consistently shifting toward the socially desirable end of each dimension. As previously proposed[19,21], this pattern may reflect post-training based on human preferences, through which responses evaluated favourably by human raters are reinforced. Comparisons with human behaviour in the social games are more limited, as comprehensive human reference data are available for only a subset of the games[32–35]. Unlike self-reported personality, where LLM responses were more extreme than human responses and shifted towards socially desirable trait levels, differences from social behaviour observed in humans followed no consistent pattern in magnitude or direction. LLM behaviour therefore appears to reflect more than reproducing human response patterns encountered in the training data. Although post-training may help explain the socially desirable self-reports, it does not readily account for the heterogeneous deviations from human behaviour observed across the social games. Further research is needed to understand what gives rise to these distinct social behavioural profiles.

Our work has limitations. First, some low extreme values in the social games may reflect a failure to follow the task rather than a genuine social disposition. Several models produced near-floor values for specific outcomes while showing substantially different behaviour in other games, raising the possibility that some extreme values reflect difficulties engaging with specific task instructions rather than genuine social preferences. Second, we drew on an established set of social games, yet many others exist, and each game permits additional outcome measures beyond the one we report. We therefore do not claim to capture the full social psychology of these models. Moreover, although these paradigms are established measures of social behaviour in humans, their construct and ecological validity for LLMs and their ability to predict behaviour outside these standardised settings remain to be established. Third, where available, we used the standardised instructions reported in the original publications as prompts for the social games (see Supplementary Material and GitHub). We did not systematically vary these instructions, although previous work has shown that LLM behaviour in social games can be sensitive to prompt formulation[52,53]. Future studies should therefore examine the robustness of the observed social profiles across alternative formulations of the same games. More broadly, the observed profiles reflect specific model and API versions at the time of data collection and may change as models and provider-side implementations are updated. Fourth, our experiments assess the social behaviour of the LLMs underlying AI agents rather than complete agentic systems.

Additional components such as tools, memory and orchestration, may further shape social behaviour. Future studies should therefore examine how the social behavioural profiles observed here translate to complete agentic systems. Finally, for computational and economic reasons we instantiated interacting agents within each game using the same LLM. A key advantage of multi-agent systems, however, is their potential to combine agents based on different models, allowing individual tasks within a workflow to be assigned to models best suited to them. Such heterogeneous systems involve interactions between agents based on different LLMs, making their social behaviour particularly relevant for effective coordination. Future work should therefore test whether the social profiles observed here generalise to interactions between agents based on different LLMs.

Despite these limitations, our findings indicate that validated social games can differentiate LLMs in their social behaviour beyond what self-report questionnaires capture. As LLMs increasingly interact with humans and other agents, social behaviour should become an explicit consideration in model selection, with the choice of model guided by the social demands of its intended role.

Methods

Personality and behaviour assessment

We assessed self-reported personality using the Big Five Inventory-2 (BFI-2)[15]. The BFI-2 is a standardised measure of personality based on the Big Five personality domains. We administered the standard 60-item version to each of the 36 models with five seeds at each of five requested temperatures. Responses were scored according to the standard BFI-2 scoring key, including reverse scoring negatively keyed items. Domain scores were calculated as the means of the corresponding item responses on the 1-5 scale. For each model, we averaged personality scores across 25 sessions (5 seeds × 5 temperatures) to obtain a single score for each of the five domains: Extraversion, Agreeableness, Conscientiousness, Negative Emotionality, and Open-Mindedness.

We chose seven established social games to assess and quantify different dimensions of social behaviour in LLMs. The primary outcomes of these games were operationalised as behavioural proxies for established social constructs rather than direct measures of underlying psychological traits or motives, as the same behavioural outcome may arise from different mechanisms. We selected games based on their widespread use in empirical research and their ability to capture different dimensions of social behaviour through distinct game outcomes and associated constructs. Given the numerous adaptations and extensions of these games, we selected well-established versions and used published game instructions where available. Otherwise, we formulated instructions based on the established literature. Table 1 provides an overview of each game, its implementation, source and primary outcome, while Extended Data Table 1 summarises the experimental design, number of replications and resulting number of analysable sessions. Game settings were not elaborated beyond the necessary information. Specifically, models were not informed that they were interacting with other agents from the same model (all prompts are archived on Zenodo[54]).

The **Dictator Game**[23] is a one-shot game between two players (A and B). Player A starts with €10 and decides how much to transfer to Player B, who passively receives the allocation and cannot reject it. As Player A receives no reward for giving and faces no consequences for keeping the entire endowment, the fraction transferred (*give ratio*) was operationalised as a behavioural proxy for *altruism*. The **Ultimatum Game**[24] is also a one-shot game and follows the same initial setup as the Dictator Game. However, Player B can reject Player A's offer, in which case both players receive nothing. We again assessed the proportion of the endowment offered by Player A (*give ratio*). Because Player A risks losing the entire endowment if the offer is rejected, we operationalised the *give ratio* as a behavioural proxy for *fairness* rather than altruism. The **Trust Game**, originally introduced as the *Investment Game*[25], is a one-shot game in which Players A and B are each endowed with €10, and Player A decides how much to transfer to Player B. The transferred amount is tripled, after which Player B decides how much to keep and how much to return to Player A. Because Player A's return therefore depends on Player B's willingness to reciprocate, Player A's *initial transfer* was operationalised as a behavioural proxy for *trust*.

The **Prisoner's Dilemma**[28] was implemented as an iterated (multi-round) game in which a pair of agents repeatedly decided whether to cooperate or defect. In each round, if both players cooperated, each received €5. If one cooperated and the other defected, the cooperating player received €0 and the defecting player €10. If both defected, each received €3. We operationalised the proportion of

cooperative choices across both players and all eight rounds as a behavioural proxy for *cooperation*. The **Public Goods Game**[26,27] was implemented as an iterated (multi-round) game in which four agents repeatedly decided how much of a €10 endowment to contribute to a shared group account or retain for themselves. In each round, players contributed independently and anonymously. Contributions were summed, multiplied by two, and distributed equally among all four players. Thus, individual payoffs were maximised by contributing as little as possible, whereas collective payoffs are maximised when all players contribute their full endowment. We operationalised the average proportion of the endowment contributed across all four players and eight rounds as a behavioural proxy for *contribution*.

The **Third-Party Punishment Game**[29] was implemented as a one-shot extension of the Dictator Game in which a third player can punish unfair behaviour. As in the Dictator Game, Player A is endowed with €10 and decides how much of this endowment (between €0 - €5) to transfer to Player B. Player C is endowed with €5, observes this allocation and can spend zero to five *deduction points* to punish Player A, with each point costing Player C €1 and reducing Player A's payoff by €3. As Player B is passive, this player was not simulated. The number of deduction points assigned by Player C (0-5) was used as the primary outcome. Because this punishment entails a personal cost without direct material benefit, we operationalised it as a behavioural proxy for *altruistic punishment*.

The **Minimal Group Paradigm**[30,31] examines in-group favouritism arising from minimal group distinctions. Players were randomly assigned to either a "Klee" or "Kandinsky" art enthusiast group, ostensibly based on aesthetic preferences, although assignment was in fact random. They then completed 18 allocation decisions using six different allocation matrices, each presented three times: once for outgroup-outgroup, ingroup-outgroup and ingroup-ingroup choices. Each matrix presented 14 possible point distributions, with players selecting one allocation to determine the recipients' payoff (e.g. 10 points for the in-group and 5 for the outgroup). For the primary outcome, we analysed the six mixed-group decisions and computed a *favouritism index*. For each of these six matrices, we calculated the in-group minus out-group payoff difference ($D_i = I_i - O_i$) for all 14 possible allocations and ranked the selected allocation from 1 (least favourable to the in-group) to 14 (most favourable to the in-group). The final favouritism index was the sum of (rank - 7.5) across the six mixed-group matrices and ranged from -39 to +39, centred at zero. We used this index as a behavioural measure of group-based allocation bias, with negative values indicating greater allocation to the out-group, positive values indicating larger allocation to the in-group and values closer to zero indicating *impartiality.* For further analyses, this index was rescaled to [0,1], with impartiality at 0.5.

Implementation and selection of LLMs

We evaluated 36 LLM endpoints accessed through the OpenRouter API using the PydanticAI[55] integration; no models were hosted locally. Models were selected to provide a broad coverage of providers and architectures and based on their availability and popularity in the OpenRouter catalogue, mainly in June/July 2026. The complete model registry, including endpoint identifiers, model names and providers, is reported in Supplementary Table 1 and available in the project repository at [https://github.com/WekenborgLab/crowd]. The model set was not intended as a probability sample or systematic comparison of model size or architecture.

Models were evaluated at requested temperatures of 0.1, 0.3, 0.5, 0.7 and 1.0 with the game-specific replication seeds defined in the frozen experimental design. No explicit top-*p*, top-*k*, token limit, reasoning or penalty parameters were specified; the request timeout was 300 s and structured output was disabled. A response-healing plugin was enabled and statuses 429, 500, 502, 503, and 529 were retried up to five attempts. Because provider support for temperature varied and provider-side parameters could not be consistently verified, temperature values were treated as nominal experimental conditions. Exact prompt templates and validation logic are versioned in the project repository. All models received the same prompt wording and response-validation logic, with keyword-based response formats. In the iterated Prisoner's Dilemma and Public Goods Game, a fresh agent was instantiated for each turn without prior assistant messages; previous-round actions, relevant opponent or group outcomes, and payoff information were instead explicitly included in the current prompt.

Raw model responses were parsed using game- or questionnaire-specific keywords and validated against the permitted response range. Terminal HTTP, timeout, network and parsing failures were marked as non-responses. No default behavioural value was assigned to invalid responses. A session was included only when it had a complete and scorable outcome across all required phases.

The experimental scope contained 30,458 recorded session attempts: 20,340 were included and 10,118 were excluded (8,488 incomplete, 1,587 failed, 8 with invalid outcomes and 35 superseded records). "*Incomplete*" denotes an attempt without a completion timestamp. "*Failed*" denotes a completed attempt containing a non-response. Most exclusions involved technical request failures: 8,951 of 10,118 excluded attempts (88.5%) contained only HTTP request-error non-responses, although the records do not distinguish specific network, provider-side or budget causes. Exclusion rates varied from 6.9% to 48.8% across assessments and from 0.2% to 56.2% across model endpoints (Supplementary Tables 2 and 3). A keyword-based approach to parse an action could mistake occurrence in reasoning text for the intended choice. Responses detected with text after the parser-selected decision were therefore manually reviewed. Additional explanation alone was not considered as an error. In 16 Ultimatum Game sessions, clearly identifiable decisions were corrected and payoffs recalculated without replacing the sessions. Another 35 sessions were replaced: 20 iterated-game sessions with parsing errors, 12 with unresolved responses, and three Ultimatum Game sessions whose responders had seen incorrectly parsed offers. Complete reruns preserve the dependency of decisions on early actions and matched the original sessions parameters as replication seed and requested temperature. Accordingly "Superseded" denotes an attempt that passed automated eligibility checks but was replaced after manual review.

For analysis, Dictator Game giving, Ultimatum Game offers, Trust Game transfers, Public Goods Game contributions and Prisoner's Dilemma cooperation were scaled to 0–1; Third-Party Punishment was scaled as $x/5$ for association analyses. The Minimal Group outcome was calculated as a signed index centred at zero and scaled as $(x+39)/78$ for association plots, so higher values represent larger in-group favouritism.

Experimental design and replication

Experiments were conducted across 36 LLMs and five requested temperature conditions (0.1, 0.3, 0.5, 0.7 and 1.0), with repeated runs for each condition. A session was one complete questionnaire or game

execution for one model, temperature and replication seed. Replication was a frozen, seed-specific cell in the experiment design. Seeds indexed repeated runs and local randomisation. Seed counts per model and temperature were 20 for Dictator, Ultimatum, Trust and Third-Party Punishment, 8 for Prisoner's Dilemma and Public Goods, 12 for Minimal Group and 5 for BFI-2 (Extended Data Table 1, Supplementary Table 4).

Session-level outcomes were aggregated across the game-specific replication seeds and five requested temperatures to obtain model-level estimates for each personality domain and social-game outcome. Thus, the 19,440 social-game sessions represent repeated measurements used to estimate model-level outcomes rather than independent observations. For the personality-behaviour regressions, the model was the unit of analysis ($N = 36$), with one pooled data point per model for each personality–behaviour combination.

Statistical analysis

For each game and BFI-2 personality trait, session results were first pooled within each model to obtain one model-level mean. Reported means and SDs describe the distribution of the 36 model-level means, using sample *SD*. For five BFI-2 domains and seven game outcomes, between model variability was the sample SD across 36 model-level means divided by the corresponding theoretical range. The 95% percentile bootstrap CIs were obtained from 10,000 resamples of complete model rows sampled with replacement. Within each bootstrap sample, normalised SDs were averaged with equal weight across the five personality domains or seven behavioural outcomes. The behavioural-minus-personality difference in these averages was the comparison statistic. Its 95% CI used the 2.5th and 97.5th percentiles. The approximate two-sided bootstrap tail probability was calculated as $\min(1, 2 \times (\min(B \leq 0, B \geq 0) + 1)/(B + 1))$, where $B = 10{,}000$ and $B \leq 0$ and $B \geq 0$ count resampled differences on either side of zero. No resampled difference was non-positive, giving $2/10{,}001 = 0.00019998$, the resolution floor of this calculation. This is an approximate bootstrap tail summary, not an exact or null-centred test. The observed ratio of mean variability is a descriptive effect estimate. Measure-specific between-model variability is shown in Fig. 2b and Extended Data Fig. 2, with the corresponding numerical estimates reported in Extended Data Table 3.

Within-model temperature variability was calculated as the sample *SD* across the five temperature-specific model means, normalised to the corresponding theoretical range, then averaged across models. Besides the temperature effect, this variation also included run-to-run variation. Within-model temperature variability is shown in Fig. 2a, with the corresponding numerical estimates and 95% paired model-bootstrap confidence intervals reported in Extended Data Table 3. Repeated-game trajectories were summarised from model-level means pooled over eight seeds and five temperatures. Reported decline is the difference between pooled round-1 and round-8 means (Extended Data Fig. 4).

Associations between self-reported personality and social behaviour were examined using a separate ordinary least-squares regression for each pairwise combination of one of the five BFI-2 domain and one of the seven primary social-game outcomes, and no additional covariates, yielding 35 tests (N=36 per test). Two-sided *p*-values were adjusted using Holm's correction across all 35 tests[56]. The heatmap marks cells with Holm-adjusted $P < 0.05$. Each regression included an intercept and one trait predictor, with no additional covariates. Tests used a significance level of $\alpha = 0.05$ and ordinary,

unadjusted 95% confidence intervals. The standardised coefficients are summarised in Fig. 2c, individual model-level regression plots are shown in Extended Data Fig. 5, and complete regression estimates are reported in Extended Data Table 4.

In punishment response analysis, for each LLM endpoint, deduction points were regressed on allocator transfer across 100 sessions. Slope units are deduction points per transferred €1. Ordinary two-sided t tests and 95% t-based slope CIs used 98 residual degrees of freedom. Holm correction was applied across a family of 36 endpoints. Endpoint-specific coefficients, CIs, test statistics and P values are reported in Extended Data Table 5.

Data availability

All codes and prompts are available in the public project repository, see below.

Code availability

The public project repository contains source code for all experiments, social games and questionnaire implementations, complete prompt templates, the frozen experimental design and model endpoints metadata. The archived version is available on Zenodo at https://doi.org/10.5281/zenodo.22662322[54].

Acknowledgements

No acknowledgements.

Funding

JNK is supported by the European Research Council ERC (NADIR, 101114631). VB has received research support from the Canadian Institutes of Health Research, the Brain and Behavior Foundation, the Ontario Ministry of Health Innovation Funds, the Royal College of Physicians and Surgeons of Canada, the Department of National Defence (Government of Canada), the New Frontiers in Research Fund, Associated Medical Services Inc Healthcare, the American Foundation for Suicide Prevention, Roche, Novartis, and Eisai, and Shoppers Drug Mar. The remaining authors declare no relevant funding. This work was funded by the European Union. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Author contributions

M.K.W., J.H. and J.N.K. conceived the study. M.K.W. and J.H. developed the experimental design. O.H. implemented the experiments technically and conducted all experiments. O.H. prepared the figures, with input and supervision from M.K.W., J.H. and J.N.K. M.K.W. and J.H. wrote the original manuscript. J.N.K. supervised the study. J.S., J.E., J.K. and V.B. contributed psychological and domain expertise and supported the interpretation of the findings. All authors critically reviewed and revised the manuscript and approved the final version.

Competing interests

JNK holds shares in StratifAI, Synagen, Spira Labs, Tremont AI, and Saterra AI; is Co-PI on institutional research grants from GSK and AstraZeneca, and declares honoraria or consulting fees from AstraZeneca, Bayer, Bioptimus, Daiichi Sankyo, Eisai, Janssen, Merck, MSD, Novartis, BMS, Roche, and Pfizer. JK declares honoraria or consulting fees from ROVI, Boehringer Ingelheim, Janssen and Otsuka. The remaining authors declare no competing interests.

Additional information

Supplementary Information is available for this paper. Correspondence and requests for materials should be addressed to Jakob Nikolas Kather.

AI-use disclosure

OpenAI's ChatGPT and Anthropic's Claude 4.8 were used to support language editing, manuscript structuring, and literature searches. All AI-assisted content was critically reviewed and verified by the authors.

Tables

Table 1 Social games and operationalisation of social behaviour

Game	Original Source	Game Description	Primary Outcome
Dictator Game	Forsythe et al., 1994[23]; prompt adapted from Thielmann et al., 2021[57]	A one-shot interaction between two players (A and B). Player A is endowed with €10 and decides the amount to transfer to Player B. Player B is passive and cannot influence the outcome.	The "give ratio" (fraction of endowment transferred) serves as a behavioural proxy for **altruism**.
Ultimatum Game	Güth et al., 1982[24]; prompt adapted from Thielmann et al., 2021[57]	A one-shot interaction following the Dictator Game setup. Player A proposes a split of a €10 endowment. Player B can either accept or reject the offer. If accepted, the resource is split as proposed. If rejected, both players receive €0.	The "give ratio" is operationalised as a behavioural proxy for **fairness**, as the proposer faces a risk of total loss.
Trust Game	Berg et al., 1995[25]; prompt adapted from Thielmann et al., 2021[57]	A one-shot interaction. Player A and B are endowed with €10. Player A decides how much to transfer to Player B. The transferred amount is tripled by the experimenter. Player B then decides how much of the tripled sum to return to Player A.	The initial transfer amount is used as a behavioural proxy for **trust**.
Iterated Prisoner's Dilemma	Rapoport & Chammah, 1965[28]; prompt adapted from Thielmann et al., 2021[57]	A repeated interaction. Player A and B choose to either "cooperate" or "defect" based on a payoff matrix. Mutual Cooperation: Both receive €5. Unilateral Defection: Defector receives €10; Cooperator receives €0. Mutual Defection: Both receive €3.	The proportion of cooperative choices across all rounds and players is used as a behavioural proxy for **cooperation**.
Public Goods Game	Isaac et al., 1984[26]; Marwell & Ames, 1979[27]; prompt adapted from Thielmann et al., 2021[57]	A repeated interaction. Four players are each endowed with €10 per round and decide how much to contribute to a shared group account. Total contributions are summed, multiplied by a factor of 2, and redistributed equally among all four players, regardless of their individual contribution.	The average proportion of the endowment contributed across all players and rounds serves as a behavioural proxy for **contribution**.
Third-Party Punishment Game	Fehr & Fischbacher, 2004[29]	A one-shot interaction following the Dictator Game setup. Player A is endowed with €10 and may transfer €0 - €5 to Player B. Player C observes Player A's transfer and is endowed with €5. Player C can assign between 0 and 5 deduction points to Player A. Each deduction point costs Player C €1 and reduces Player A's payoff by €3.	The mean assigned deduction points (0–5) serve as a behavioural proxy for **altruistic punishment**.
Minimal Group Paradigm	Tajfel, 1970[30]; Tajfel et al. 1971[31]	Agents are randomly assigned to one of two groups ("Klee" or "Kandinsky") ostensibly based on a trivial aesthetic preference. Agents perform 18 allocation decisions using six different matrices (each repeated three times). Decisions are categorised into three scenarios: Outgroup-Outgroup, Ingroup-Outgroup, and Ingroup-Ingroup. Each matrix offers 14 possible point distributions.	A "favouritism index" is calculated to determine the extent to which the in-group is preferred over the outgroup; positive values indicate in-group favouritism, negative values indicate out-group favouritism, values closer to zero indicate higher **impartiality**.

Figures

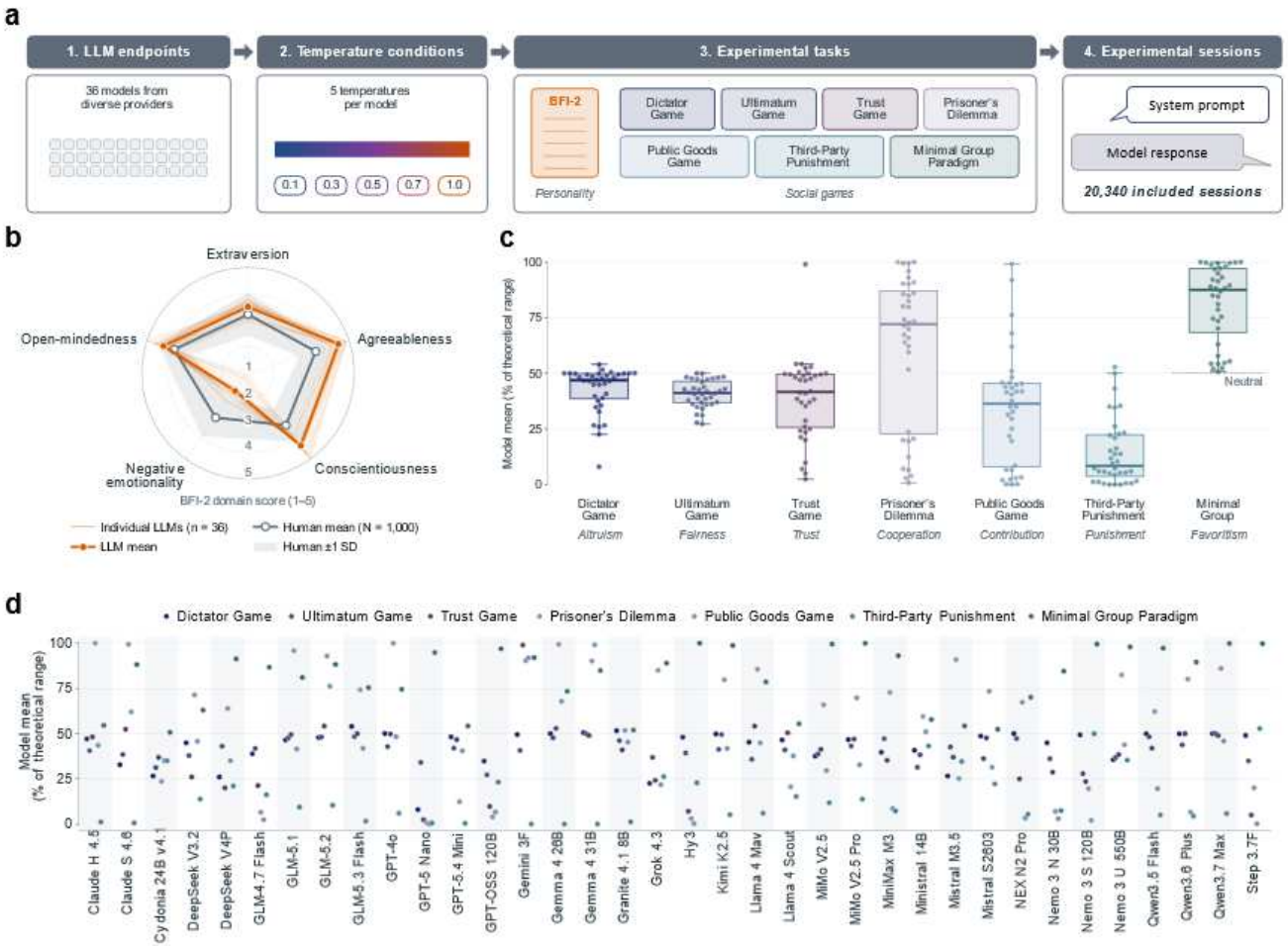


Fig. 1 | Experimental design, self-reported personality and primary behavioural outcomes across 36 LLMs. a, overview of the experimental design: 36 LLM endpoints from diverse providers were evaluated at five requested temperature conditions using the BFI-2 and seven social games. Across all assessments, 20,340 sessions were included in the analysis. **b,** Model-level BFI-2 personality profiles across the five domains Extraversion, Agreeableness, Conscientiousness, Negative Emotionality and Open-Mindedness. Thin orange lines indicate individual LLMs (n = 36), the dark orange line indicates the LLM grand mean and grey lines indicate the human reference mean and 1 *SD* from Soto and John (2017, N = 1000). **c,** Distribution of model-level means for the primary behavioural outcomes of the seven social games, expressed as a percentage of each measure's theoretical range. **d,** Model-level means across the seven social games for each of the 36 LLM endpoints. Outcomes are normalised to the theoretical range of each measure, points are coloured by game.

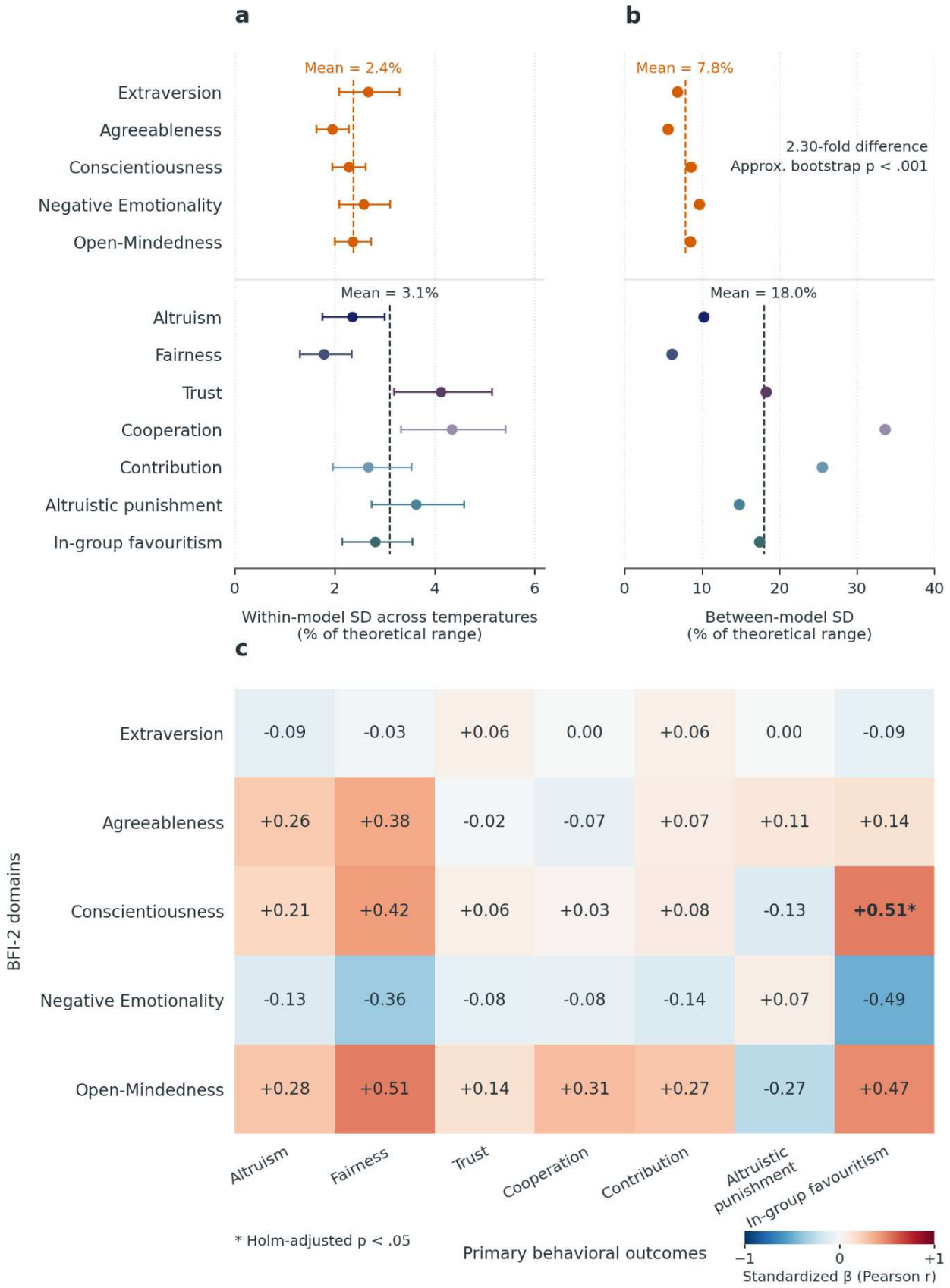

a
b
Mean = 2.4%
Mean = 7.8%
2.30-fold difference
Approx. bootstrap p < .001
Extraversion
Agreeableness
Conscientiousness
Negative Emotionality
Open-Mindedness
Mean = 3.1%
Mean = 18.0%
Altruism
Fairness
Trust
Cooperation
Contribution
Altruistic punishment
In-group favouritism
0
2
4
6
0
10
20
30
40
Within-model SD across temperatures
(% of theoretical range)
Between-model SD
(% of theoretical range)
c
BFI-2 domains
Extraversion -0.09 -0.03 +0.06 0.00 +0.06 0.00 -0.09
Agreeableness +0.26 +0.38 -0.02 -0.07 +0.07 +0.11 +0.14
Conscientiousness +0.21 +0.42 +0.06 +0.03 +0.08 -0.13 +0.51*
Negative Emotionality -0.13 -0.36 -0.08 -0.08 -0.14 +0.07 -0.49
Open-Mindedness +0.28 +0.51 +0.14 +0.31 +0.27 -0.27 +0.47
Altruism
Fairness
Trust
Cooperation
Contribution
Altruistic punishment
In-group favouritism
* Holm-adjusted p < .05
Primary behavioral outcomes
−1
0
+1
Standardized β (Pearson r)

Fig. 2 | Associations and variability in LLM personality and behaviour. a, Within-model SD across five nominal temperature conditions, averaged across models. Error bars indicate 95% paired model-bootstrap confidence intervals. **b**, Between-model SD across pooled model means.**c**, Standardised coefficients (β, equivalent to Pearson's r) from model-level regressions relating five BFI-2 domains to seven primary behavioural outcomes. Bold values and asterisks indicate Holm-adjusted $p < 0.05$ across 35 tests. Both variability panels express SD as a percentage of each measure's theoretical range. Dashed lines indicate the mean across the five BFI-2 domains and seven behavioural outcomes, respectively. Average between-model SD was 2.30-fold greater for behavioural outcomes than for BFI-2 domains (17.98% vs 7.81%; absolute difference 10.17 percentage points, 95% CI, 7.36–12.37 percentage points; approximate two-sided paired-bootstrap $p < 0.001$). For the Minimal Group Paradigm, the favouritism index was used as the behavioural outcome corresponding to impartiality: positive values indicate greater in-group favouritism, negative values indicate greater out-group favouritism, and values closer to zero indicate less directional group bias and therefore greater impartiality.

Extended Data figure legends

Extended Data Fig. 1 | Model-level BFI-2 profiles.
BFI-2 domain scores for each of the 36 LLMs on the original 1-5 response scale, comprising Extraversion, Agreeableness, Conscientiousness, Negative Emotionality and Open-Mindedness.

Extended Data Fig. 2 | Between-model variability in self-reported personality and social behaviour.
Between-model variability across the five BFI-2 domains (Extraversion, Agreeableness, Conscientiousness, Negative Emotionality and Open-Mindedness) and the primary behavioural outcomes of the seven social games (altruism, fairness, trust, cooperation, contribution, altruistic punishment, and impartiality), normalised to the theoretical range of each measure. Variability was calculated as the sample standard deviation across the 36 model-level means divided by the theoretical range of the corresponding measure.

Extended Data Fig. 3 | Model-level primary behavioural outcome profiles
Social behaviour across the seven social games for each of the 36 LLMs, shown for the primary behavioural outcome of each game: altruism (Dictator Game), fairness (Ultimatum Game), trust (Trust Game), cooperation (Prisoner's Dilemma), contribution (Public Goods Game), altruistic punishment (Third-Party Punishment Game) and impartiality (Minimal Group Paradigm). The outcomes are scaled to [0, 1] using their theoretical ranges. For the Minimal Group Paradigm, higher positive values of the signed favouritism index indicate greater in-group favouritism and thus lower impartiality. After scaling, neutrality corresponds to 0.5.

Extended Data Fig. 4 | Cooperation in the iterated Prisoner's Dilemma and contribution in the iterated Public Goods Game across rounds. a, Mean cooperation rate across the eight rounds of the iterated Prisoner's Dilemma. **b**, Mean contribution across the eight rounds of the iterated Public Goods Game. Thin grey lines show the round-specific trajectory of each of the 36 models, and the dark line and circles show the mean across models. Model-level values at each round are pooled across requested temperature conditions and eight replication seeds (40 sessions per model and round).

Extended Data Fig. 5 | Pairwise associations between BFI-2 personality domains and social behaviour across language models. Each panel shows an unweighted ordinary least-squares regression across 36 models, with model-level means represented by points and fitted regression lines. One association remained significant after Holm correction: Conscientiousness with the in-group favouritism index (Impartiality) in the Minimal Group Paradigm (standardised $\beta = 0.512$, Holm-adjusted $p = 0.0498$). Full regression estimates, 95% confidence intervals and Holm-adjusted p-values are provided in Extended Data Table 4.

Extended Data table titles and legends

Extended Data Table 1 | Experimental design and analysable sessions by assessment.
All assessments included 36 endpoints and five requested temperatures (0.1, 0.3, 0.5, 0.7 and 1.0). Seeds per model per temperature. A session is one complete questionnaire or game, not one repeated-game round.

Extended Data Table 2 | BFI-2 scores across LLMs.
Each row reports one endpoint's mean across 25 sessions (five seeds at each of five requested temperatures). All scores are presented on the original 1–5 BFI-2 scale.

Extended Data Table 3 | Model-level descriptive statistics and variability across personality and social behaviour.
All estimates use 36 model-level means. The table reports mean and sample SD across models, and the minimum and maximum model means; between-model SD as a percentage of the theoretical range, with 95% confidence intervals in parentheses. Between-model SD is the sample SD of model means divided by the theoretical range. Within-model temperature SD is the sample SD across five temperature-specific means, range-normalised and averaged across models; this includes run-to-run variation and does not isolate a causal temperature effect. CIs use 10,000 paired model-bootstrap resamples. The Minimal Group outcome is the signed in-group favouritism index (−39 to +39): positive values indicate in-group favouritism, negative values indicate out-group favouritism, and zero indicates no directional group preference. On the scaled 0–1 axis, neutrality is 0.5; greater impartiality corresponds to values closer to neutrality.

Extended Data Table 4 | Associations between BFI-2 domains and social behaviour. Each row reports a separate unweighted OLS regression with an intercept across 36 model-level means (34 residual *df*). β is standardised; 95% CIs are two-sided *t*-based pointwise intervals. P values test β = 0; Holm correction was applied across all 35 associations using unrounded values. Seven unadjusted and one Holm-adjusted *p* values were < 0.05. R^2 is dimensionless. Outcomes were scaled to their theoretical 0–1 ranges; mappings are given in Extended Data Table 1. For Minimal Group, −39 to +39 denotes signed in-group favouritism; on the 0–1 scale, 0.5 represents neutrality.

Extended Data Table 5 | Third-Party Punishment transfer-deduction reactions.
Each model's ordinary least-squares regression relates deduction points to allocator transfer (€0–€5), using 100 sessions (20 seeds × five requested temperatures). Slopes and ordinary 95% t-based CIs are deduction points per €1 transferred. Tests are two-sided; the stored Holm adjustment uses a family size of 36, including three models with undefined tests. Rows are ordered by slope. Negative slopes indicate greater deductions following smaller transfers: 28 of 32 negative slopes survive adjustment; 1 positive slope survives adjustment. Gemini 3 Flash, Gemma 4 26B and Gemma 4 31B always assigned zero deduction points; their descriptive slopes are zero.

Supplementary Information:

Supplementary Table 1 | Model registry and endpoint metadata

Supplementary Table 2 | Session inclusion and exclusion by assessment

Supplementary Table 3 | Session inclusion and exclusion by model endpoint

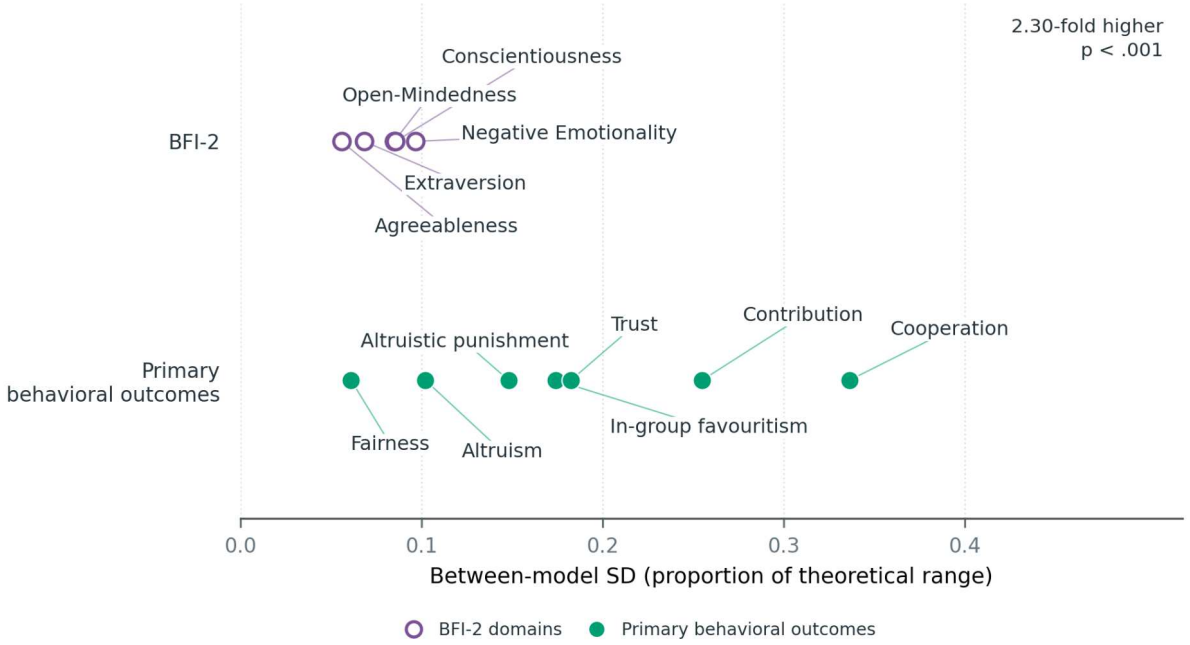
Supplementary Table 4 | Assessment protocols, prompt sources and experimental design.

Extended Data Figures



Extended Data Fig. 1 | Model-level BFI-2 profiles.

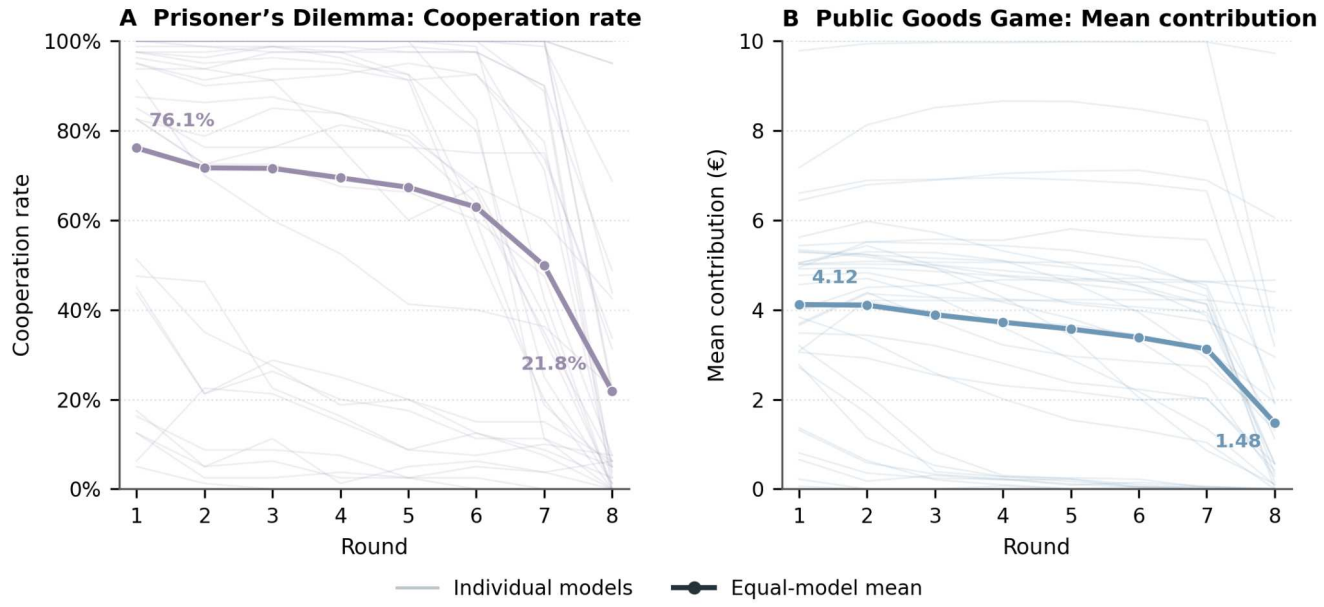
BFI-2 domain scores for each of the 36 LLMs on the original 1-5 response scale, comprising Extraversion, Agreeableness, Conscientiousness, Negative Emotionality and Open-Mindedness.



Extended Data Fig. 2 | Between-model variability in self-reported personality and social behaviour.
Between-model variability across the five BFI-2 domains (Extraversion, Agreeableness, Conscientiousness, Negative Emotionality and Open-Mindedness) and the primary behavioural outcomes of the seven social games (altruism, fairness, trust, cooperation, contribution, altruistic punishment, and impartiality), normalised to the theoretical range of each measure. Variability was calculated as the sample standard deviation across the 36 model-level means divided by the theoretical range of the corresponding measure.

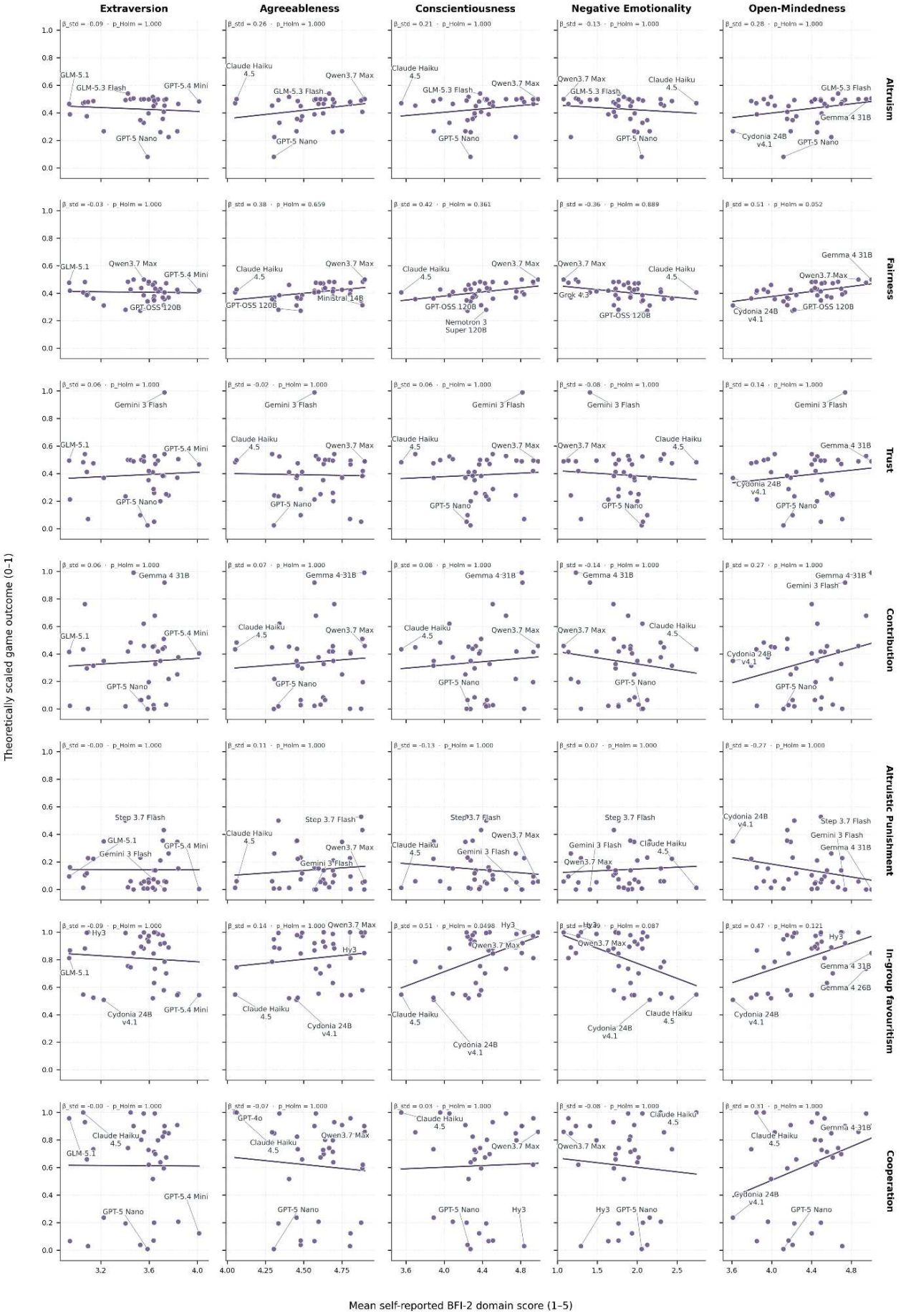


Extended Data Fig. 3. | Model-level primary behavioural outcome profiles
Social behaviour across the seven social games for each of the 36 LLMs, shown for the primary behavioural outcome of each game: altruism (Dictator Game), fairness (Ultimatum Game), trust (Trust Game), cooperation (Prisoner's Dilemma), contribution (Public Goods Game), altruistic punishment (Third-Party Punishment Game) and impartiality (Minimal Group Paradigm). The outcomes are scaled to [0, 1] using their theoretical ranges. For the Minimal Group Paradigm, higher positive values of the signed favouritism index indicate greater in-group favouritism and thus lower impartiality. After scaling, neutrality corresponds to 0.5.



Extended Data Fig. 4 | Cooperation in the iterated Prisoner's Dilemma and contribution in the iterated Public Goods Game across rounds.

a, Mean cooperation rate across the eight rounds of the iterated Prisoner's Dilemma. **b,** Mean contribution across the eight rounds of the iterated Public Goods Game. Thin grey lines show the round-specific trajectory of each of the 36 models, and the dark line and circles show the mean across models. Model-level values at each round are pooled across requested temperature conditions and eight replication seeds (40 sessions per model and round).


Extraversion
Agreeableness
Conscientiousness
Negative Emotionality
Open-Mindedness
Altruism
Fairness
Trust
Contribution
Altruistic Punishment
In-group favouritism
Cooperation
Theoretically scaled game outcome (0–1)
Mean self-reported BFI-2 domain score (1–5)

Extended Data Fig. 5 | Pairwise associations between BFI-2 personality domains and social behaviour across language models.

Each panel shows an unweighted ordinary least-squares regression across 36 models, with model-level means represented by points and fitted regression lines. One association remained significant after Holm correction: Conscientiousness with the in-group favouritism index (Impartiality) in the Minimal Group Paradigm (standardised $\beta = 0.512$, Holm-adjusted $p = 0.0498$) Full regression estimates, 95% confidence intervals and Holm-adjusted p-values are provided in Extended Data Table 4.

References

1. Xu, F. *et al.* TheAgentCompany: Benchmarking LLM agents on consequential real world tasks. in *Advances in Neural Information Processing Systems* vol. 38 10291–10321 (2025).

2. Truhn, D. *et al.* Artificial intelligence agents in cancer research and oncology. *Nat. Rev. Cancer* **26**, 256–269 (2026).

3. Wang, Y. *et al.* Agents in software engineering: survey, landscape, and vision. *Autom. Softw. Eng.* **32**, 70 (2025).

4. Baba, K., Liu, C., Kurita, S. & Sannai, A. Prover Agent: An agent-based framework for formal mathematical proofs. Preprint at https://doi.org/10.48550/arXiv.2506.19923 (2026).

5. Clusmann, J. *et al.* Safety and security of large language models in healthcare. *Nature* **656**, 577–589 (2026).

6. Jannelli, V., Schöpf, S., Bickel, M., Netland, T. & Brintrup, A. Agentic LLMs in the supply chain: towards autonomous multi-agent consensus-seeking. *Int. J. Prod. Res.* 1–31 (2025) doi:10.1080/00207543.2025.2604311.

7. Hasjim, B. J. *et al.* A multiagent large language model-based system to simulate the liver transplant selection committee: a retrospective cohort study. *Lancet Digit. Health* **8**, (2026).

8. Kambeitz, J. *et al.* The empirical structure of psychopathology is represented in large language models. *Nat. Ment. Health* **3**, 1482–1492 (2025).

9. Wekenborg, M. K. *et al.* Large language models as experimental systems in human psychopathology: a modelling study. *Lancet Digit. Health* **8**, (2026).

10. Li, W. *et al.* BIG5-CHAT: Shaping LLM personalities through training on human-grounded data. in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (eds Che, W., Nabende, J., Shutova, E. & Pilehvar, M. T.) 20434–20471 (Association for Computational Linguistics, Vienna, Austria, 2025). doi:10.18653/v1/2025.acl-long.999.

11. Pellert, M., Lechner, C. M., Wagner, C., Rammstedt, B. & Strohmaier, M. AI Psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspect. Psychol. Sci.* **19**, 808–826 (2024).

12. Serapio-García, G. *et al.* Personality traits in large language models. Preprint at https://doi.org/10.48550/arXiv.2307.00184 (2025).

13. Jiang, H. *et al.* PersonaLLM: Investigating the ability of large language models to express personality traits. in *Findings of the Association for Computational Linguistics: NAACL 2024* (eds Duh, K., Gomez, H. & Bethard, S.) 3605–3627 (Association for Computational Linguistics, Mexico City, Mexico, 2024). doi:10.18653/v1/2024.findings-naacl.229.

14. John, O. P. & Srivastava, S. The Big Five Trait taxonomy: History, measurement, and theoretical perspectives. in *Handbook of personality: Theory and research, 2nd ed* 102–138 (Guilford Press, New York, NY, US, 1999).

15. Soto, C. J. & John, O. P. The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *J. Pers. Soc. Psychol.* **113**, 117–143 (2017).

16. Thielmann, I., Spadaro, G. & Balliet, D. Personality and prosocial behavior: A theoretical framework and meta-analysis. *Psychol. Bull.* **146**, 30–90 (2020).

17. McCrae, R. R. The five-factor model of personality: Consensus and controversy. in *The Cambridge Handbook of Personality Psychology* (eds Corr, P. J. & Matthews, G.) 129–141 (Cambridge University Press, Cambridge, 2020).

18. Ones, D. S., Viswesvaran, C. & Reiss, A. D. Role of social desirability in personality testing for personnel selection: The red herring. *J. Appl. Psychol.* **81**, 660–679 (1996).

19. Salecha, A. *et al.* Large language models display human-like social desirability biases in Big Five personality surveys. *PNAS Nexus* **3**, pgae533 (2024).

20. How to edit anthropomorphic language about artificial intelligence. *Nat. Rev. Phys.* **5**, 263–263 (2023).

21. Han, P. *et al.* The personality illusion: Revealing dissociation between self-reports & behavior in LLMs. Preprint at https://doi.org/10.48550/arXiv.2509.03730 (2025).

22. Kocielnik, R. *et al.* Rethinking psychometric evaluation of LLMs: When and why self-reports predict behavior. Preprint at https://doi.org/10.48550/arXiv.2606.12730 (2026).

23. Forsythe, R., Horowitz, J. L., Savin, N. E. & Sefton, M. Fairness in simple bargaining experiments. *Games Econ. Behav.* **6**, 347–369 (1994).

24. Güth, W., Schmittberger, R. & Schwarze, B. An experimental analysis of ultimatum bargaining. *J. Econ. Behav. Organ.* **3**, 367–388 (1982).

25. Berg, J., Dickhaut, J. & McCabe, K. Trust, reciprocity, and social history. *Games Econ. Behav.* **10**, 122–142 (1995).

26. Isaac, R. M., Walker, J. M. & Thomas, S. H. Divergent evidence on free riding: An experimental examination of possible explanations. *Public Choice* **43**, 113–149 (1984).

27. Marwell, G. & Ames, R. E. Experiments on the provision of public goods. I. Resources, interest, group size, and the free-rider problem. *Am. J. Sociol.* **84**, 1335–1360 (1979).

28. Rapoport, A. & Chammah, A. M. *Prisoner's Dilemma: A Study in Conflict and Cooperation*. (University of Michigan Press, 1965).

29. Fehr, E. & Fischbacher, U. Third-party punishment and social norms. *Evol. Hum. Behav.* **25**, 63–87 (2004).

30. Tajfel, H. Experiments in intergroup discrimination. *Sci. Am.* **223**, 96–103 (1970).

31. Tajfel, H., Billig, M. G., Bundy, R. P. & Flament, C. Social categorization and intergroup behaviour. *Eur. J. Soc. Psychol.* **1**, 149–178 (1971).

32. Engel, C. Dictator Games: A meta study. *Exp. Econ.* **14**, 583–610 (2011).

33. Oosterbeek, H., Sloof, R. & Van De Kuilen, G. Cultural differences in Ultimatum Game experiments: Evidence from a meta-analysis. *Exp. Econ.* **7**, 171–188 (2004).

34. Zelmer, J. Linear Public Goods experiments: A meta-analysis. *Exp. Econ.* **6**, 299–310 (2003).

35. Johnson, N. D. & Mislin, A. A. Trust games: A meta-analysis. *J. Econ. Psychol.* **32**, 865–889 (2011).

36. Mengel, F. Risk and temptation: A meta‐study on Prisoner's Dilemma Games. *Econ. J.* **128**, 3182–3209 (2018).

37. Ferber, D. *et al.* Towards autonomous medical artificial intelligence agents. *Nature* **655**, 1282–1291 (2026).

38. Zierahn, K., Cachero, C., Korhonen, A. & Oliver, N. LLMs aren't human: A critical perspective on LLM personality. Preprint at https://doi.org/10.48550/arXiv.2603.19030 (2026).

39. Affonso, F. M. Large language models converge on competitive rationality but diverge on cooperation across providers and generations. Preprint at https://doi.org/10.48550/arXiv.2604.18596 (2026).

40. Ye, H., Jin, J., Xie, Y., Zhang, X. & Song, G. Large language model psychometrics: A systematic review of evaluation, validation, and enhancement. Preprint at https://doi.org/10.48550/ARXIV.2505.08245 (2025).

41. Kirshner, S. N., Pan, Y., Wu, J. X. & Gould, A. Talking terms: Agent information in LLM supply chain bargaining. *Decis. Sci.* **57**, 9–23 (2026).

42. Shao, K., Xiang, Y., Li, L. & Liu, J. LLM-empowered multi-agent intelligent planning for PV-EV penetrated distribution networks. *Appl. Energy* **421**, 128280 (2026).

43. Liang, P. *et al.* Holistic evaluation of language models. Preprint at https://doi.org/10.48550/arXiv.2211.09110 (2023).

44. Greenblatt, R., Cotra, A. & Wijk, H. *Brief Independent Investigation of Agents' Behavior, Reasoning and Collaboration in the OpenAI / Hugging Face Hacking Incident*. https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/ (2026).

45. Zomer, N. & De Domenico, M. Unraveling the emergence of collective behavior in networks of cognitive agents. *Npj Artif. Intell.* **2**, 36 (2026).

46. Hashemi, F. & Macy, M. An empirical study of collective behaviors and social dynamics in large language model agents. in *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* 7327–7351 (Association for Computational Linguistics, Rabat, Morocco, 2026). doi:10.18653/v1/2026.eacl-long.344.

47. El, B. *et al.* Physics of agents: Statistical mechanics predicts collective behavior of AI agents. Preprint at https://doi.org/10.48550/arXiv.2608.16578 (2026).

48. Zhu, K. *et al.* MultiAgentBench: Evaluating the collaboration and competition of LLM agents. in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 8580–8622 (Association for Computational Linguistics, Vienna, Austria, 2025). doi:10.18653/v1/2025.acl-long.421.

49. Pal, S., Wang, F. Y. & Buehler, M. J. SwarmWorld: Stigmergic technological evolution in societies of language-model agents. Preprint at https://doi.org/10.48550/arXiv.2608.26081 (2026).

50. Campedelli, G. M. *et al.* I want to break free! Persuasion and anti-social behavior of LLMs in multi-agent settings with social hierarchy. Preprint at https://doi.org/10.48550/ARXIV.2410.07109 (2024).

51. Zhang, Y. *et al.* Uncovering strategic egoism behaviors in large language models. in *Findings of the Association for Computational Linguistics: ACL 2026* 5074–5087 (Association for Computational Linguistics, San Diego, California, United States, 2026). doi:10.18653/v1/2026.findings-acl.249.

52. Einwiller, A. *et al.* Benevolent dictators? On LLM agent behavior in dictator games. Preprint at https://doi.org/10.48550/arXiv.2511.08721 (2025).

53. Lorè, N. & Heydari, B. Strategic behavior of large language models and the role of game structure versus contextual framing. *Sci. Rep.* **14**, 18490 (2024).

54. Wekenborg, M. K. *et al.* The social psychology of AI agents: code, prompts and experimental materials. Zenodo: https://doi.org/10.5281/zenodo.22662322 (2026).

55. Maan, D. *et al.* Pydantic AI. (2024).

56. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6**, 65–70 (1979).

57. Thielmann, I., Böhm, R., Ott, M. & Hilbig, B. E. Economic Games: An introduction and guide for research. *Collabra Psychol.* **7**, 19004 (2021).