Agent psychonomicus: Psychological mechanisms emerge in cognitive agents simulated by large language models

Jack Chen[1*], Weiheng Xiao[2], Lexi Chen[3]

[1]Columbia Business School, Columbia University in the City of New York, New York, NY 10027, USA.

[2]Tasteful AI, Jersey City, NJ 07306, USA.

[3]The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA.

*Corresponding author. Email: cchen27@gsb.columbia.edu

This version: November 9, 2026

Copyright © 2026 Jack Chen, Weiheng Xiao, and Lexi Chen. All rights reserved. This working paper is distributed for the purposes of comment and discussion only. It may not be reproduced, distributed, or used, in whole or in part, without the prior written permission of the authors.

Abstract

Researchers increasingly use large language models (LLMs) to simulate participants in psychological experiments. Yet researchers observe only the LLM's answer and cannot tell whether a manipulation acted through the psychological mechanism that a theory proposes. We therefore introduce *Agent psychonomicus*, a class of cognitive models for testing each step of the mechanism proposed by a process theory. We implement it as the Minimal Viable Cognitive Agent (MVCA), in which an ensemble of 12 LLM agents carries out general cognitive processes, such as appraising how a situation bears on the self, and states the results as first-person process statements. Using three different LLMs, we conducted 57,147 cognitive simulations of 60 published experiments and analyzed over six hundred thousand process statements. We found that the theorized psychological mechanism, or the mediator, emerged from general processes simulated by MVCA. Inducing the theorized mechanism directly in process statements had a causal effect on the experiments' outcomes. Among experiments re-examined in large preregistered replications, MVCA's cognitive agents showed a significantly larger share of the original effects that replicated than of those that did not, suggesting that MVCA's cognitive simulations of human participants are diagnostic and address the false-positive results of silicon samples in existing research.

Keywords: large language models; cognitive agents; process theories; mediation; moderation; experimental causal chain; trait activation; consumer psychology

Introduction

Consumers who touch a mug given to them by the experimenter are willing to pay more for it, because touching, according to the theory of psychological ownership, makes them feel that the mug is their own[1,2]. Effects of this kind have been reproduced with large language models (LLMs) that simulate the participants of experiments, also referred to as silicon samples[3,4]. Do LLMs, then, also feel that the mug is their own?

This question matters in practice, because researchers increasingly use LLMs to simulate the participants of social science research. In such a simulation, researchers describe to an LLM the profile of a participant, such as age, gender, occupation, and values, present the LLM with the questions and stimuli that a human participant would see, and analyze the LLM's answers as the participant's responses[3,5–7]. In research on silicon samples, for example, researchers create many profiles that match the respondents of a survey and examine whether the LLM answers as these respondents did[5,8]. Building on this approach, researchers have created digital twins of particular participants, each based on one participant's interviews or earlier survey responses, with the goal of predicting how this participant would respond to new questions and experiments[9,10].

Both research streams aim to reproduce the responses of human participants as closely as possible. Toward this end, researchers have compared the answers of silicon samples and digital twins with the responses of human participants, with mixed results[11]. Silicon samples reproduced classic findings of behavioral economics and social psychology[3], and predicted the treatment effects of 70 preregistered survey experiments with nationally representative samples in the United States[7]. In scenario-based experiments from psychology and management, however, silicon samples also produced significant effects where the original studies had found none in human participants[4], and the answers of digital twins correlated only weakly with the responses of the human participants from whose answers these twins had been built[10].

For psychologists, however, describing how human decision-makers respond to an intervention or behave in a social setting is only a first step toward explaining why they respond

as they do. Such explanations often take the form of a process theory, which holds that a factor, such as an action, a feature of the situation, or a property of a stimulus, alters a psychological state within the individual, which in turn affects the individual's response[12,13]. This psychological state, often called the mediator, constitutes the psychological mechanism that the theory proposes. In the mug example, the feeling that the mug is one's own is the mediator through which touching has a positive effect on willingness to pay.

To test a process theory, psychologists typically follow the experimental causal chain approach[14], in which they manipulate the factor that the theory proposes as the cause and examine four propositions: first, that the manipulation affects the mediator; second, that the mediator affects the response; third, that features of the situation moderate the mediator's effect on the response; and fourth, that individual characteristics produce heterogeneity in the manipulation's effect (Fig. 1A). They usually test the first proposition by measuring the mediator after the manipulation (a measurement-of-mediation design)[14], and the second by manipulating the mediator directly in another experiment (a manipulation-of-mediator design)[15]. To test the third proposition, psychologists sometimes adopt the moderation-of-process design, in which they vary another feature of the situation[14]. To test the fourth proposition, finally, psychologists measure individual characteristics and examine whether the manipulation affects some individuals more than others[16,17].

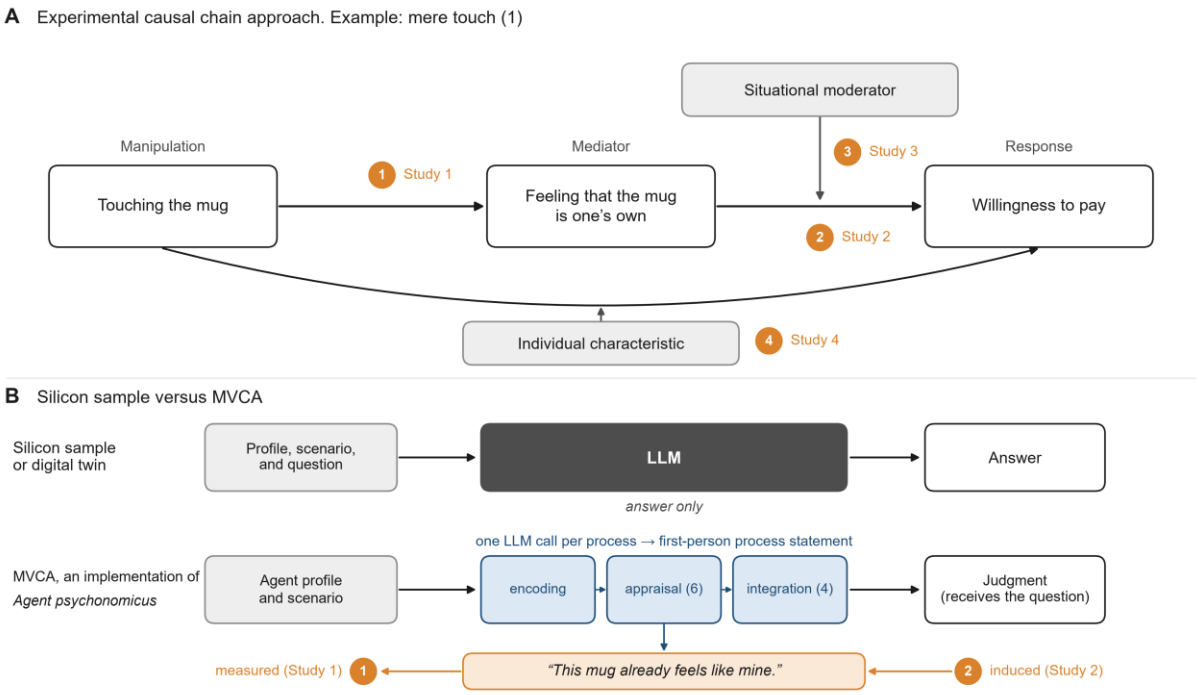


Fig. 1 | Testing a process theory in cognitive agents. (**A**) The experimental causal chain approach, illustrated with the theory of mere touch[1], according to which touching a mug leads consumers to feel that the mug is their own, which in turn has a positive effect on their willingness to pay. Numbered circles mark the four propositions and the studies that test them. (**B**) A silicon sample or digital twin, in which the researcher observes only the LLM's answer, compared with MVCA, an implementation of *Agent psychonomicus*, in which the LLM carries out each cognitive process in a separate LLM call and states its result as a first-person process statement. The mediator can be measured in these process statements (Study 1) or induced as a process statement (Study 2).

With silicon samples or digital twins, however, these tests are hard to conduct. Because researchers present the LLM with a scenario, a text that describes one experimental condition, together with a question, and observe only the LLM's answer, they cannot easily tell whether the manipulation's effect on the LLM's answers arose through the psychological mechanism that the theory proposes, nor manipulate this mechanism before the LLM answers. Moreover, the determinants of the LLM's answers are often encoded in the parameters of the LLM, where researchers cannot easily inspect them[18]. Thus, even if the LLM's answers matched the responses of human participants exactly, such a study could not provide evidence for the psychological mechanism that the theory proposes.

To test a process theory with participants simulated by LLMs, researchers therefore need to observe and manipulate the psychological mechanism that the theory proposes, to test whether

a situational moderator alters the mediator's effect, and to test whether individual characteristics produce heterogeneity in the manipulation's effect. We call any model that meets these requirements an *Agent psychonomicus* and the individuals that it simulates cognitive agents.

Being able to observe and manipulate this mechanism alone does not, however, mean that the responses of cognitive agents yield causal evidence for the theory. Following the received view of experimental methods in psychology, we take the perspective that an experiment tests a theory only together with assumptions about its measures, conditions, and participants[19,20]. Drawing on this perspective and on the conception of the person in social cognition research[21,22], we propose that the cognitive processes of an *Agent psychonomicus* must have four functional properties, each of which corresponds to an assumption that the experimental causal chain approach makes about human participants (Fig. 2A). These four functional properties are that (1) the cognitive agents' process statements result from how they interpret, or construe, the situation; (2) their responses are based on the content of the process statements that precede them; (3) they act on a process statement only when the situation makes it relevant to the decision; and (4) their individual characteristics shape their responses in situations that offer cues to which these characteristics respond. We call these properties construal, reliance, conditionality, and individuality. In Supplementary Note 1, we derive each property from the experimental causal chain approach[14,15,23].

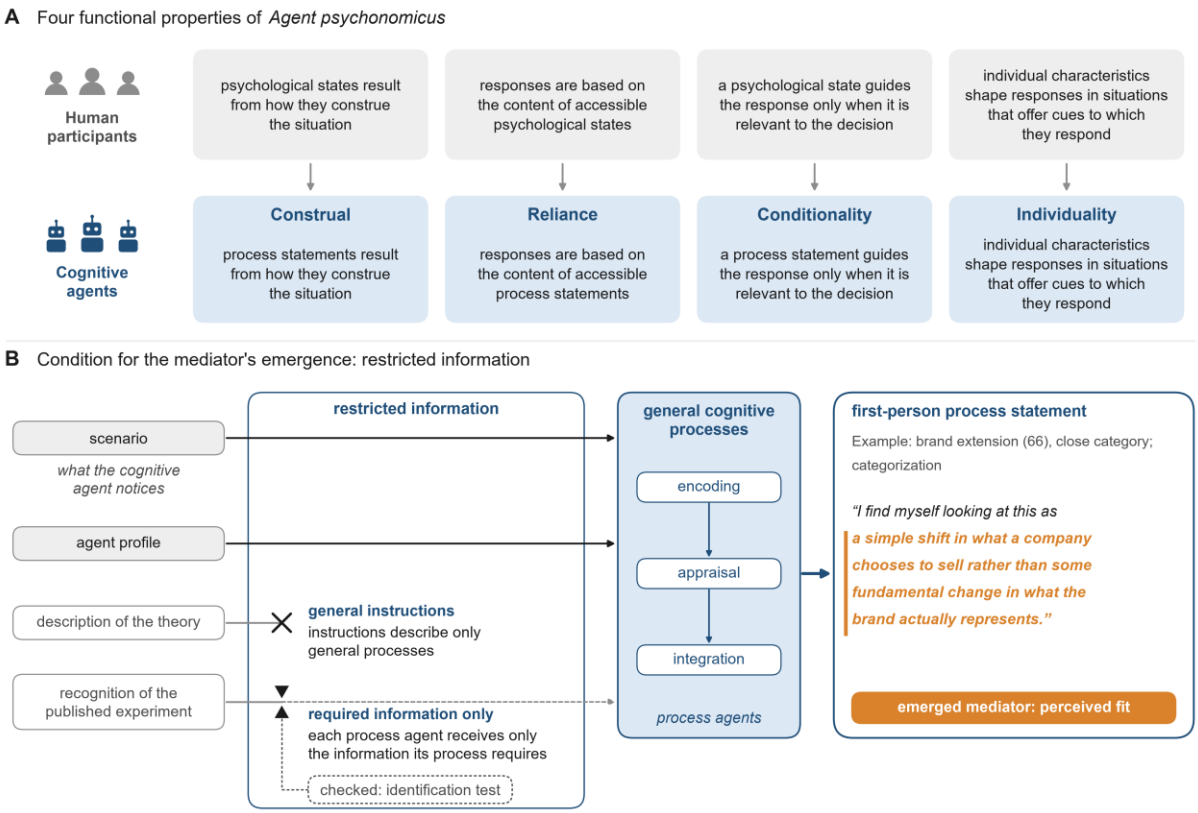


Fig. 2 | Functional properties of *Agent psychonomicus* and the condition for the mediator's emergence. (**A**) Functional properties of the cognitive agents (blue) and the corresponding assumptions that the experimental causal chain approach makes about human participants (gray)[22,35,37,42,60]. (**B**) Restricted information. No process agent receives a description of the theory (cross), and each process agent receives only the information that its process requires, so that it has little basis for recognizing the published experiment (narrowed dashed arrow). On the right is a process statement from the categorization process of one cognitive agent in the close-category condition of the brand-extension experiment[66] (Gemma 4 26B), with the passage that conveys the mediator, perceived fit, in orange.

In addition to the four functional properties, testing a process theory in cognitive agents requires that the psychological states that the theory proposes arise in the cognitive agents as they arise in human decision-makers. We therefore propose that an *Agent psychonomicus* be built on psychological constructionism, which holds that psychological states arise from general cognitive processes that operate in every situation rather than from processes dedicated to particular states[24]. In this paper, we construct and test the Minimal Viable Cognitive Agent (MVCA), an implementation of *Agent psychonomicus* using LLMs, whose cognitive agents we designed to have the four functional properties. Following psychological constructionism and research on language agents, which proposes dividing an agent's work into separate tasks[25,26], we

designed MVCA so that the LLM carries out each general cognitive process, such as appraising how a situation bears on the self, in a separate LLM call governed by its own instructions, the system prompt. In each LLM call, the LLM states the result of the process as a short first-person process statement, similar to thought-listing procedures in which human participants list their thoughts[27,28].

We then supply this sequence of processes with an agent profile, which specifies an individual's demographic attributes and levels on 139 individual characteristics studied in psychology, sampled from population norms and other distributions (Methods), so that MVCA simulates one cognitive agent from each profile. Researchers can then measure the mediator in the process statements that precede a cognitive agent's response and induce the mediator as a process statement, such as "this mug already feels like mine," while the scenario and the agent profile stay the same (Fig. 1B). They can also test whether a feature of the situation moderates the mediator's effect by varying it in the scenario, and whether an individual characteristic produces heterogeneity in the manipulation's effect by varying its level in the agent profile.

Because LLMs are trained on text that includes published research, an LLM that receives the materials of a published experiment can also draw on its knowledge of the published finding[29]. In an *Agent psychonomicus* built with LLMs, such as MVCA, each process agent, which carries out one process, should therefore receive no information about the theory, or even about the experiment under examination, with the exception of the scenario for the process agent for encoding and the question for the process agent for judgment. Each process agent should instead receive only the information that its process requires, such as what the cognitive agent has noticed in the situation, a condition that we call restricted information (Fig. 2B), similar to the way in which human participants are intentionally kept unaware of the hypothesis to avoid experimenter effects[30]. When the psychological state that a theory proposes as the mediator appears in the process statements of cognitive agents whose process agents receive restricted information, we propose that the mediator has *emerged* from the general processes.

We tested MVCA in 60 influential experiments in social and consumer psychology, published in journals such as the *Journal of Consumer Research*, the *Journal of Personality and Social Psychology*, and *Psychological Science*. These experiments concern mediators of different kinds, such as a view of oneself, a desire, a feeling, or an inference about others, and include

experiments whose original effects have been re-examined in large preregistered replications. Using three LLMs from different developers, we conducted 57,147 cognitive simulations of the 60 experiments and analyzed 628,617 first-person process statements. Across four studies, we found that the cognitive agents of MVCA supported each of the four propositions of a process theory. Importantly, we found that the theorized psychological mechanism, or the mediator, emerged from the general processes, and that inducing the mediator directly in the process statements had a causal effect on the cognitive agents' responses. In addition, among the experiments that large preregistered replications had re-examined, the cognitive agents showed a significantly larger share of the original effects that had replicated than of those that had not, suggesting that MVCA's cognitive simulations are diagnostic and address the false-positive results of silicon samples in existing research[4]. Where an original effect had not replicated, the manipulation had a much smaller positive effect on the mediator than where it had replicated, whereas the induced mediator had a positive effect on responses in both groups of experiments.

Research on silicon samples and digital twins has mostly asked whether an LLM's answers reproduce the responses of human participants[31], whereas we ask whether researchers can test the psychological mechanism that a theory proposes in participants simulated by LLMs. By constructing MVCA and testing it against 60 published experiments, we show that an *Agent psychonomicus* can be built from existing LLMs without fine-tuning on the responses of human participants, so that researchers can test each step of the mechanism in cognitive agents, and locate the step at which a theory fails, before they run an experiment with human participants.

Results

Cognitive architecture of MVCA

Following psychological constructionism, we built MVCA from 12 general processes, each carried out by a process agent, a separate LLM call governed by its own system prompt (Fig. 3A; Table 1). In the first process, the LLM describes what the cognitive agent notices in the scenario; this description is the cognitive agent's encoding of the situation[21]. The cognitive agent then evaluates what it noticed in six appraisal processes, derived from appraisal theories and psychological constructionism[24,32,33], combines these appraisals in four integrating processes, and finally, in the process of judgment, responds to the question of the experiment on the basis of these process statements. We call the process statements and the response that MVCA produces

for one cognitive agent in one experimental condition a cognitive simulation. We present details of the full architecture in Supplementary Note 2 and the instructions and information that each process agent receives in Supplementary Note 3.

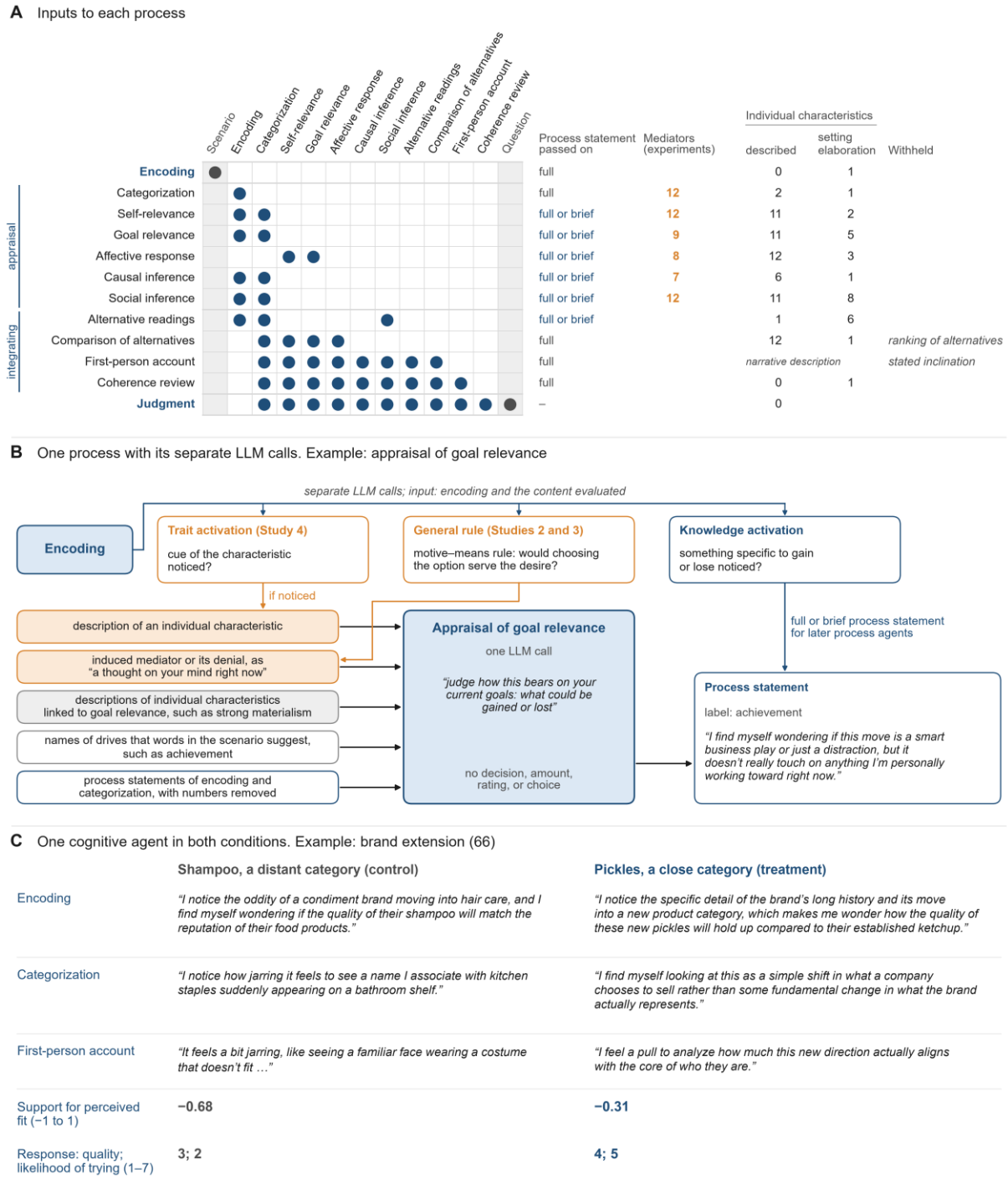


Fig. 3 | Cognitive architecture of MVCA. (**A**) Information that the LLM receives for each process, marked by dots, namely earlier process statements and, in the gray columns, the

scenario (encoding only) and the question (judgment only). The columns on the right show, for each process, how later process agents receive its process statement (in full, or under knowledge activation in full or as a brief version), the number of experiments whose mediator arises in it (Supplementary Table 1), the numbers of individual characteristics described to the LLM and of those that make the process statement fuller or briefer, and the parts withheld from later process agents. (**B**) Appraisal of goal relevance, with the separate LLM calls for trait activation (Study 4), the general rules (Studies 2 and 3), and knowledge activation; orange marks parts used only in Studies 2 to 4. (**C**) Process statements and responses of the cognitive agent with the median effect of the manipulation on support for perceived fit in the brand-extension experiment[66] (Gemma 4 26B), in the distant-category (control) and close-category (treatment) conditions. Support for perceived fit (Study 1) ranges from −1 (the process statements contradict the statements of perceived fit) to 1 (they imply them).

We model MVCA with construal and reliance, corresponding with how human decision-makers respond to a situation as they construe it[22] and base their responses on the content of accessible thoughts[34], by ordering the processes and limiting what each process agent receives. The process agents for the appraisals receive the encoding or process statements derived from it but not the scenario, so that later process statements build on what the cognitive agent noticed rather than on the wording of the scenario. Similarly, the process agent for judgment receives the process statements of the ten processes between encoding and judgment but not the scenario, so that the cognitive agent's response is based on these process statements.

Human decision-makers use stored knowledge only when it is accessible and applicable to the situation[35], and we built this conditionality into MVCA in two ways. First, we present later process agents with the full process statement of each of six processes (marked in Table 1) only when the cognitive agent has noticed something to which that process applies, and with a brief version otherwise, a procedure that we call knowledge activation (Fig. 3B). Second, we apply three general rules, each derived from a broader psychological theory and stated identically for every experiment, which determine from what the cognitive agent has noticed whether and in what form an induced process statement shapes later process statements. Under the discounting rule, for example, an induced feeling is set aside if the cognitive agent has noticed a source of it that is unrelated to the decision (Table 1).

Whereas all cognitive agents share the same general processes, each is created from its own agent profile, which specifies an individual's demographic attributes and levels on 139 individual characteristics studied in psychology (Methods; Supplementary Table 2), and each

process agent receives descriptions only of the characteristics that, according to published research, affect its process[36,37]. Because an individual characteristic shapes behavior mainly in situations that offer cues to which the characteristic responds[37], we built individuality into MVCA through a procedure that we call trait activation. In this procedure, we present one or two first-person sentences describing each characteristic varied in Study 4 (Supplementary Table 3) to the process agent for the appraisal that the characteristic affects, but only after the cognitive agent has noticed such a cue (Fig. 3B).

To meet the condition of restricted information, we confined the system prompt of each process agent to the description of a general process derived from theories of cognition (Table 1), and we provided each process agent with only the information that its process requires (Fig. 3A). Because process agents receive descriptions of some individual characteristics, we count a mediator or an individual characteristic in the process statements as emergent when no process agent received a description of it in any form, and we call this criterion the no-description criterion (Supplementary Note 2).

Table 1 | Constitutive processes of MVCA and the theories from which they are derived. Each process is carried out in a separate LLM call whose system prompt describes the general process[24].

Process	Description	Theoretical reference
Encoding		
Encoding	Describes what the cognitive agent notices in the scenario, without evaluating it; only this process agent receives the scenario	Encoding and comprehension of social information[21]
Appraisal‡		
Categorization†	Identifies what kind of object or situation the cognitive agent encounters	Categorization as the basis of construal[21,59]
Self-relevance*	Appraises how the situation bears on the cognitive agent's self	Appraisal of relevance to the self[32,33]
Goal relevance* †	Appraises what the cognitive agent could gain or lose in the situation with respect to its current goals	Appraisal of goal relevance[33]; goal systems theory[60]
Affective response* †	States how the cognitive agent feels, following from its appraisals of self-relevance and goal relevance	Appraisal theories of emotion[32]; affect as information[42]
Causal inference* †	Infers why the situation arose and who or what is responsible	Causal attribution[61]
Social inference*	Infers what others in the situation intend or signal	Subjective construal in social inference[62]
Integration		
Alternative readings*	Considers how else the situation could be construed	Generation of alternative interpretations and motivated closure[63]
Comparison of alternatives	Compares the options that the cognitive agent noticed in the situation	Construction of value through the comparison of options[28]

Process	Description	Theoretical reference
First-person account	Summarizes, in the first person, what the cognitive agent thinks and is inclined to do	Global Workspace Theory, in which specialized processes make their results available to one another[64,65]
Coherence review	Examines whether the process statements are consistent with one another	Integration of inferences into a coherent representation[21]
Response		
Judgment	Answers the question of the experiment on the basis of the process statements of the ten processes after encoding; only this process agent receives the question	Judgment constructed from accessible and applicable process statements[34,35]

* Under knowledge activation[35], later process agents receive the full process statement only when the LLM determines that the encoding contains content to which the process applies, and otherwise a brief version.

† General rules for an induced process statement apply in categorization (standard-comparison rule[59]), goal relevance (motive–means rule[60]), and affective response and causal inference (discounting rule[42]).

‡ In trait activation (Study 4)[37], once the encoding contains a cue to which an individual characteristic responds, MVCA introduces sentences describing the characteristic into the appraisal process that it affects.

Four studies testing the propositions of a process theory

In four studies, we tested each proposition of a process theory in MVCA's cognitive agents, together with the functional property that the cognitive agents must have for this test to yield evidence for the theory (Fig. 4). We conducted the four studies in the 60 published experiments, 32 of them mostly from consumer psychology and the 28 with replication outcomes mostly from research on social cognition and decision making (Supplementary Table 1). Because we could simulate the same cognitive agent in both conditions of an experiment, we compared each cognitive agent's process statements or responses between the two conditions, as in a within-participants design. We report each effect as Cohen's d_z, the mean of these within-agent differences divided by their standard deviation across cognitive agents, and summarize each study by pooling d_z over its experiments or tests with a random-effects model (Methods). We simulated 50 cognitive agents with Gemma 4 26B (Google) and the first 30 of them also with Qwen 3.7 Flash (Alibaba) and Ling 3.0 Flash (inclusionAI), three LLMs from different developers. Unless stated otherwise, we report results averaged over the three LLMs for these 30 cognitive agents, because averaging over LLMs reduces chance variation specific to one LLM, and we state whether an effect was significant with each LLM separately (Supplementary Table 4).

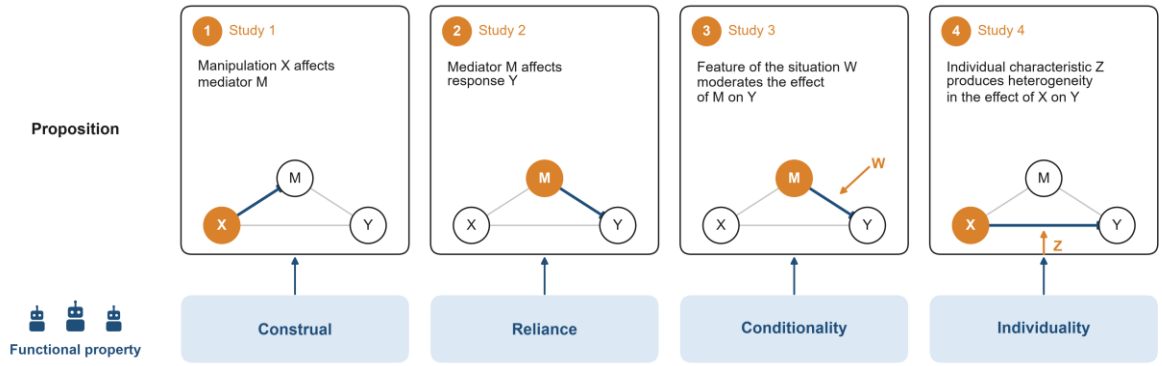


Fig. 4 | Four studies testing the propositions of a process theory and the functional properties of *Agent psychonomicus*. The top row shows the proposition of a process theory that each study tests (navy arrow, relation tested; orange, variable manipulated, induced, or varied), and the bottom row shows the functional property of the cognitive agents that this test presupposes.

Study 1. Effects of manipulations on the mediator (construal)

In Study 1, we tested the first step of the mechanism, whether the manipulation affects the mediator, and examined construal, the functional property that this test requires. Questions that measure the mediator in human participants can reveal the hypothesis[30] or influence the subsequent response[38], whereas we could measure the mediator in the process statements that preceded each cognitive agent's response after the cognitive simulations had ended (Methods). To measure the mediator, a separate LLM rated how strongly the process statements reflected it (from 1 to 7), and a natural language inference model[39,40] estimated how strongly the process statements implied, rather than contradicted, two first-person statements of the mediator, an estimate that we call support for the mediator (Fig. 3C).

On both measures, the manipulation had a positive effect on the mediator in the cognitive agents' process statements. Its effect on the ratings of the mediator was positive in 44 of the 60 experiments, significantly so in 31 (pooled $d_z = 0.77$, 95% confidence interval (CI) [0.45, 1.09]; Fig. 5A), and its effect on support for the mediator was positive in 46, significantly so in 23 (pooled $d_z = 0.44$, 95% CI [0.25, 0.63]). Both pooled effects were significant with each LLM separately (Supplementary Tables 4 and 5). The effect on support was already positive in the process statement of the appraisal process in which, according to the theory, the mediator arises ($d_z = 0.31$, 95% CI [0.17, 0.45]). The process agent for this appraisal receives what the cognitive

agent noticed but not the scenario, so this effect shows that the cognitive agents construed the situation that the manipulation altered.

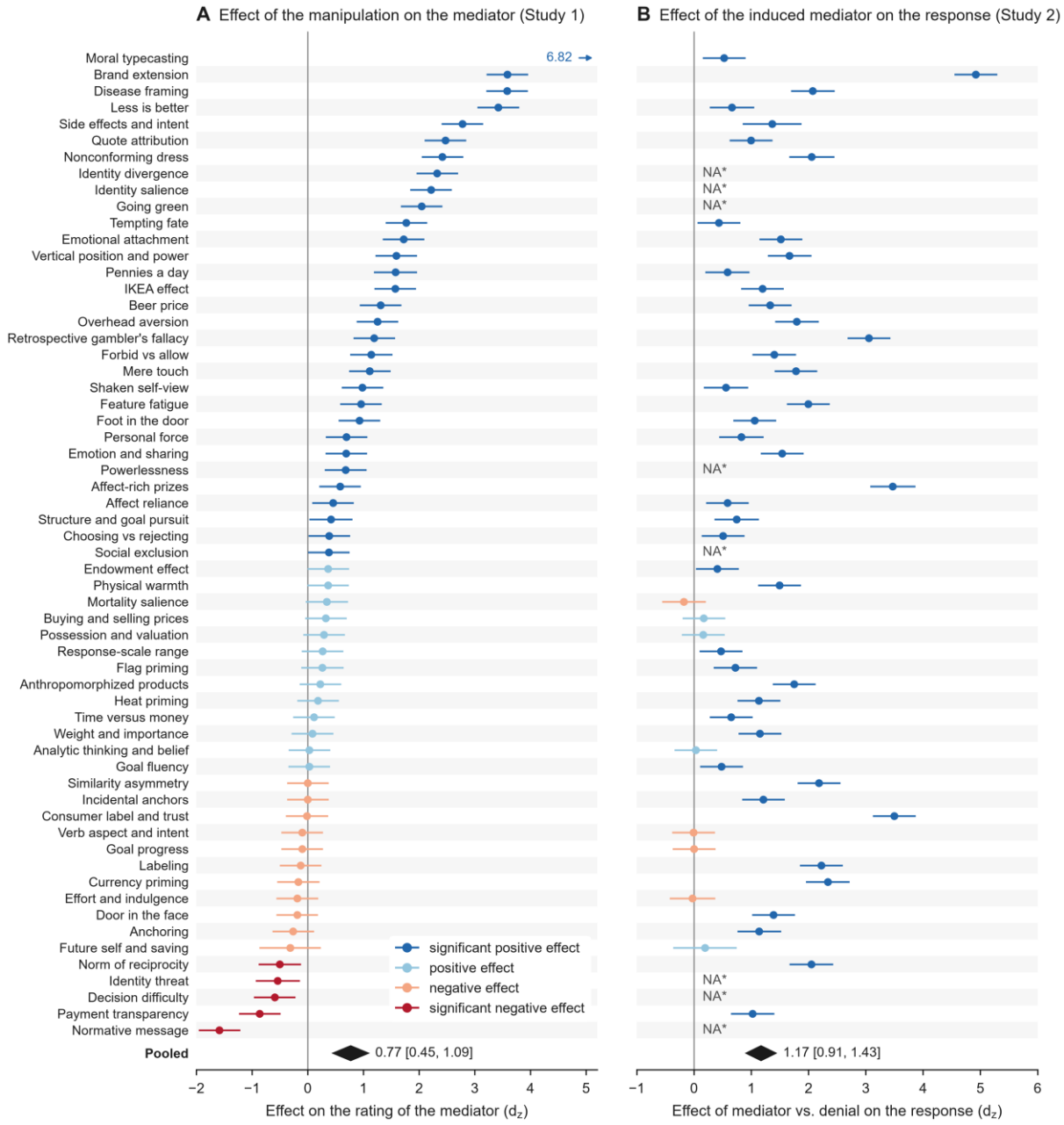


Fig. 5 | Effects of the manipulation on the mediator and of the mediator on the response (Studies 1 and 2). Points show d_z with 95% CI after averaging each cognitive agent's difference between conditions over the three LLMs (15 to 30 cognitive agents per experiment; fewer where the scenario presupposes an attribute, such as a gender, that only some agent profiles have). Dark blue and dark red mark positive and negative effects whose CI excludes zero, and light blue and light red those whose CI includes zero; diamonds show pooled random-effects estimates with Hartung–Knapp–Sidik–Jonkman 95% CIs. Experiments are ordered by the effect in A and

labeled as in Supplementary Table 1. (**A**) Effect of the manipulation on a separate LLM's rating of how strongly the process statements reflected the mediator; an arrow marks a value beyond the axis. (**B**) Effect of the induced mediator, rather than its denial, on the response. NA* marks the eight experiments whose theories predict that a situational moderator alters the mediator's effect, which we tested in Study 3 at both levels of the moderator. Results with each LLM are in Supplementary Tables 5 and 7.

The mediators also emerged from the general processes. In the 53 experiments whose mediators meet the no-description criterion, the manipulation had a positive effect on support for the mediator (pooled d_z = 0.36, 95% CI [0.17, 0.54]); in the other seven, three fixed examples that the system prompts contain identically in every experiment and condition name constructs related to the mediator (Supplementary Notes 3 and 4). Human decision-makers construct psychological states from conceptual knowledge[24], so if the mediator emerged in the same way, the process statements should also draw on the knowledge connected with the mediator and concern constructs related to it in meaning, including constructs that the LLM had not received in any form. To test this prediction, we measured the activation of a construct, the similarity in meaning between a process statement and the construct's description, with a sentence-embedding model[41] after removing the part of the process statement's meaning that it shared with the scenario (Methods). In 54 of the 60 experiments, the closer one of the 69 individual characteristics that the LLM had not received in any form was in meaning to the experiment's mediator, the more strongly it was activated in the cognitive agents' process statements (mean Spearman ρ = 0.32; Fig. 6, A and B). For 52 of the 60 experiments, this relation was weaker in the process statements of the other 59 experiments, which concern different decisions (mean ρ = 0.19; Wilcoxon P < 0.001; Fig. 6B). The experiment's own mediator, moreover, was the most strongly activated of the 60 mediators in 24 experiments, whereas random ranks would have placed it first in about one (Fig. 6C).

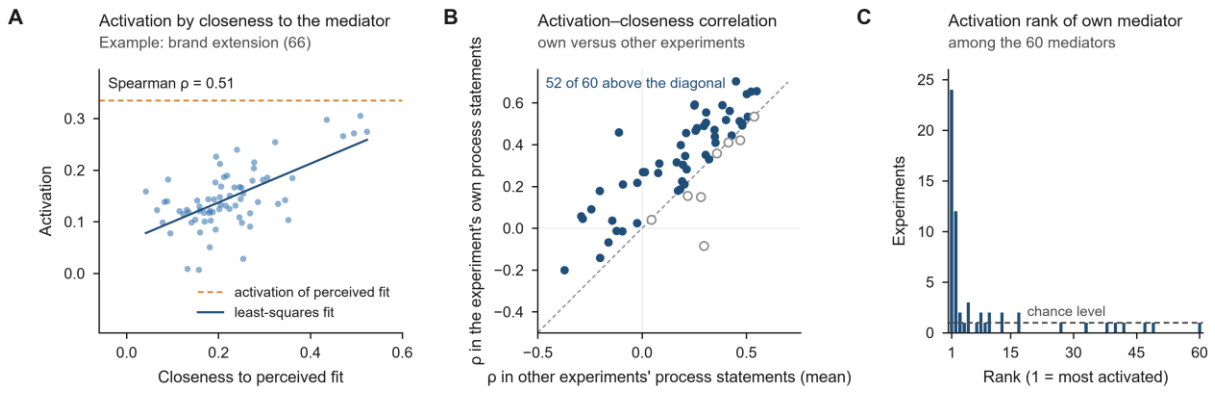


Fig. 6 | Activation of constructs in the cognitive agents' process statements (Study 1). Values are averaged over the three LLMs (15 to 30 cognitive agents per experiment). (**A** and **B**) The 69 characteristics that the LLM did not receive in any form. (**A**) Activation of each characteristic against its closeness in meaning to perceived fit in the brand-extension experiment[66]; the solid line shows the least-squares fit, and the dashed line the activation of perceived fit. (**B**) Spearman correlation (ρ) between the characteristics' closeness to an experiment's mediator and their activation, in the experiment's own process statements against the mean in the process statements of the other 59 experiments; filled points, above the diagonal, mark experiments with a higher correlation in their own process statements. (**C**) Number of experiments in which the experiment's own mediator held each rank among the 60 mediators (rank 1, most strongly activated); the dashed line shows the number expected by chance.

Study 2. Effects of induced mediators on responses (reliance)

In Study 2, we turned to the second step of the mechanism and tested whether the mediator in turn affects the response. When psychologists manipulate the mediator directly in human participants, the manipulation often alters other psychological states as well[15,23], but in MVCA we could induce the mediator as a single process statement while the scenario, the agent profile, and the system prompts stayed the same. We simulated each cognitive agent twice in the treatment condition of 52 of the 60 experiments, once with an induced process statement that states the mediator and once with one that denies it, and presented this process statement in the appraisal process in which, according to the theory, the mediator arises (Supplementary Table 1; Methods).

Inducing the mediator, rather than its denial, had a positive effect on the cognitive agents' responses in 48 of the 52 experiments, significantly so in 44 (pooled $d_z = 1.17$, 95% CI [0.91, 1.43]; Fig. 5B), and the pooled effect was significant with each LLM separately (Supplementary Table 4). In the IKEA-effect experiment, for example, the induced process statement that building a box oneself made one feel capable, rather than its denial, had a positive effect on the price that the cognitive agents were willing to pay for the box ($d_z = 1.19$). A process statement that favored the predicted response as strongly as the mediator without stating the mediator had about as large an effect on the responses as the induced mediator (Supplementary Table 7; Supplementary Note 8). Like human decision-makers, who base their responses on the content of the thoughts and feelings that are accessible at the time[35,42], the cognitive agents showed reliance, basing their responses on the content of the process statements that preceded them.

Agreement with large preregistered replications

The cognitive agents also distinguished the original effects that replicated in large preregistered replications from those that did not. These replications had re-examined the original effects of 28 of the experiments[43–50]. The cognitive agents showed the original effect, a significant difference in the response between the treatment and the control condition in the published direction, in 9 of the 14 experiments whose effects replicated but in only 3 of the 14 whose effects did not (one-sided Fisher's exact test, P = 0.03; Fig. 7A). Pooled within each group, the manipulation's effect on the response was larger where the original effect had replicated (d_z = 0.95, 95% CI [0.16, 1.75]) than where it had not (d_z = 0.07, 95% CI [−0.10, 0.24]; difference, P = 0.02).

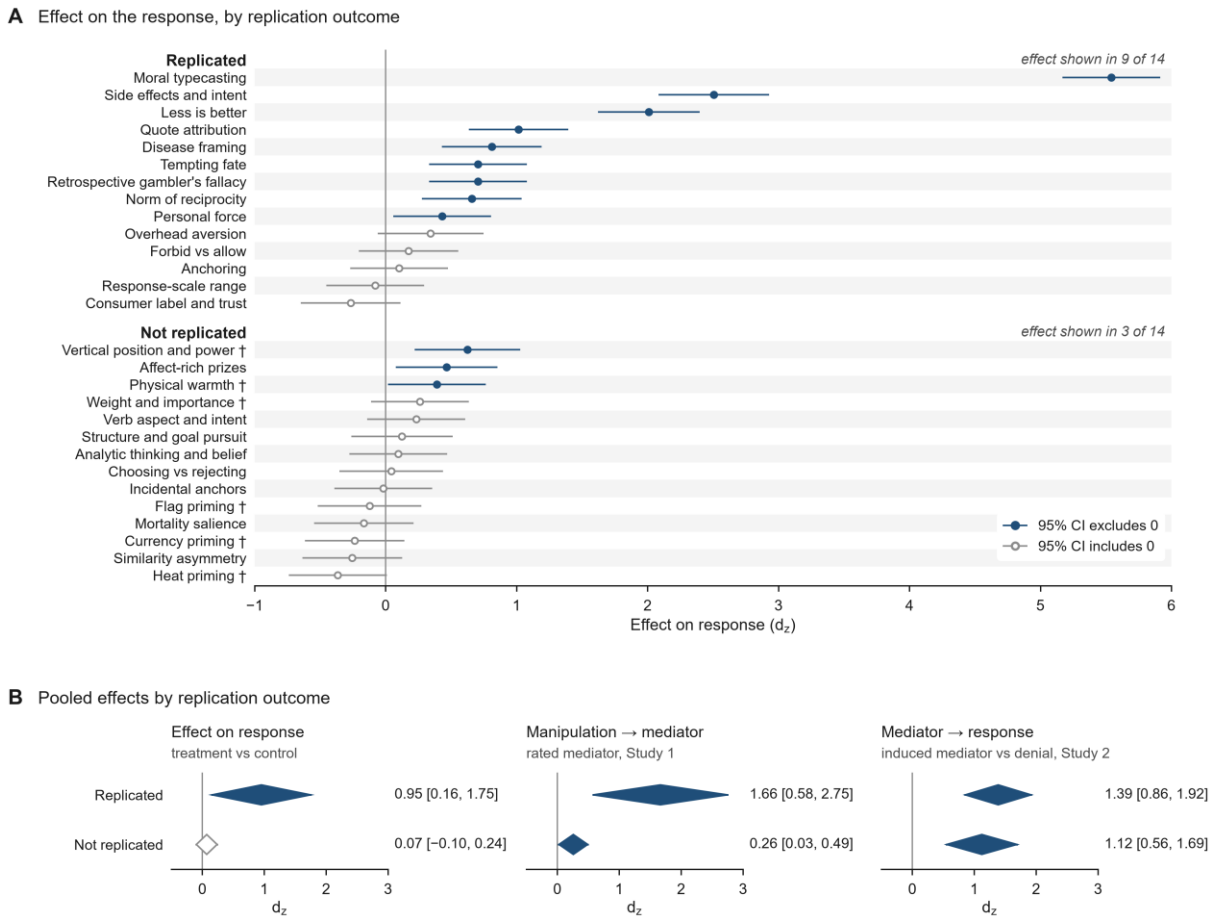


Fig. 7 | Effects in the cognitive agents compared with the outcomes of large preregistered replications. The 28 experiments whose original effects were re-examined in large preregistered replications, in which 14 effects replicated and 14 did not. Values are averaged over the three LLMs (30 cognitive agents). (**A**) Effect of the manipulation on the response (d_z, 95% CI), ordered by size within each replication outcome. Filled points mark a CI that excludes 0, which we counted as showing the original effect, and † marks a manipulation that is an incidental cue.

(**B**) Pooled random-effects d_z (95% CI) by replication outcome for the effects of the manipulation on the response and on the rating of the mediator and for the effect of the induced mediator, rather than its denial, on the response; filled diamonds mark a CI that excludes 0. Results for each experiment are in Supplementary Table 8.

Because Studies 1 and 2 tested the two steps of the mechanism separately, we could trace this difference to the first step, the manipulation's effect on the mediator. Where the original effect had replicated, the manipulation had a positive effect on the ratings of the mediator (d_z = 1.66, 95% CI [0.58, 2.75]); where it had not, this effect was small (d_z = 0.26, 95% CI [0.03, 0.49]; difference, P = 0.006; Fig. 7B). For example, none of the original effects replicated in the six experiments whose theories propose that an incidental cue, such as a faint image of banknotes behind a questionnaire, evokes the mediator without playing any role in the decision, and in four of these experiments, the manipulation had no significant positive effect on either measure of the mediator (Supplementary Table 8). Induced mediators, by contrast, had a positive effect on the responses in both groups (d_z = 1.39 where the effect had replicated and 1.12 where it had not; difference, P = 0.46; Fig. 7B), so the second step of the mechanism, the mediator's effect on the response, held whether or not the original effect had replicated.

Study 3. Situational moderation of the mediator's effect (conditionality)

In Study 3, we tested whether a feature of the situation moderates the mediator's effect on the response in the 10 experiments whose theories propose such moderation, eight of which we had not tested in Study 2 (11 tests, two of them in the social-exclusion experiment; Supplementary Table 9). According to these theories, the situational moderator alters the mediator's effect on the response rather than the manipulation's effect on the mediator[51], so we used a moderation-of-process design, in which we induced the mediator as in Study 2 at both levels of the situational moderator.

On average across the 11 tests, the situational moderators altered the mediator's effect on the response as the theories predict (pooled interaction d_z = 0.31, 95% CI [0.04, 0.57]), with an interaction that was positive in 8 of the 11 tests and significant in five (Fig. 8A; Supplementary Tables 4 and 9). With each LLM separately, the pooled interaction was positive (Supplementary Table 4), and with Gemma 4 26B, five of the 11 interactions were significant in the predicted direction and none in the opposite direction, more than chance would produce (permutation test, P < 0.0001; Supplementary Note 11).

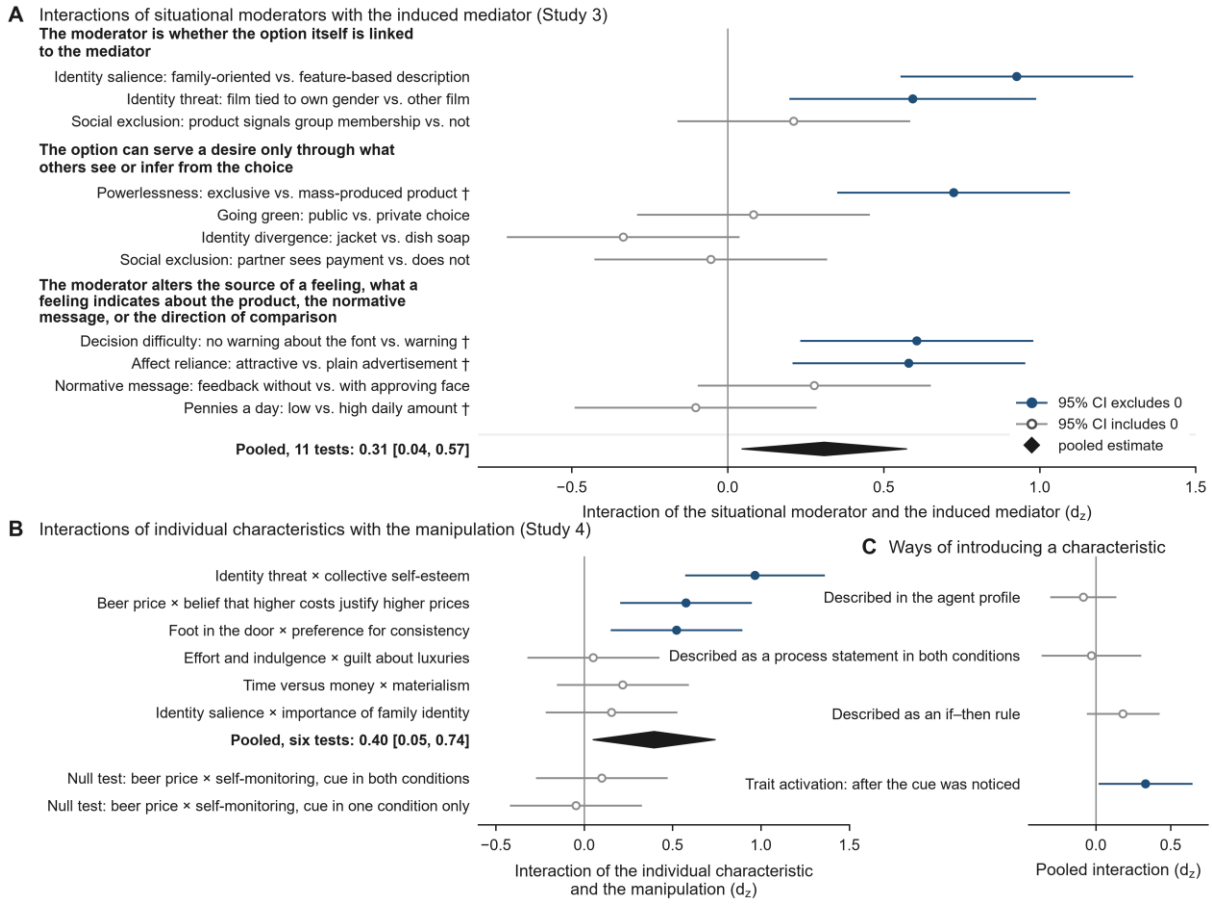


Fig. 8 | Interactions with situational moderators and individual characteristics (Studies 3 and 4). Points and bars show d_z and 95% CI; filled points mark a CI that excludes zero, and diamonds show pooled random-effects estimates. (**A** and **B**) Values are averaged over the three LLMs (27 to 30 cognitive agents per test); results with each LLM are in Supplementary Tables 3, 4 and 9. (**A**) Effect of the induced mediator, rather than its denial, at the level of the situational moderator named first, for which the theory predicts the larger effect, minus that at the other level. Headings group the tests by how, according to the theory, the situational moderator makes the mediator relevant to the decision (Supplementary Table 9), and † marks the tests in which a general rule of MVCA applied (Table 1). (**B**) Effect of the manipulation at the level of the individual characteristic for which the larger effect is predicted minus that at the other level (Supplementary Table 3). The two null tests cross the beer-price manipulation with self-monitoring, for which no interaction is predicted. (**C**) Pooled interaction for four ways of introducing a characteristic, in cognitive agents created from other agent profiles (Gemma 4 26B; Supplementary Note 12).

Because the interactions differed widely between tests (with Gemma 4 26B, $I^2 = 84\%$), we examined the cognitive agents' process statements to understand why. Where the situational moderator altered the mediator's effect, the cognitive agents had noticed a feature that made the

mediator relevant to the decision, either an option that was itself linked to the mediator or a feature to which one of the three general rules applied (Supplementary Notes 9 and 10). In the identity-salience test, for example, the induced process statement that family is central to who one is had a positive effect on the likelihood of buying a product designed for staying in touch with family but not on that of buying the same product described by its features (interaction d_z = 0.93). In the powerlessness test, the situational moderator was whether the product was exclusive or mass-produced, and the induced desire to feel more powerful shaped the cognitive agents' later process statements only when the LLM determined from their encoding that the product would serve this desire (d_z = 0.72; Supplementary Note 20).

The cognitive agents therefore showed conditionality where what they had noticed made an induced process statement relevant to the decision. In three tests whose theories assume that human decision-makers take into account what others will conclude from their choice[52,53], an inference that MVCA does not carry out, the predicted interaction was not significant with any LLM (Supplementary Table 9; see Discussion).

Study 4. Individual heterogeneity in the manipulation's effect (individuality)

In Study 4, we varied the cognitive agents themselves rather than the situation and tested whether individual characteristics produce heterogeneity in the manipulation's effect, as they should if the cognitive agents show individuality. We conducted six tests, each of which crossed the manipulation of one experiment with one individual characteristic, such as identity threat with collective self-esteem[54] and the place where a beer is bought with the belief that higher costs justify higher prices[55]. According to published research or the theory of the experiment, the manipulation should affect human decision-makers more strongly at one level of the characteristic than at the other (Supplementary Table 3). We simulated each cognitive agent with the same agent profile at a high and a low level of the characteristic, so that the two cognitive simulations differed only in the sentences, introduced through trait activation, that described this characteristic.

As predicted, the level of the individual characteristic that the sentences described produced heterogeneity in the manipulation's effect, with an interaction that was positive in all six tests and significant in three (pooled d_z = 0.40, 95% CI [0.05, 0.74]; Fig. 8B), and the pooled interaction was significant with each LLM except Ling 3.0 Flash (Supplementary Tables 3 and

4). The heterogeneity was specific to the characteristic that the sentences described. Sentences about self-monitoring[56], a characteristic that theories of price fairness do not invoke, produced no significant interaction with any LLM, whether MVCA introduced them in response to the same cue as in the beer-price test, a fancy or upscale place of purchase, or in response to a different cue (two null tests; Supplementary Table 3).

When we presented the description of a characteristic also mattered. In cognitive agents created from other agent profiles, heterogeneity was significant when the description followed a cue that the cognitive agent had noticed, as in trait activation. It was not significant with any of three procedures that presented the description regardless of such cues, one of which described the characteristic in the agent profile, as silicon samples and digital twins do[5,9] (Fig. 8C; Supplementary Notes 12 and 13).

Identification test for recognition of the published experiments

Finally, we tested whether restricted information leaves the LLM little basis for recognizing the published experiments[29]. We asked a separate LLM which published experiment, if any, the information for each process came from, presenting this information exactly as Gemma 4 26B had received it in Study 1 and, for comparison, the scenario and the question, the materials that an LLM receives in silicon samples and digital twins (Methods).

The separate LLM identified the experiment in 38% of its answers from the scenario and the question and in 37% from the information for encoding, which contains the scenario, but less often for every later process (median, 18% across the ten processes between encoding and judgment; 16% for judgment; Fig. 9A; Supplementary Table 11). Restricted information thus had a negative effect on identification in the processes after encoding.

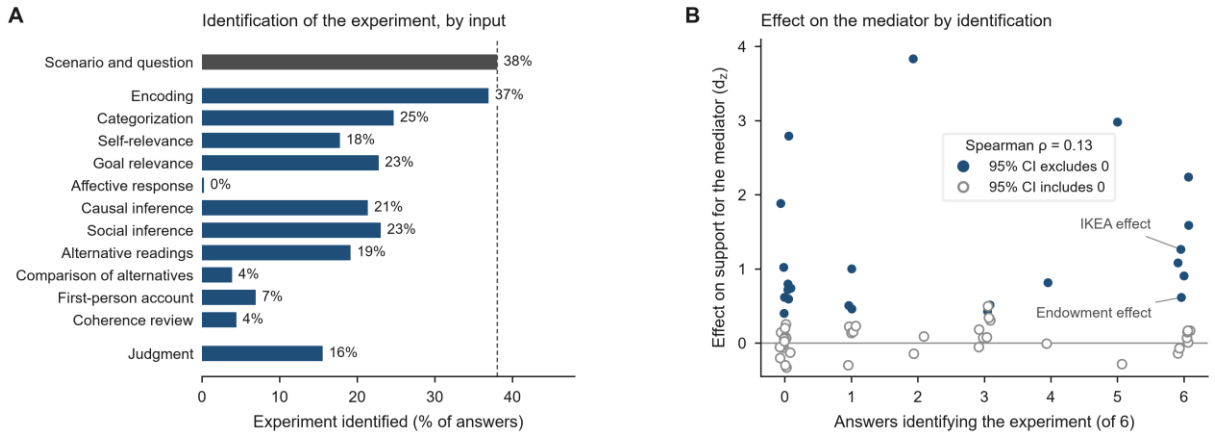


Fig. 9 | Identification of the published experiments by a separate LLM. A separate LLM (DeepSeek V4.1 Flash) named, if it could, the published experiment from the information that Gemma 4 26B received for each process in Study 1 and from the scenario and the question (360 answers each). (**A**) Percentage of answers that identified the experiment, by process; the gray bar and the dashed line show the percentage for the scenario and the question. (**B**) Effect of the manipulation on support for the mediator (d_z, averaged over the three LLMs) against the number of answers, of six, that identified the experiment from the scenario and the question. Points show experiments, jittered horizontally; filled points mark a 95% CI that excludes zero, and labeled points mark two of the 13 experiments identified in all six answers.

The positive effect of the manipulation on support for the mediator did not depend on identification. Across the 60 experiments, this effect was not significantly correlated with how often the separate LLM identified an experiment from the scenario and the question (Spearman $\rho = 0.13$, $P = 0.33$; Fig. 9B), and it was also about the same size in the 23 experiments that the separate LLM did not identify in any answer (pooled $d_z = 0.37$, 95% CI [0.07, 0.66]) as in the 37 experiments identified at least once (pooled $d_z = 0.49$, 95% CI [0.22, 0.75]; Supplementary Table 11).

Discussion

With each of the three LLMs, the manipulation had a positive effect on the mediator in the process statements that preceded the response of MVCA's cognitive agents, and the induced mediator had a positive effect on the response, so the cognitive agents showed construal and reliance. Like human decision-makers, who act on accessible thoughts only to the degree that these thoughts are applicable to the decision[35], the cognitive agents showed conditionality when what they had noticed made the mediator relevant to the decision. Their individuality appeared when the manipulation provided a cue to which an individual characteristic responds, which is where trait activation theory expects individual characteristics of human decision-makers to shape their responses[37].

Beyond these four functional properties, the theorized psychological mechanisms emerged in the cognitive agents' process statements from the general processes (Study 1 and the identification test). According to psychological constructionism, human decision-makers construct psychological states from conceptual knowledge acquired through experience[24], and we propose that the LLM, in carrying out a general process, similarly draws on everyday knowledge contained in the text produced by humans on which it was trained. If the LLM were

fine-tuned on the responses of human participants in published experiments, as models designed to reproduce these responses are[57], it would instead receive the findings at which the cognitive agents should arrive by construing the situation. MVCA requires no such responses, because its processes are derived from general theories of cognition and its LLMs are used without fine-tuning.

Because the mediators emerge from the general processes, the cognitive agents should show a predicted effect only if the general processes produce every step of the mechanism that the theory proposes. The cognitive agents therefore rarely showed the original effects that had failed to replicate in large preregistered replications, whereas silicon samples showed significant effects in most scenario-based experiments in which the original studies had found none in human participants[4]. Where an original effect had not replicated, the theory failed in the cognitive agents mainly at the first step of the mechanism. Researchers can thus test a theory in cognitive agents one step of the mechanism at a time before they conduct an experiment or a large replication with human participants, beginning with whether the manipulation has the predicted effect on the mediator, and locate the step at which the theory's prediction fails.

When a predicted moderation does not appear in the cognitive agents, researchers can also identify an inference on which the theory relies without stating it. Theories of signaling and self-presentation, for example, predict that the effect of a desire on choice depends on whether others can observe the choice[52,53], but in the cognitive agents, this situational moderator did not significantly alter the effect of the desire (Study 3). These theories require an inference about what observers will conclude from one's own choice, whereas in MVCA's process of social inference, the cognitive agent infers only what others in the situation intend or signal. We therefore attribute the absence of the predicted moderation to this missing inference.

We see the most immediate use of MVCA in experiments that have not yet been conducted. If researchers simulate such an experiment before they run it and then measure the mediator both in the cognitive agents and in the human participants, they can compare the step of the mechanism at which a theory fails in the cognitive agents with the step at which it fails in human participants.

Methods

Study design and prespecified analyses

The objective of this research was to examine whether researchers can test, in the cognitive agents of MVCA, the four propositions of a process theory: that the manipulation affects the mediator, that the mediator affects the response, that a situational moderator alters the mediator's effect, and that an individual characteristic produces heterogeneity in the manipulation's effect. We conducted Studies 1 to 4, which tested these propositions in turn, in 60 influential published experiments in social and consumer psychology (Supplementary Table 1; Supplementary Note 21). We also examined whether the mediators emerged from the general processes and whether the results agreed with the outcomes of large preregistered replications where these existed. The research involved no human participants or animals and used no human data other than published results.

Before conducting the studies reported here, we deposited the specification of MVCA and of the primary analyses in a time-stamped repository (7 October 2026, repository version f1ae3f2). For each experiment, the specification defined the appraisal process in which, according to the theory, the mediator arises and the general rule, if any, that applies to it; it also defined the cognitive agents of the studies (see Agent profiles), the designs, the primary analyses, and the criteria for each study. The criteria, which served as prespecified hypotheses, concerned the results with Gemma 4 26B; in every study, one criterion was a pooled effect whose 95% confidence interval (CI) excludes zero, and Studies 1, 2, and 4 had further criteria (Supplementary Note 23). With Gemma 4 26B, Studies 1, 2, and 4 met these criteria in the 32 experiments that the specification comprised (Supplementary Note 23), whereas Study 3 did not, because the CI of its positive pooled interaction included zero (Supplementary Table 4).

For the other 28 of the 60 experiments, large preregistered replications had re-examined the original effect; 14 effects replicated, and 14 did not[43–50] (Supplementary Note 18). We included these 28 experiments in Studies 1 and 2, which require neither a situational moderator nor an individual characteristic. Before conducting any cognitive simulation in them, we prepared their materials by the procedure used for the other 32 experiments (see Experiments and materials), assigned each mediator to the appraisal process in which, according to the theory, the mediator arises, following the assignment rule in the deposited specification, and deposited these materials and assignments in time-stamped files (13 October 2026). After conducting the

cognitive simulations in 10 of these 28 experiments and before conducting any in the other 18, we deposited, in the same way, the analysis plan comparing the experiments whose effects replicated with those whose effects did not, so that this plan was prespecified for these 18 experiments.

The following cognitive simulations and analyses were post hoc and were not part of the specification of 7 October. We conducted the cognitive simulations with Qwen 3.7 Flash and Ling 3.0 Flash and computed the estimates averaged over LLMs. In Study 1, we conducted the construct-activation analysis and computed support for the mediator and the mediator-minus-denial score in the process statement of the appraisal process in which, according to the theory, the mediator arises. In addition, we conducted the identification test, the search for names of published findings (Supplementary Table 12; Supplementary Note 17), and the analysis of Study 1 without the experiments in which fixed example words in the system prompts name a construct related to the mediator (Supplementary Note 3). In Studies 3 and 4, we repeated the tests that had not reached significance, conducted the joint permutation test of Study 3, estimated the pooled interaction of Study 4 without the beer-price test, and computed the correlation of individual cognitive agents' effects between the two levels of self-monitoring in a null test of Study 4. Finally, among the 28 experiments with replication outcomes, we examined as a group the experiments in which the manipulation is an incidental cue, one that, according to the theory, evokes the mediator without playing any role in the decision.

Experiments and materials

In each experiment, the original authors manipulated a factor and measured a response, namely a judgment, a choice, or an amount such as willingness to pay, and the original authors or later researchers proposed a mediator that explains the finding. For each experiment, we prepared materials that follow the original procedure: the scenario of each condition, phrased in the second person, and the question with its response options. To measure the mediator, we formulated two first-person statements of the mediator from the original authors' measure or description of it (the mediator statements; Supplementary Table 1), two statements that deny the mediator (the denial statements), and a definition of the mediator for a separate LLM that rated the process statements (see Language models).

The materials also contain the process statements that we induced in Studies 2 and 3, including those with which we compared the induced mediator in Study 2, such as a pleasant mood (Supplementary Tables 1 and 9; Supplementary Note 8), and, for Study 4, the cue and the sentences that describe each level of the individual characteristic (Supplementary Table 3). We compared each set of materials with the original article and revised every set that departed from the original procedure or whose scenario revealed the mediator or the predicted response.

Agent profiles

Each of the 150 agent profiles describes a simulated individual by age, gender, occupation, city, Big Five personality traits, and a short biography. The individuals are adults aged 22 to 74 (75 men, 64 women, and 11 non-binary adults), most of whom live in US cities. In addition, we specified each individual's 139 individual characteristics (Supplementary Table 2), namely a score from 0 to 1 on each of 124 numeric characteristics and 15 characteristics that are short texts. To compile the list of these characteristics and, for each, the kinds of processes that published research links to it, we used an LLM (Claude), whose training data include published research. We drew most numeric levels from published population norms or from normal distributions, and an LLM (gpt-4o-mini) rated ten beliefs and composed some of the short texts from the biography (Supplementary Note 4). The cognitive agents of the four studies were created from 50 of the 150 profiles (Supplementary Note 7).

Cognitive model

For each process before judgment (Fig. 3A), the LLM states the result of the process as a process statement, a short label followed by first-person sentences (Fig. 3B). Before later process agents receive a process statement, MVCA removes all numbers from it and withholds the parts that resemble a response, so that the process agent for judgment cannot adopt them as the response (Supplementary Note 3).

To induce a process statement, such as one that states the mediator or one that denies it, we presented it, framed as "a thought on your mind right now," among the earlier process statements that the LLM receives for the appraisal process in which, according to the theory, the mediator arises. Every later process agent that receives the process statement of this appraisal also receives the induced process statement, unless a general rule withholds it. In Study 4, MVCA presents the sentences that describe the level of the individual characteristic in the same

way, in the appraisal process that, according to published research, the characteristic affects (Supplementary Table 3).

To determine whether knowledge activation, a general rule, or trait activation applies, the LLM answers, in a separate LLM call, a question about the cognitive agent's encoding; an answer counts only if the LLM supports it with a passage of at least two words that appears verbatim in the encoding. Depending on the answers in knowledge activation and on 17 individual characteristics that affect how extensively individuals think about a situation but are not described to the LLM, later process agents receive the process statement of a process in full, as a brief version consisting of its label and first sentence, or not at all (Supplementary Table 2; Supplementary Note 3).

The general rules apply only to process statements induced in particular appraisal processes (Table 1). If the encoding does not show that choosing the option would serve an induced desire (motive–means rule), or if it mentions a source of an induced feeling or impression that is unrelated to the decision (discounting rule), MVCA withholds the induced process statement from the process agent for the appraisal and from every later process agent. Under the standard-comparison rule, when an induced process statement compares the option with a standard, MVCA attaches to this process statement the LLM's conclusion that the option falls within this standard or clearly outside it (Supplementary Note 3).

Language models

In all four studies, Gemma 4 26B (Google) carried out the processes of all 50 cognitive agents, and Qwen 3.7 Flash (Alibaba) and Ling 3.0 Flash (inclusionAI) carried out those of the first 30. All three LLMs, which we used without fine-tuning, received identical materials, agent profiles, system prompts, and rules (Supplementary Table 13; Supplementary Note 15). For every LLM call of a cognitive simulation, the temperature was 0.2, and we disabled the optional step-by-step reasoning mode. The ratings of the mediator in Study 1 and the answers in the identification test came from a separate LLM, DeepSeek V4.1 Flash (temperature 0). We computed the measures based on natural language inference with the model cross-encoder/nli-deberta-v3-large[39] and construct activation with two sentence-embedding models[41].

Across Studies 1 to 4, we conducted 57,147 cognitive simulations (Study 1, 12,983; Study 2, 28,772; Study 3, 10,716; Study 4, 4,676), 28,542 with Gemma 4 26B, 14,304 with

Qwen 3.7 Flash, and 14,301 with Ling 3.0 Flash. The four studies required 792,337 LLM calls in all: 742,911 for the 12 processes and knowledge activation, 16,610 for the general rules and trait activation, 25,966 for the ratings by the separate LLM, and 6,850 for the identification test. The cognitive simulations produced 628,617 first-person process statements. In the 28,197 cognitive simulations whose process statements we counted (all of Study 1 and the Study 2 cognitive simulations of the 28 experiments with replication outcomes), a process statement contained 109 words on average and a cognitive simulation 1,198 words.

Measures

In Studies 1 to 4, the response was the choice, rating, or amount that the process agent for judgment stated in reply to the original question (Supplementary Note 21), and we signed each within-agent difference so that positive values indicate the predicted direction.

In Study 1, support for the mediator was the probability that a process statement implied a mediator statement minus the probability that it contradicted the statement. We averaged it over the two statements and over the process statements of the ten processes between encoding and judgment, and we also computed it in the process statement of the appraisal process in which, according to the theory, the mediator arises (Supplementary Table 5). In addition, the separate LLM, which received neither the scenario nor the condition, rated, once for each cognitive simulation, how strongly the complete process statements of these ten processes reflected the mediator as defined in the materials. For a criterion of the deposited specification, we also computed, for each process statement, the probability that it implied the mediator statements minus the probability that it implied the denial statements, a score that we call the mediator-minus-denial score (Supplementary Note 5).

Construct activation

We computed the activation of a construct (one of the 139 individual characteristics or one of the 60 mediators) in a cognitive simulation as the mean cosine similarity between the representation of the construct's description (Supplementary Tables 1 and 2) and the representations of the process statements of the ten processes between encoding and judgment, computed with all-mpnet-base-v2 after removing from each process statement's representation its projection onto the scenario's representation (Supplementary Note 6). A construct's closeness in

meaning to an experiment's mediator was the cosine similarity between their descriptions, computed with a different model, all-MiniLM-L6-v2.

For each experiment, we computed the Spearman correlation between the closeness of the characteristics to the experiment's mediator and their activation in the experiment's own process statements, averaged over all of its cognitive simulations, and the mean of the corresponding correlations in the process statements of the other 59 experiments; across the 60 experiments, we compared these two correlations with a Wilcoxon signed-rank test. To meet the no-description criterion, we based these analyses on the 69 characteristics that the LLM did not receive in any form (Supplementary Note 4), and we repeated them with all 139 characteristics (Supplementary Table 6). We also ranked each experiment's own mediator among the 60 mediators by its standardized activation in the experiment's process statements (Fig. 6C), and we computed the manipulation's effect on the activation of the experiment's own mediator, both alone and minus the mean activation of the other 59 mediators (Supplementary Table 6).

Identification test

We asked DeepSeek V4.1 Flash to name the published experiment, if any, from which an input came, and presented it with the exact input that Gemma 4 26B had received for each of the 12 processes, in both conditions, in the Study 1 cognitive simulations of the first three cognitive agents that took part in each experiment and, for comparison, with the scenario followed by the question (4,680 inputs). An answer counted as identifying the experiment only if DeepSeek V4.1 Flash, in a second LLM call, judged the experiment that it had named to be the actual one (Supplementary Note 17).

Statistical analysis

For each experiment and comparison, we computed Cohen's d_z from the differences within each cognitive agent, with the 95% CI $d_z \pm t(0.975, n - 1)/\sqrt{n}$, where n is the number of cognitive agents. The standard deviation of these differences reflects both differences between cognitive agents and chance variation from one cognitive simulation to the next (Supplementary Note 16). We pooled the d_z values of the experiments or tests of each study with a random-effects model estimated by restricted maximum likelihood, with $1/n + d_z^2/(2n)$ as the sampling variance of d_z, and computed CIs with the Hartung–Knapp–Sidik–Jonkman method[58]. We call an effect significant when its two-sided 95% CI excludes zero ($\alpha = 0.05$); because the primary test

of each study is its pooled estimate, we did not adjust the intervals for single experiments for multiple comparisons. We describe the variation among the interactions of Study 3 with I^2, the share of the variation that sampling error does not explain (Supplementary Note 22). To test whether the number of significant interactions in Study 3 with Gemma 4 26B exceeded what chance would produce, we used a joint sign-flip permutation test that preserves the dependence between tests that share cognitive agents (Supplementary Note 11).

For the 28 experiments with replication outcomes, we followed the deposited analysis plan (Supplementary Notes 18 and 19). We counted the cognitive agents as showing an experiment's original effect when the 95% CI of d_z for the treatment-minus-control difference in the response, averaged over the three LLMs, lay above zero, and we tested whether showing the effect was associated with the replication outcome with a one-sided Fisher's exact test, because the deposited plan stated the directional prediction that the cognitive agents would show the effect where it had replicated rather than where it had not. Separately for the experiments whose effects replicated and for those whose effects did not, we pooled the effects of the manipulation on the response, on support for the mediator, and on the ratings of the mediator (Study 1), and the effect of the induced mediator rather than its denial on the response (Study 2), and we tested the difference between the two pooled estimates with a two-sided z test, $z = (\hat{d}_1 - \hat{d}_2)/\sqrt{(SE_1^2 + SE_2^2)}$.

The estimates averaged over LLMs and the comparisons between LLMs use the 30 cognitive agents that all three LLMs simulated, whereas the estimates with Gemma 4 26B in Supplementary Table 4 use all 50; for the averaged estimates, we averaged each cognitive agent's difference or interaction, or each construct's activation, over the three LLMs (Supplementary Note 15). We conducted the tests of Studies 3 and 4 that had not reached significance twice more with Gemma 4 26B and the same cognitive agents and materials, to assess how much their results varied by chance (Supplementary Table 10; Supplementary Note 14). The primary pooled estimates are based on the first time each test was conducted; for the repeated tests, we also averaged each cognitive agent's interaction over the three times before computing d_z.

Use of artificial intelligence

Besides the LLMs that are the object of this research (Supplementary Table 13) and those used to construct the agent profiles (see Agent profiles), the authors used an AI assistant (Claude,

Anthropic) to help develop the analysis code and to draft and edit the manuscript; the authors checked all code, numbers, and text and take full responsibility for the content.

Data availability

The scored data of each cognitive agent that underlie all results and figures, the experimental materials, and the agent profiles will be released on GitHub and archived in Zenodo with a DOI upon publication. Researchers can conduct the cognitive simulations again, with complete process statements, using the released code and their own access to the LLM provider.

Code availability

The code of MVCA, together with the specification deposited before the reported studies were conducted (repository version f1ae3f2), the system prompts, and the analysis and figure scripts, will be released on GitHub under the MIT license and archived in Zenodo with a DOI upon publication.

References

1. Peck, J. & Shu, S. B. The effect of mere touch on perceived ownership. *J. Consum. Res.* **36**, 434–447 (2009).

2. Pierce, J. L., Kostova, T. & Dirks, K. T. The state of psychological ownership: integrating and extending a century of research. *Rev. Gen. Psychol.* **7**, 84–107 (2003).

3. Aher, G. V., Arriaga, R. I. & Kalai, A. T. Using large language models to simulate multiple humans and replicate human subject studies. In *Proc. 40th International Conference on Machine Learning* Vol. 202, 337–371 (PMLR, 2023).

4. Cui, Z., Li, N. & Zhou, H. A large-scale replication of scenario-based experiments in psychology and management using large language models. *Nat. Comput. Sci.* **5**, 627–634 (2025).

5. Argyle, L. P. et al. Out of one, many: using language models to simulate human samples. *Polit. Anal.* **31**, 337–351 (2023).

6. Horton, J. J. Large language models as simulated economic agents: what can we learn from Homo silicus? Preprint at https://arxiv.org/abs/2301.07543 (2023).

7. Ashokkumar, A., Hewitt, L., Ghezae, I. & Willer, R. Large language models can predict the results of social science experiments. *Nature* **656**, 115–122 (2026).

8. Brand, J., Israeli, A. & Ngwe, D. Using GPT for market research. Preprint at https://doi.org/10.2139/ssrn.4395751 (2023).

9. Park, J. S. et al. Generative agent simulations of 1,000 people. Preprint at https://arxiv.org/abs/2411.10109 (2024).

10. Peng, T. et al. Digital twins are funhouse mirrors: five systematic distortions. *Sci. Adv.* **12**, eaeh8260 (2026).

11. Dillion, D., Tandon, N., Gu, Y. & Gray, K. Can AI language models replace human participants? *Trends Cogn. Sci.* **27**, 597–600 (2023).

12. Baron, R. M. & Kenny, D. A. The moderator–mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.* **51**, 1173–1182 (1986).

13. Zhao, X., Lynch Jr., J. G. & Chen, Q. Reconsidering Baron and Kenny: myths and truths about mediation analysis. *J. Consum. Res.* **37**, 197–206 (2010).

14. Spencer, S. J., Zanna, M. P. & Fong, G. T. Establishing a causal chain: why experiments are often more effective than mediational analyses in examining psychological processes. *J. Pers. Soc. Psychol.* **89**, 845–851 (2005).

15. Pirlott, A. G. & MacKinnon, D. P. Design approaches to experimental mediation. *J. Exp. Soc. Psychol.* **66**, 29–38 (2016).

16. Bolger, N., Zee, K. S., Rossignac-Milon, M. & Hassin, R. R. Causal processes in psychology are heterogeneous. *J. Exp. Psychol. Gen.* **148**, 601–618 (2019).

17. Bryan, C. J., Tipton, E. & Yeager, D. S. Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nat. Hum. Behav.* **5**, 980–989 (2021).

18. Brockers, V. C., Ehrlich, D. A. & Priesemann, V. Disentangling interaction and bias effects in opinion dynamics of large language models. *Nat. Commun.* **17**, 10077 (2026).

19. Lakatos, I. Falsification and the methodology of scientific research programmes. In *Criticism and the Growth of Knowledge* (eds Lakatos, I. & Musgrave, A.) 91–196 (Cambridge Univ. Press, Cambridge, 1970).

20. Meehl, P. E. Appraising and amending theories: the strategy of Lakatosian defense and two principles that warrant it. *Psychol. Inq.* **1**, 108–141 (1990).

21. Wyer Jr., R. S. & Srull, T. K. Human cognition in its social context. *Psychol. Rev.* **93**, 322–359 (1986).

22. Ross, L. & Nisbett, R. E. *The Person and the Situation: Perspectives of Social Psychology* (McGraw-Hill, New York, 1991).

23. Bullock, J. G., Green, D. P. & Ha, S. E. Yes, but what's the mechanism? (Don't expect an easy answer). *J. Pers. Soc. Psychol.* **98**, 550–558 (2010).

24. Barrett, L. F. The future of psychology: connecting mind to brain. *Perspect. Psychol. Sci.* **4**, 326–339 (2009).

25. Sumers, T. R., Yao, S., Narasimhan, K. & Griffiths, T. L. Cognitive architectures for language agents. *Trans. Mach. Learn. Res.* (2024).

26. Webb, T., Mondal, S. S. & Momennejad, I. A brain-inspired agentic architecture to improve planning with LLMs. *Nat. Commun.* **16**, 8633 (2025).

27. Cacioppo, J. T., von Hippel, W. & Ernst, J. M. Mapping cognitive structures and processes through verbal content: the thought-listing technique. *J. Consult. Clin. Psychol.* **65**, 928–940 (1997).

28. Johnson, E. J., Häubl, G. & Keinan, A. Aspects of endowment: a query theory of value construction. *J. Exp. Psychol. Learn. Mem. Cogn.* **33**, 461–474 (2007).

29. Ludwig, J., Mullainathan, S. & Rambachan, A. Large language models: an applied econometric framework. Preprint at https://arxiv.org/abs/2412.07031 (2024).

30. Orne, M. T. On the social psychology of the psychological experiment: with particular reference to demand characteristics and their implications. *Am. Psychol.* **17**, 776–783 (1962).

31. Karetnikov, N., Rahwan, I. & Svetinovic, D. Large language models as human proxies. *Nat. Comput. Sci.* https://doi.org/10.1038/s43588-026-01060-3 (2026).

32. Smith, C. A. & Ellsworth, P. C. Patterns of cognitive appraisal in emotion. *J. Pers. Soc. Psychol.* **48**, 813–838 (1985).

33. Moors, A., Ellsworth, P. C., Scherer, K. R. & Frijda, N. H. Appraisal theories of emotion: state of the art and future development. *Emot. Rev.* **5**, 119–124 (2013).

34. Schwarz, N. Attitude construction: evaluation in context. *Soc. Cogn.* **25**, 638–656 (2007).

35. Higgins, E. T. Knowledge activation: accessibility, applicability, and salience. In *Social Psychology: Handbook of Basic Principles* (eds Higgins, E. T. & Kruglanski, A. W.) 133–168 (Guilford Press, New York, 1996).

36. Mischel, W. & Shoda, Y. A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychol. Rev.* **102**, 246–268 (1995).

37. Tett, R. P. & Guterman, H. A. Situation trait relevance, trait expression, and cross-situational consistency: testing a principle of trait activation. *J. Res. Pers.* **34**, 397–423 (2000).

38. Feldman, J. M. & Lynch Jr., J. G. Self-generated validity and other effects of measurement on belief, attitude, intention, and behavior. *J. Appl. Psychol.* **73**, 421–435 (1988).

39. He, P., Gao, J. & Chen, W. DeBERTaV3: improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. Preprint at https://arxiv.org/abs/2111.09543 (2021).

40. Williams, A., Nangia, N. & Bowman, S. R. A broad-coverage challenge corpus for sentence understanding through inference. In *Proc. 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* 1112–1122 (Association for Computational Linguistics, 2018).

41. Reimers, N. & Gurevych, I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. In *Proc. 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* 3980–3990 (Association for Computational Linguistics, 2019).

42. Schwarz, N. & Clore, G. L. Mood, misattribution, and judgments of well-being: informative and directive functions of affective states. *J. Pers. Soc. Psychol.* **45**, 513–523 (1983).

43. Klein, R. A. et al. Investigating variation in replicability: a "many labs" replication project. *Soc. Psychol.* **45**, 142–152 (2014).

44. Lynott, D. et al. Replication of “Experiencing physical warmth promotes interpersonal warmth” by Williams and Bargh (2008). *Soc. Psychol.* **45**, 216–222 (2014).

45. Ebersole, C. R. et al. Many Labs 3: evaluating participant pool quality across the academic semester via replication. *J. Exp. Soc. Psychol.* **67**, 68–82 (2016).

46. Eerland, A. et al. Registered replication report: Hart & Albarracín (2011). *Perspect. Psychol. Sci.* **11**, 158–171 (2016).

47. Camerer, C. F. et al. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nat. Hum. Behav.* **2**, 637–644 (2018).

48. Klein, R. A. et al. Many Labs 2: investigating variation in replicability across samples and settings. *Adv. Methods Pract. Psychol. Sci.* **1**, 443–490 (2018).

49. Bago, B. et al. Situational factors shape moral judgements in the trolley dilemma in Eastern, Southern and Western countries in a culturally diverse sample. *Nat. Hum. Behav.* **6**, 880–895 (2022).

50. Klein, R. A. et al. Many Labs 4: failure to replicate mortality salience effect with and without original author involvement. *Collabra Psychol.* **8**, 35271 (2022).

51. Edwards, J. R. & Lambert, L. S. Methods for integrating moderation and mediation: a general analytical framework using moderated path analysis. *Psychol. Methods* **12**, 1–22 (2007).

52. Griskevicius, V., Tybur, J. M. & Van den Bergh, B. Going green to be seen: status, reputation, and conspicuous conservation. *J. Pers. Soc. Psychol.* **98**, 392–404 (2010).

53. Berger, J. & Heath, C. Who drives divergence? Identity signaling, outgroup dissimilarity, and the abandonment of cultural tastes. *J. Pers. Soc. Psychol.* **95**, 593–607 (2008).

54. Luhtanen, R. & Crocker, J. A collective self-esteem scale: self-evaluation of one's social identity. *Pers. Soc. Psychol. Bull.* **18**, 302–318 (1992).

55. Bolton, L. E., Warlop, L. & Alba, J. W. Consumer perceptions of price (un)fairness. *J. Consum. Res.* **29**, 474–491 (2003).

56. Snyder, M. Self-monitoring of expressive behavior. *J. Pers. Soc. Psychol.* **30**, 526–537 (1974).

57. Binz, M. et al. A foundation model to predict and capture human cognition. *Nature* **644**, 1002–1009 (2025).

58. IntHout, J., Ioannidis, J. P. A. & Borm, G. F. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Med. Res. Methodol.* **14**, 25 (2014).

59. Bless, H. & Schwarz, N. Mental construal and the emergence of assimilation and contrast effects. *Adv. Exp. Soc. Psychol.* **42**, 319–373 (2010).

60. Kruglanski, A. W. et al. A theory of goal systems. *Adv. Exp. Soc. Psychol.* **34**, 331–378 (2002).

61. Kelley, H. H. The processes of causal attribution. *Am. Psychol.* **28**, 107–128 (1973).

62. Griffin, D. W. & Ross, L. Subjective construal, social inference, and human misunderstanding. In *Advances in Experimental Social Psychology* Vol. 24 (ed. Zanna, M. P.) 319–359 (Academic Press, San Diego, 1991).

63. Kruglanski, A. W. & Webster, D. M. Motivated closing of the mind: "seizing" and "freezing." *Psychol. Rev.* **103**, 263–283 (1996).

64. Baars, B. J. *A Cognitive Theory of Consciousness* (Cambridge Univ. Press, Cambridge, 1988).

65. Dehaene, S. & Naccache, L. Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition* **79**, 1–37 (2001).

66. Aaker, D. A. & Keller, K. L. Consumer evaluations of brand extensions. *J. Mark.* **54**, 27–41 (1990).

Acknowledgements

We thank Bernd Schmitt and Kristen Lane for helpful comments and Chunli An for research assistance.

Author contributions

J.C. and W.X. conceived the study and developed the methodology. L.C. conducted the investigation. J.C. and W.X. wrote the original draft, and all authors reviewed and edited the manuscript.

Competing interests

The authors declare no competing interests.

Correspondence and requests for materials

Correspondence and requests for materials should be addressed to Jack Chen (cchen27@gsb.columbia.edu).

Funding

J.C. discloses support for the research of this work from Columbia Business School. W.X. and L.C. declare no relevant funding.

Supplementary Information for

Agent psychonomicus: Psychological mechanisms emerge in cognitive agents simulated by large language models

Jack Chen, Weiheng Xiao, and Lexi Chen

*Corresponding author. Email: cchen27@gsb.columbia.edu

Contents: Supplementary Notes 1 to 23; Supplementary Tables 1 to 13; Supplementary References.

Supplementary Notes

Supplementary Note 1. *Agent psychonomicus*, cognitive models for testing process theories

This note gives the full derivation of the two conditions of *Agent psychonomicus* that the main text summarizes in the Introduction.

We modeled the name *Agent psychonomicus* on *Homo economicus*, the model of the decision-maker in economic theory, whose choices follow from the principles of rational choice. By analogy, the responses of the cognitive agents of an *Agent psychonomicus* follow from general cognitive processes that psychological theories describe, so that psychologists can study these responses as they study the responses of human participants, by testing hypotheses about the processes that precede them.

Being able to observe and manipulate these processes alone does not, however, mean that the responses of cognitive agents yield causal evidence for the theory. In this section, we propose two additional conditions. The first condition is that the cognitive processes that precede the cognitive agents' responses have four functional properties, each of which corresponds to an assumption that the experimental causal chain approach makes about human participants (Fig. 2A). These four functional properties are that (1) the cognitive agents' process statements result from how they interpret, or construe, the situation; (2) their responses are based on the content of the process statements that precede them; (3) they act on a process statement only when the situation makes it relevant to the decision; and (4) their individual characteristics shape their responses in situations that offer cues to which these characteristics respond.

The second condition, which we call restricted information, is that each process agent, the component of the model that carries out one process, receives only the information that its process requires, and thus no information about the theory or about any aspect of the experiment that its process does not need. Under restricted information, a mediator in the cognitive agents' process statements arises from the general cognitive processes (Fig. 2B).

Four functional properties of Agent psychonomicus

We derive the four functional properties from the received view of experimental methods in psychology and from the conception of the person on which process theories in social and consumer psychology build. With regard to experimental methods, we take the perspective that an experiment tests a theory only in conjunction with auxiliary assumptions about the measures, the experimental conditions, and the participants[1–4]. The experimental causal chain approach, too, makes assumptions about how human participants respond when researchers measure the mediator, manipulate it, vary a situational moderator, or compare individuals[5–7]. For cognitive agents to take the place of human participants in these tests, an *Agent psychonomicus* must therefore be designed with these assumptions in mind. Thus, drawing on the conception of the person in social cognition research[8,9], we propose that an *Agent psychonomicus* must have four functional properties, which specify how its cognitive agents arrive at their responses: construal, reliance, conditionality, and individuality.

We derive the first functional property, construal, from the long-standing view in social psychology that a situation affects behavior through the person's subjective construal of it rather than through its objective features alone[9,10]. Researchers draw on this view when they measure the mediator after the manipulation and take a difference between conditions as evidence that the manipulation produced a psychological state in human participants[5]. This conclusion holds only

if the measure captures the psychological state rather than features of the experimental procedure, a quality of measures also referred to as construct validity[11,12]. The same holds for cognitive agents, whose process statements serve as this measure. If their process statements merely repeated the wording of the scenario, a process statement that conveys the mediator would show only that the scenario's wording suggested the mediator, and not that the manipulation produced it in the cognitive agent. In an *Agent psychonomicus*, construal therefore requires that the process statements of cognitive agents result from how they construe the situation that the scenario describes.

We derive the second functional property, reliance, from the conception that a person's response is constructed from the content of the thoughts and feelings that are accessible at the time[8,13,14]. Researchers draw on this conception when they manipulate the mediator directly and take the effect of this manipulation on the response as evidence that the mediator affects the response[7]. This conclusion holds only if this effect arises from the content of the induced mediator rather than from other aspects of the induction[6]. Suppose, for example, that a cognitive agent would pay more for a product after any process statement about the product was induced, even a process statement that denies the mediator. A higher price after the researcher induced the mediator would then not show that the mediator raised the price. In an *Agent psychonomicus*, reliance therefore requires that the responses of cognitive agents be based on the content of the process statements that precede them.

We derive the third functional property, conditionality, from the conception that accessible thoughts and feelings shape a response only to the degree that they are applicable, or diagnostic, for the decision at hand[15,16]. Human decision-makers, for example, stop relying on a feeling once they attribute it to a source unrelated to the decision[13]. Building on this conception, a process theory may propose that a situational moderator, a feature of the situation other than the manipulated factor, alters the mediator's effect by making the mediator relevant or irrelevant to the decision. Researchers test such a proposition by varying the moderator in a moderation-of-process design and take the resulting interaction as evidence of when and how the mediator affects the response[5,17]. In an *Agent psychonomicus*, conditionality therefore requires that whether a cognitive agent acts on a process statement depend on whether the situation makes the process statement relevant to the decision; otherwise, a situational moderator could not alter the mediator's effect as the theory proposes.

We derive the fourth functional property, individuality, from the interactionist view that an individual characteristic shapes behavior mainly in situations that offer cues to which the characteristic responds[18–20]. A threat to consumers' identity, for example, should lower evaluations of identity-linked products only among consumers low in collective self-esteem[21]. In fact, recent work suggests that such differences between individuals in the effects of manipulations are more common than psychologists typically assume[22,23]. For a process theory, these differences are also evidence about the process. If a manipulation operates by presenting a cue to which a characteristic responds, its effect should depend on the level of this characteristic, and finding this dependence therefore supports the process that the theory proposes[17,20]. In an *Agent psychonomicus*, individuality therefore requires that the individual characteristics of cognitive agents shape their responses in situations that offer such cues. A characteristic that had the same effect on responses in every situation, however, would have this effect equally in both conditions and could not produce the heterogeneity in the manipulation's effect that the theory predicts.

The condition for the mediator's emergence: restricted information

For a mediator to emerge in the cognitive agents as it does in human decision-makers, it must arise from general processes applied to what the cognitive agents notice in the situation. In an *Agent psychonomicus* built with LLMs, two sources of information bear on this condition, namely the instructions that govern each process agent, such as the system prompt of an LLM, and the knowledge of published research that an LLM acquires in training[24].

In fact, experimenters face a similar problem with human participants: participants who guess the hypothesis may respond as they believe the experimenter expects, and experimenters therefore keep participants unaware of the hypothesis[25]. We thus propose the second condition for an *Agent psychonomicus*: each process agent must receive no information about the theory and no information about the experiment beyond what its process requires, a condition that we call restricted information. Next, we describe how restricted information can be built into an *Agent psychonomicus* through two architectural requirements.

Because psychological constructionism holds that psychological states arise from general processes[26], the first architectural requirement is that the instructions that govern each process agent describe only a general process, such as appraising how the situation bears on the self. The instructions must therefore never name the experiment, the theory, the hypothesis, or the mediator, or reveal which response the researcher expects. The second architectural requirement is that each process agent receive only the information that its process requires, such as what the cognitive agent has noticed and the individual characteristics that bear on this process, so that it has little basis for recognizing the experiment from its training data (Fig. 2B).

When a mediator appears in a cognitive agent's process statements and both architectural requirements are met, we propose that this mediator has *emerged* in the cognitive agent's process statement processes. By emergence, we mean that a psychological state arises from the interplay of general processes, none of which is dedicated to producing it[27]. In this view, individuals categorize what they perceive and appraise how it bears on the self and on their goals, and different psychological states arise when individuals apply these general processes to different situations[26,28,29]. From this theoretical perspective, the mediators that process theories propose arise in exactly this way. The feeling that an object is one's own, for example, develops when individuals control an object or come to know it intimately, as when they touch it, and thereby experience it as part of the self[30–32]. Thus, we propose that in an *Agent psychonomicus* that meets both architectural requirements, a mediator found in the cognitive agents' process statements arises from general processes applied to what the cognitive agents noticed in the situation. Such a mediator makes the cognitive agents' process statements psychologically plausible, because it presumably arose in the same way as the theory assumes it arises in human participants. Researchers can therefore interpret the predicted effects in cognitive agents, as they would in human participants, as the result of a mediator that arose from construing the situation and in turn shaped the response.

Supplementary Note 2. Architecture of MVCA

This note gives the full description of MVCA that the main text summarizes in the Results.

In this section, we describe MVCA, an implementation of *Agent psychonomicus* based on a multi-agent setup of LLMs, and explain how we designed it to meet the conditions proposed in Supplementary Note 1. Existing work has used multi-agent setups of LLMs to simulate human

behavior, such as generative agents that simulate the daily lives of the residents of a small town[33]. MVCA differs from this existing work in two important respects. Conceptually, whereas the LLM agents in existing work are designed to produce humanlike behavior, MVCA is designed to simulate the cognitive processes that precede a response, derived from general theories of cognition (Table 1). In its implementation, MVCA does not have a single LLM agent respond as a participant; instead, the LLM carries out each cognitive process of a cognitive agent in a separate LLM call.

We imbue MVCA with the four functional properties through three design choices: the order of the processes, together with what each process agent receives, establishes construal and reliance; knowledge activation and the general rules, which make the influence of a process statement depend on what the cognitive agent noticed in the situation, establish conditionality; and the way in which MVCA introduces an individual characteristic establishes individuality. To meet the condition of restricted information, we built both architectural requirements into the system prompts and into the information that each process agent receives, as we will show shortly.

Cognitive architecture of MVCA

Following psychological constructionism, we built MVCA from 12 general processes rather than from processes dedicated to particular psychological states; in these processes, the cognitive agent encodes the situation, appraises it, integrates the resulting process statements, and responds (Fig. 3A). Each process is carried out by a process agent, a separate LLM call with its own system prompt, in which the LLM states the result of the process as a short first-person process statement. In the first process, the LLM encodes the situation that the cognitive agent faces by describing what the cognitive agent notices in the scenario; we call this process statement the encoding[8]. After encoding the situation, the cognitive agent evaluates what it noticed in six appraisal processes, derived from appraisal theories and psychological constructionism[26,34,35]: in categorization, it identifies what kind of object or situation it encounters; in the appraisals of self-relevance and goal relevance, it considers how the situation bears on the self and on its current goals; in affective response, it registers how it feels about the situation; in causal inference, it asks why the situation arose; and in social inference, it asks what others intend. Having appraised the situation, the cognitive agent combines these appraisals in four integrating processes: it considers how else the situation could be construed (consideration of alternative readings), compares the options that it noticed (comparison of alternatives), summarizes what it thinks and is inclined to do (first-person account), and examines whether its process statements are consistent with one another (coherence review). Finally, in judgment, the cognitive agent answers the question of the experiment on the basis of these process statements. We call the process statements and the response that MVCA produces for one cognitive agent in one experimental condition a cognitive simulation.

The cognitive processes of MVCA, as described above, have the properties of construal and reliance because of the order of the processes and the information that each process agent receives. Following Global Workspace Theory, according to which specialized processes make their results available to one another through a workspace of limited capacity[36,37], later process agents receive the process statements of earlier ones. Because human decision-makers respond to a situation as they construe it[9], the process agents for the appraisals never receive the scenario, only the encoding or process statements derived from it and, for the appraisal of goal relevance, the names of drives that words in the scenario suggest (see below; Fig. 3A). Later process statements thus build on what the cognitive agent noticed rather than on the wording of the

scenario, as construal requires. Similarly, the process agent for judgment receives the process statements of the ten processes between encoding and judgment but not the scenario, so that the cognitive agent's response is based on these process statements, as reliance requires.

We imbue MVCA with conditionality in two ways, through knowledge activation and through general rules. Knowledge activation follows the principle that human decision-makers use stored knowledge when it is accessible and applicable to the situation[15]. Accordingly, for six of the processes (Table 1), later process agents receive the full process statement only when the cognitive agent has noticed something to which the process applies, as the LLM determines from the encoding in a separate LLM call (Fig. 3B), and otherwise only a brief version of this process statement (Methods). The general rules determine whether, or in what form, a process statement that the researcher induces shapes the cognitive agent's later process statements. Under the motive–means rule, which follows goal systems theory[38], an induced desire shapes later process statements only if the cognitive agent has noticed that choosing an option would serve the desire. Under the discounting rule, which follows affect-as-information theory[13], an induced feeling is set aside if the cognitive agent has noticed a source of the feeling that is unrelated to the target of the judgment. Under the standard-comparison rule, which follows the inclusion/exclusion model[39], an induced process statement that compares the target of the judgment with a standard leads the cognitive agent to consider whether the target in fact falls within this standard or clearly outside it, a distinction that, according to this model, determines whether the judgment moves toward the standard or away from it (Table 1; Methods). Beyond knowledge activation and the general rules, the process agents for most appraisals receive the encoding, so that the cognitive agent can consider an induced mediator together with what it noticed about the option and its circumstances (Study 3). A situational moderator that the cognitive agent noticed in the scenario can therefore alter the mediator's effect, as conditionality requires.

With regard to restricted information, we built the architectural requirement on the instructions into MVCA by restricting the system prompt of each process agent to a general process derived from theories of cognition rather than from the process theory under test (Table 1). Three fixed examples in the system prompts, identical in every experiment and condition, name constructs related to the mediators of seven experiments, among them psychological ownership, the mediator of the mere-touch experiment (Supplementary Note 3). Under the no-description criterion, these seven mediators therefore do not count as emergent (Study 1). Also note that we stated each general rule in the same terms for every experiment and without naming any particular option, so that a rule describes only a principle of a broader psychological theory. To test the mediator's effect on the response, we induced the mediator by presenting a process statement that states it, as one of the cognitive agent's own process statements, to the process agent for the appraisal in which, according to the theory, the mediator arises (Studies 2 and 3; Supplementary Table 1).

Following the architectural requirement on the information, we provided each process agent with only the information that its process requires (Fig. 3A; Supplementary Information). This means that no process agent receives the name of the experiment, the theory, the hypothesis, the other experimental condition, or the direction in which the theory predicts that the conditions will differ. Also note that the process agent for encoding, the only one that receives the scenario, does not receive the question with which the experiment measures the response, so that this measure cannot influence what the cognitive agent notices; only the process agent for judgment receives the question and its response options. In addition, the process agent for the appraisal of goal

relevance receives the names of drives, the characteristics that describe what an individual strives for (e.g., achievement or need to belong), but not the cognitive agent's levels on these drives; MVCA selects the drives from words in the scenario (Supplementary Information).

Meeting both architectural requirements means that in MVCA, a mediator that appears in a cognitive agent's process statements has emerged from the cognitive agent's construal of the situation through general processes. In the brand-extension experiment[40], for example, the manipulation had a positive effect on support for perceived fit, the degree to which a cognitive agent's process statements imply rather than contradict this mediator; for the cognitive agent shown in Fig. 3C, this support rose from −0.68 to −0.31.

Individual characteristics of the cognitive agents

Whereas all cognitive agents share the same general processes, each is created from its own agent profile, which describes one simulated individual by age, gender, occupation, city, a short biography, and 139 individual characteristics studied in psychology (Methods; Supplementary Information). For each characteristic, we listed the cognitive processes through which, according to published research, it affects judgments and choices (Supplementary Table 2), and in MVCA a characteristic can shape a cognitive agent's process statements only through these processes[19,20].

Individual characteristics shape a cognitive agent's process statements in three main ways (Fig. 3, A and B). First, process agents receive descriptions of the characteristics linked to their process (in all, 41 of the 139 characteristics); each description states how high or low the cognitive agent scores on the characteristic, such as a high level of materialism, and what this score means (Supplementary Information). Second, because individual characteristics also affect how extensively individuals think about a situation[41,42], 17 characteristics, which are never described to any process agent, make the cognitive agent's process statements in particular processes fuller or briefer (Supplementary Table 2). Third, for the characteristics that we varied in Study 4 (Supplementary Table 3), MVCA introduces one or two first-person sentences describing the characteristic into the appraisal process that, according to published research, the characteristic affects, but only once the cognitive agent has noticed a cue to which the characteristic responds (Fig. 3B). The characteristic thus shapes the cognitive agent's process statements only in situations that offer such a cue, as individuality requires; following trait activation theory[20], we call this procedure trait activation.

Because process agents receive descriptions of some individual characteristics, we count a construct that the process statements convey, such as a mediator or an individual characteristic, as emergent only if no process agent received a description of it in any form, a criterion that we call the no-description criterion (Study 1).

Supplementary Note 3. Cognitive processes of MVCA and their inputs

This section describes the instructions and the inputs for each cognitive process of MVCA. Each cognitive simulation comprises 13 LLM calls, one for each of the 12 processes and one for knowledge activation. In Study 4, the LLM carries out trait activation in one further LLM call, and whenever a general rule applies to an induced process statement or to the sentences introduced by trait activation, the LLM determines in one further LLM call whether the condition of the rule is met. For each process before judgment, the LLM states the result of the process as a process statement, which consists of a short label naming its main content and one or more first-person sentences; for judgment, it states the response.

The instructions for each process consist of the system prompt and the instruction lines of the user prompt. They include one line that states the task of the process ("At this step, do only this: …"), followed by a line that forbids the LLM to decide what the cognitive agent will do or to state an answer, amount, rating, or choice. The table below lists, for each cognitive process, this task line, the earlier processes whose process statements the LLM receives for it, and what determines whether later process agents receive its process statement in full, as a brief version, or not at all. No process agent after the one for encoding receives the scenario, and no process agent before the one for judgment receives the question or its response options.

Cognitive process	Task line in its instructions	Receives the process statements of	Full process statement, brief version, or none
Encoding	(instruction, paraphrased) describe the situation faithfully, without evaluating it or mentioning any question or decision	the scenario only	full
Categorization	say what kind of thing and situation this is	encoding	full; with high need for cognitive closure, later process agents receive only its first candidate category
Self-relevance	judge how this bears on you and your sense of self	encoding, categorization	knowledge activation: "Does the encoding mention something you yourself did earlier, or something said about you?"
Goal relevance	judge how this bears on your current goals: what could be gained or lost	encoding, categorization	knowledge activation: "Is there something specific you could lose or gain?"
Affective response	say what you feel, as it follows from how this bears on you and on your goals	self-relevance, goal relevance	knowledge activation: "Is there something clearly pleasant or unpleasant?"; low need for cognition counts as one step up among the shifts described below
Causal inference	judge why things are as they are and who or what is responsible	encoding, categorization	knowledge activation: "Is there something unusual that needs explaining (for example a price, a request, or the way someone acts)?"
Social inference	judge what other people involved intend or signal	encoding, categorization	knowledge activation: "Are other people present, or will anyone see or learn about what you choose or do?"
Alternative readings	consider other ways the situation could be read	encoding, categorization, social inference	knowledge activation: "Is there any sign that things may not be what they seem?"; none with high need for cognitive closure; with low need for cognition, later process agents receive only its first reading
Comparison of alternatives	compare the alternatives that are present in the situation (not answer options)	categorization, self-relevance, goal relevance, affective response	full; brief with low need for cognition; later process agents do not receive its ranking of the alternatives
First-person account	tell, in the first person, how you are experiencing this moment	the six appraisal processes, alternative readings, comparison of alternatives	full; later process agents do not receive its statement of what the cognitive agent is inclined to do
Coherence review	point out anything in your notes that does not fit together	the six appraisal processes, alternative readings, comparison of alternatives, first-person account	full; brief with low cognitive reflection
Judgment	(instruction, paraphrased) answer the question on the basis of the process statements, without introducing facts that are not in them	the question and the process statements of the ten processes after encoding (not the encoding)	—

For the first-person account, the LLM also receives a narrative description, a fixed text that MVCA completes with content from the agent profile (Supplementary Note 4). The brief version

of a process statement consists of its short label (words only, at most 40 characters) and its first sentence; when a process produces no process statement, later process agents receive “(not on your mind right now)” in its place. Before later process agents receive a process statement, MVCA removes all numbers from it, so that no process statement states a price or a rating that the process agent for judgment could adopt as the response. For the same reason, later process agents do not receive two parts of the process statements that resemble a response: the ranking of the alternatives in the comparison of alternatives and the first-person account's statement of what the cognitive agent is inclined to do. When we induce a process statement in an appraisal process (the process statement that states the mediator, the process statement that denies it, or one of the process statements with which we compared the induced mediator in Study 2; Supplementary Note 8), or when trait activation introduces sentences that describe an individual characteristic, every later process agent that receives the process statement of this appraisal, including the process agent for judgment, receives this process statement in full together with the induced process statement or the sentences, unless a general rule withholds them.

Knowledge activation, the general rules, and trait activation each depend on what the cognitive agent's encoding contains, which the LLM determines in a separate LLM call. In each of these calls, the LLM receives the encoding and a question about it (for a general rule, also a process statement; see below), and it states its answer in a fixed format together with the passage of the encoding that supports it. An answer counts only if this passage consists of at least two words and appears verbatim in the encoding; otherwise, MVCA treats the answer as negative. In knowledge activation, the LLM determines in a single LLM call, for each of the six processes subject to knowledge activation, whether the encoding contains content to which the process applies.

For each of these six processes, later process agents receive the full process statement if the LLM answers the question listed in the table in the affirmative and supports this answer with a passage of at least two words that appears verbatim in the encoding; otherwise, they receive a brief version. What later process agents receive of a process's process statement lies on a three-step scale, from the full process statement through a brief version to no process statement, and it also depends on the 17 individual characteristics that are never described to the LLM but set how fully particular processes are elaborated (Supplementary Table 2). For a process linked to such a characteristic, MVCA moves what later process agents receive one step along this scale, toward the full process statement or away from it, depending on the cognitive agent's level and on whether published research links a high level to more or to less elaboration; for the six processes subject to knowledge activation, the step starts from the outcome of knowledge activation. For the characteristics listed below, a high level moves what later process agents receive one step toward a fuller process statement and a low level one step toward a briefer one; the shifts are summed, and the sum is limited to one step for each process. The characteristics and the processes they shift are self-monitoring, need to belong, susceptibility to descriptive norms, and interdependent self-construal (social inference); persuasion knowledge (social inference and alternative readings); status motive (social inference and goal relevance); empathic concern (social inference and affective response); independent self-construal (self-relevance); prevention focus and promotion focus (goal relevance); trait anxiety (alternative readings and goal relevance); and reactance proneness (alternative readings). For three of the 17 characteristics, MVCA alters the wording of the linked process's task, in addition to or instead of shifting what later process agents receive, as when the instructions for the consideration of alternative readings ask a cognitive agent low in need for cognition to consider only one other reading of the situation

(Supplementary Note 4). When the agent profile states high arousal, a process subject to knowledge activation produces no process statement unless the LLM's answer to its question was affirmative and supported by a passage that appears verbatim in the encoding.

The general rules apply only to induced process statements, and which rule applies depends on the appraisal process in which the process statement is induced. The motive–means rule applies to desires induced in the appraisal of goal relevance, such as a desire to feel more powerful; under this rule, the LLM determines whether the encoding states something about the option that would make choosing it serve the desire. The discounting rule applies to any process statement induced in affective response or causal inference, which the rule's question refers to as a feeling or an impression; under this rule, the LLM determines whether the encoding mentions a source of the feeling or impression that is unrelated to the target of the judgment. For these two rules, the LLM receives, besides the rule's question, only the encoding and the process statement that states the mediator, even when the process statement that denies the mediator was induced, so that its answer cannot depend on which of the two process statements was induced. If this answer is negative under the motive–means rule or positive under the discounting rule, MVCA withholds the induced process statement from the process agent for the appraisal and from every later process agent, and the LLM then carries out the appraisal as if no process statement had been induced. The standard-comparison rule applies to process statements induced in categorization that compare the option with a standard, such as the process statement that a daily donation is as small as the price of a coffee. Under this rule, the LLM receives the induced process statement itself, because the rule concerns the standard that this process statement invokes, and determines whether, according to the numbers or facts in the encoding, the option falls within the standard or clearly outside it. If the LLM reaches either conclusion, MVCA attaches it to the induced process statement (for example, "compared with that, this option is clearly beyond it"), and the process agent for categorization and every later process agent receive the induced process statement together with this conclusion; otherwise, MVCA presents the induced process statement without a conclusion. The system prompt of each rule does not name the theory under test, the predicted interaction, or the situational moderator.

In trait activation, the LLM determines whether the encoding contains the cue to which the characteristic under study responds (Supplementary Table 3). If it does, MVCA presents the sentences describing the characteristic, framed like an induced process statement, among the earlier process statements that the LLM receives for the appraisal process that, according to published research, the characteristic affects. A general rule that applies to process statements induced in this appraisal also applies to these sentences.

Apart from the task line, the instructions contain fixed examples and terms that are the same in every condition, every experiment, and every LLM, and three of them name constructs related to a mediator. First, the instructions of every process agent, including the one for judgment, ask the LLM to list the characteristics in the agent profile that it weighed, and they illustrate how to name such a characteristic with two examples, one of which is psychological ownership (which appears in the instructions as "stores.self.psychological_ownership"). Psychological ownership is the mediator of the mere-touch experiment and is related to the mediators of the possession-and-valuation, emotional-attachment, endowment-effect, and IKEA-effect experiments. Second, the output format for the consideration of alternative readings includes the example label "costly-nonconformity-as-status," which is close to the mediator of the nonconforming-dress experiment. Third, the instructions for affective response ask the LLM to update valence and arousal, and

arousal is the mediator of the emotion-and-sharing experiment. Because these examples and terms are the same in both conditions of every experiment, they cannot by themselves produce the difference between conditions that Studies 1 to 4 test. In the experiments named above, however, they could account for the presence of a construct related to the mediator in the process statements; for this reason, the construct-activation analysis with the 69 characteristics excludes psychological ownership, the belief that consumption signals status, and current arousal (Supplementary Note 4). They may also have made it easier for the LLM to produce such a construct once the manipulation made it applicable to the decision. To examine this possibility, we excluded the seven experiments named above, the only ones of the 60 whose mediators are related to psychological ownership, to status signaled by nonconformity, or to arousal; without these seven experiments, the increase in support for the mediator in Study 1 remained (averaged over the three LLMs, d_z = 0.36, 95% CI [0.17, 0.54], 53 experiments; with Gemma 4 26B, 0.16 [0.04, 0.28]; with Qwen 3.7 Flash, 0.20 [0.09, 0.32]; with Ling 3.0 Flash, 0.22 [0.10, 0.35]). Excluding the nonconforming-dress experiment alone, the increase averaged over the three LLMs was d_z = 0.39 (95% CI [0.23, 0.56]).

Supplementary Note 4. Agent profiles and process-specific information about the cognitive agent

This section describes the agent profiles and what the LLM receives about the cognitive agent for each process. The 150 agent profiles were generated by a script for an earlier version of our cognitive model. The script drew each adult's age, gender, occupation, and city independently from lists of plausible values, drew the Big Five personality traits and six further characteristics by Latin hypercube sampling from uniform distributions between 0.05 and 0.95, and composed a short biography and a description of cognitive style from sentences keyed to these values. We retained these values and then specified 139 individual characteristics for each profile: 124 levels on rating scales, such as need for cognition and need to belong, and 15 short texts, such as the individual's actual, ideal, and ought selves[43]. Of the 124 numeric individual characteristics, 32 were drawn by Latin hypercube sampling from published population norms, with means shifted according to the Big Five scores; three (need for cognition, need for uniqueness, and self-monitoring) were taken from the earlier profiles; ten beliefs, such as the belief that a higher price indicates higher quality, were rated from 0 to 1 by an LLM (gpt-4o-mini) on the basis of the biography; and 79 were drawn independently for each profile from normal distributions with means mostly between 0.45 and 0.60 and standard deviations between 0.16 and 0.20, shifted according to the Big Five scores. Among the 15 text characteristics, which also include the generational cohort, the descriptions of the actual, ideal, and ought selves and five formative memories were composed by gpt-4o-mini from the biography, identity domains were derived from the occupation and the biography, and five self-views (moral identity, hoped-for possible self, feared possible self, self-brand connection, and role-identity centrality) are the same sentence in every profile. We used an LLM (Claude), whose training data include published research, to compile the list of characteristics, the kinds of processes that published research links to each, and the confidence in each link; Supplementary Table 2 lists the characteristics and how each enters MVCA.

The instructions of every process agent, including the one for judgment, open with the cognitive agent's first name, age, gender, occupation, and city. Each process agent other than the one for judgment then receives, under the heading "What you bring to this step (facts about you; draw on one only if the situation in your notes makes it relevant)," one line for each characteristic that is

described for its process (Supplementary Table 2). The line states the characteristic's name, the cognitive agent's score on a scale from 0 to 1, its percentile and band relative to the norms, and the meaning of this level, for example "price quality: 0.50 (25th pct, mod-low) — expensive doesn't necessarily mean good; suspicious of premium claims." A numeric characteristic appears only if its level differs from the norm by at least 0.6 standard deviations, and a text characteristic always appears. In each experiment, no process agent receives the cognitive agent's current mood or its levels on the current states that correspond to constructs of the theory under test. For the first-person account, the LLM also receives a narrative description with the descriptions of the selves, the identity domains, the generational cohort, and the five formative memories.

Unlike these described characteristics, the 17 characteristics that shift how fully a process is elaborated are not described by name or score. Some of them, however, alter the wording of a task: a cognitive agent high in need for cognitive closure is asked to adopt a single reading in categorization, a cognitive agent low in need for cognition is asked to consider only one other reading in the consideration of alternative readings and to compare the alternatives only on their single most noticeable aspect, and construal level determines one sentence of the instructions for encoding.

For the appraisal of goal relevance, the LLM also receives the names of some drives, selected because words in the scenario match them or, if no word matches, because they are the five drives on which the cognitive agent scores highest. These drives are 18 characteristics of the agent profile, whose names the LLM receives without the cognitive agent's levels. Apart from the encoding and the process statements that build on it, these names are the only information derived from the scenario that the LLM receives for a process other than encoding. Because they are derived from the scenario, the names of drives can differ between the conditions of an experiment; in Study 1, they were the same in both conditions of 51 of the 60 experiments. By contrast, the descriptions of the cognitive agent's characteristics and the system prompts are the same in both conditions of every experiment, so the process statements after encoding could differ between the conditions of these 51 experiments only through what the cognitive agent noticed in the scenario. In these 51 experiments, the pooled increase in support for the mediator, averaged over the three LLMs, was $d_z = 0.48$ (95% CI [0.26, 0.70]; with Gemma 4 26B, 0.26 [0.10, 0.41]); this increase therefore arose from what the cognitive agents noticed in the scenario, as construal requires. In the other nine experiments, the increase was 0.23 (95% CI [−0.19, 0.64]). In Studies 3 and 4, the list of names of drives was the same for the induced mediator and its denial and for the two levels of an individual characteristic, because these contrasts leave the scenario unchanged; the list could differ, however, between the levels of a situational moderator and between the conditions of a manipulation.

To meet the no-description criterion, the construct-activation analysis of Study 1 is based on the 69 characteristics that the LLM never received in any form. Of the 98 characteristics never described to the LLM (the 139 minus the 41 described for some process), 29 reached it in another form, as follows. Excluding the 41 described characteristics, the 11 other characteristics whose names can appear among the names of drives (the remaining 7 of the 18 drives are among the 41 described characteristics), and construal level, which alters the wording of the instructions for encoding, leaves 86 characteristics. When we examined the exact information that the LLM received in the Study 1 cognitive simulations with Gemma 4 26B in the 32 experiments of the deposited specification (3,138 cognitive simulations, each with 12 LLM calls for the processes), we found that a further 17 of these 86 characteristics nevertheless reached the LLM in another

form, so that 29 characteristics that were never described reached the LLM in some form. The narrative description that the LLM receives for the first-person account states the cognitive agent's generational cohort and five formative memories (success, failure, rejection, belonging, and loss) in every cognitive simulation. In addition, the names of seven characteristics appear in the instructions for some processes in every cognitive simulation: psychological ownership (as an example of a label in the output format of every process), the belief that consumption signals status, signal sophistication, life satisfaction, the moral meaning of money, willpower theory (named mind–body dualism), and current arousal. Finally, the self-descriptions and memories of some cognitive agents express the level of four characteristics: self-monitoring (10 of the 50 cognitive agents), trait anxiety (12), empathic concern (2), and private self-consciousness (1). Excluding these 17 leaves the 69 characteristics that the LLM received in no form. With these 69, the correlation between closeness to the mediator and activation, averaged over the three LLMs, was positive in 54 of 60 experiments and larger with the experiment's own process statements in 52 of 60 (mean $\rho = 0.32$ compared with 0.19 with the process statements of the other experiments; Wilcoxon $P < 0.001$; results with each LLM in Supplementary Note 6).

Supplementary Note 5. Measures of the process statements (Study 1)

Supplementary Table 5 reports the effect of the manipulation on several measures of the process statements of the ten processes between encoding and judgment. The support for the mediator, the measure reported in Study 1 and Supplementary Table 4, is the probability, estimated by the natural language inference model, that a process statement implied the mediator statements minus the probability that it contradicted them; we averaged this support over the ten process statements or computed it in the process statement of the appraisal process in which, according to the theory, the mediator arises. A second score, the mediator-minus-denial score, weighs the mediator statements against their denials: it is the probability that a process statement implied the mediator statements minus the probability that it implied the denial statements. For the deposited criterion of Study 1, we averaged this score over the ten process statements; with each LLM, we also computed it in the process statement of the appraisal process in which, according to the theory, the mediator arises (Supplementary Table 5). To test whether the manipulation affected the experiment's own mediator in particular, we computed the mediator-minus-denial score in the process statements for the statements of the two experiments that follow the experiment in Supplementary Table 1, continuing from the top of the table for the last two experiments; we chose these experiments by their position so that the choice would not depend on what their statements say. For each cognitive agent, we then subtracted the treatment-minus-control difference in the score for the other experiments' statements from the treatment-minus-control difference in the score for the experiment's own statements. The result is the score beyond other experiments' mediators, which we likewise computed in the process statement of that appraisal process with each LLM (Supplementary Table 5).

In panel B of the same table, a second measure of the mediator is its rating by a separate LLM (DeepSeek V4.1 Flash), which was blind to the scenario and the condition. This LLM made one rating of the mediator for each cognitive simulation, using the definition of the mediator and the anchors of the rating scale in each experiment's materials, and it received the complete process statements, including the parts that later process agents do not receive. In the brand-extension experiment[40], for example, a new product in a close rather than a distant category increased, averaged over the three LLMs, both support for the statement that the product is a natural fit for

what the brand makes (d_z = 2.98) and the ratings of perceived fit (d_z = 3.58; Supplementary Table 5).

With two further measures, we tested whether the manipulation merely shifted the process statements toward favorable or pleasant content, which could have raised support for many mediators without reflecting the particular psychological state that each theory proposes. We defined pleasantness as the probability that the process statements implied "I feel good and pleasant right now" minus the probability that they implied "I feel bad and unpleasant right now," averaged over the ten process statements as support was (Gemma 4 26B). In addition, panel B reports the treatment–control difference in the rating of the mediator minus the same difference in the separate LLM's rating of a favorable evaluation of the target of the judgment. With Gemma 4 26B, pleasantness did not increase on average (d_z = −0.07, 95% CI [−0.19, 0.06]) and shifted in opposite directions in different experiments (Supplementary Table 5), so a shift in mood cannot explain why the mediator increased in most experiments. Averaged over the three LLMs, the ratings of the mediator increased more than the ratings of how favorably the process statements evaluated the target of the judgment (difference d_z = 0.54, 95% CI [0.26, 0.82]; results with each LLM in Supplementary Table 5 and Supplementary Note 15), so the manipulation did more than make the process statements more favorable.

Supplementary Note 6. Construct activation (Study 1)

With each LLM separately, as in the average over the three LLMs reported in Study 1, the closer a characteristic was in meaning to the experiment's mediator, the more strongly it was activated in the cognitive agents' process statements in most experiments, both with the 69 characteristics that the LLM received in no form and with all 139 characteristics (Supplementary Table 6). With the 69 characteristics, which Fig. 6, A and B, shows averaged over the three LLMs, the correlation between closeness and activation was positive in 57 of 60 experiments with Gemma 4 26B, 55 with Qwen 3.7 Flash, and 51 with Ling 3.0 Flash. It was larger with the experiment's own process statements than with the process statements of the other experiments in 53, 47, and 44 of the 60 experiments, respectively (mean ρ 0.32 compared with 0.17, 0.32 compared with 0.19, and 0.29 compared with 0.20; Wilcoxon $P < 0.001$ with each LLM). In the brand-extension experiment, whose characteristics Fig. 6A shows, the Spearman correlation between the 69 characteristics' closeness in meaning to the mediator (perceived fit) and their activation in the experiment's own process statements was ρ = 0.51, averaged over the three LLMs. With all 139 characteristics, the correlation between a construct's closeness in meaning to the mediator and its activation in the process statements was positive with each LLM in most experiments (54 to 58 of 60), and it was larger with the experiment's own process statements than with the process statements of other experiments in 46 to 52 of the 60 experiments.

Beyond constructs related to the mediator, the experiment's own mediator was the most activated of the 60 mediators in its own process statements in 27 experiments with Gemma 4 26B, 24 with Qwen 3.7 Flash, and 20 with Ling 3.0 Flash, and it was among the three most activated in 32, 33, and 31 experiments, respectively. The manipulation increased the activation of the experiment's own mediator with Gemma 4 26B (d_z = 0.32, 95% CI [0.14, 0.50]) and with Qwen 3.7 Flash (d_z = 0.35, 95% CI [0.16, 0.54]); with Ling 3.0 Flash, it increased the activation, but not significantly (d_z = 0.09, 95% CI [−0.04, 0.22]). This increase exceeded the increase in the activation of the other 59 mediators with Gemma 4 26B (difference d_z = 0.30, 95% CI [0.12, 0.48]) and with Qwen 3.7 Flash (d_z = 0.35, 95% CI [0.15, 0.54]); with Ling 3.0 Flash, the difference was positive but not significant (d_z = 0.11, 95% CI [−0.02, 0.24]). Averaged over the

three LLMs, the manipulation increased the activation of the experiment's own mediator (d_z = 0.37, 95% CI [0.15, 0.60]; Supplementary Table 6), and the difference from the increase in the activation of the other 59 mediators was d_z = 0.37 (95% CI [0.14, 0.59]).

In Supplementary Table 6, activation is the cosine similarity between a construct's description and the process statements. The description of a characteristic consisted of its name and one or two sentences, except for seven characteristics, whose description was their name alone (Supplementary Table 2; descriptions in the deposited materials); the description of a mediator was its wording in Supplementary Table 1. With the sentence-embedding model all-mpnet-base-v2, we computed a representation (a vector of numbers) for each process statement of the ten processes between encoding and judgment, for each construct's description, and for each scenario text. So that a process statement would not count as activating a construct merely because it repeats the scenario, we projected each process statement's representation onto the representation of the scenario in the same cognitive simulation and removed this projection from it, which removes from the process statement the part of its meaning that it shares with the scenario. We computed closeness in meaning with all-MiniLM-L6-v2 so that closeness and activation would not be measured with the same model. For each experiment, we averaged the activation of each characteristic over all cognitive simulations of the experiment, in both conditions, and for the results averaged over the three LLMs, we averaged the activations obtained with the three LLMs for the 30 cognitive agents that all three simulated.

The effect of the manipulation on activation was the treatment-minus-control difference within each cognitive agent, signed so that positive values indicate the predicted direction, and the manipulation's effect on the activation of the experiment's own mediator is expressed as d_z across cognitive agents. We computed this effect in all ten process statements and in the process statement of the appraisal process in which, according to the theory, the mediator arises, and also as the same effect minus the effect on the other 59 mediators. For each experiment, the Spearman ρ between closeness and activation with the experiment's own process statements is compared with the mean ρ over the process statements of the other 59 experiments, and the rank of the own mediator among the 60 mediators is computed after standardizing the activations first within each experiment across the 60 mediators and then within each mediator across the 60 experiments.

Supplementary Note 7. Results in cognitive agents created from other agent profiles (Study 1)

For 50 of 100 cognitive agents created from other agent profiles, the effects of the manipulation in Study 1 with Gemma 4 26B, in the 32 experiments of the deposited specification, were of similar size to those of the cognitive agents of the four studies in the same experiments (support for the mediator, d_z = 0.23, 95% CI [0.07, 0.39], compared with 0.25; ratings of the mediator by the separate LLM, d_z = 0.26, 95% CI [0.02, 0.50], compared with 0.18).

Supplementary Note 8. Other induced process statements compared with the induced mediator (Study 2)

Study 2 comprised 52 of the 60 experiments. The theories of the other eight predict that the mediator's effect on the response depends on a situational moderator, so we tested these experiments only in Study 3, which varied this moderator (Supplementary Tables 1 and 9).

The induced mediator could have shifted the response merely because any additional process statement shifts the response, or because it favored the predicted response whether or not it stated the mediator. To examine these possibilities, we also induced other process statements in the treatment condition, in the appraisal process in which, according to the theory, the mediator arises (Supplementary Table 7). To examine whether a process statement shifts the response merely by being on the cognitive agent's mind, we compared two of these process statements with each other: the neutral process statement, which states a neutral fact about the target of the judgment (for example, "I read that the new product will reach stores next month"), and the unrelated process statement, which concerns something unrelated to the decision (for example, "I like the fresh new color of my kitchen walls"). We also compared the mediator with an equally favorable process statement, which favors the predicted response as strongly as the mediator does without stating the mediator. In the beer-price experiment[44], for example, the process statement that states the mediator was that a high price is fair at a place where beer costs more, whereas the equally favorable process statement ("A cold bottle of my favorite beer would be just perfect right now") favors a high price without stating this mediator. In addition, we induced a pleasant mood in the five experiments in which a pleasant mood could also explain the original finding (Supplementary Table 7). We report the comparison of the mediator with the neutral process statement descriptively.

Supplementary Note 9. Rationale for the general rules (Study 3)

We built general rules into MVCA because, in cognitive agents created from other agent profiles, some situational moderators did not alter the mediator's effect as the theories predict without such rules, which indicates that these predictions rely on principles of broader psychological theories. In the decision-difficulty experiment, for example, the theory holds that a feeling of difficulty makes a choice seem important[45], and the moderator alters what the feeling of difficulty is attributed to. We therefore induced the feeling itself, the process statement that the choice was harder than expected, rather than the belief that the choice was important, which is the mediator of this experiment in Study 1 (Supplementary Tables 1 and 9). When a warning stated that the website's small font would make the flight details hard to read, human participants could attribute the feeling of difficulty to the font rather than to the choice and should then not treat the feeling as a sign that the choice was important. The cognitive agents, however, did not discount the feeling: without any rule, the induced process statement that the choice was harder than expected still increased the time they would spend on the choice even after the warning about the small font (interaction $d_z = -0.21$, opposite to the prediction).

At that point, MVCA did not contain a general principle on which the theory relies without stating it: human decision-makers discount a feeling that they attribute to an unrelated source. With this principle built into MVCA as the discounting rule, the interaction was $d_z = 0.27$, in the predicted direction: in 99% of cognitive simulations with the warning and in none without it, the LLM determined that the encoding mentioned a source of the feeling unrelated to the judgment, so that MVCA withheld the induced feeling. Like the feeling of difficulty, several other mediators induced in Study 3 are desires, feelings, or standards of comparison, and for each of these, a broader psychological theory specifies when it affects the response. Specifically, according to goal systems theory, a desire guides choice when the option is seen as a means of serving it[38]; according to affect-as-information theory, a feeling is discounted when it is attributed to a source unrelated to the judgment, as in the decision-difficulty experiment[13,46]; and according to the inclusion/exclusion model, human decision-makers evaluate an option against a

standard according to whether it falls within the standard or clearly outside it[39]. MVCA's general rules follow these theories, and each rule applies to one kind of mediator (Methods).

The motive–means rule produces the predicted moderation only if the situational moderator alters whether the encoding shows that choosing the option would serve the desire. This was the case in the powerlessness experiment: in cognitive agents created from other agent profiles, MVCA withheld the induced desire to feel powerful under this rule in 56 to 62% of cognitive simulations for the exclusive product and in 97 to 99% for the mass-produced product, because in these cognitive simulations the encoding did not show that choosing the product would serve this desire. It was not the case for the desires that concern how one is seen by others. When we applied the motive–means rule to the desire for status in the going-green experiment, the encoding showed that choosing the green product would serve this desire in none of the public and none of the private choices. In the identity-divergence experiment, the encoding showed in none of the cognitive simulations, whether for the jacket or for the dish soap, that choosing the option would serve the desire not to be taken for one of the new buyers. For the desire to be accepted in the social-exclusion experiment, it showed that the option would serve the desire in 44% of cognitive simulations in which the other person would see the payment and in 46% of those in which the person would not. In Study 3, where the motive–means rule did not apply to these tests, the induced desire to be accepted increased the price that cognitive agents were willing to pay both when the other person would see the payment and when the person would not (d_z = 0.56 and 0.54; Supplementary Table 9).

Supplementary Note 10. Classes of situational moderator (Study 3)

Study 3 comprised 11 tests from 10 experiments: the eight experiments tested only in Study 3, one of which, the social-exclusion experiment, provided two tests, and two experiments of Study 2, affect reliance and pennies a day, whose theories also specify a situational moderator (Supplementary Tables 1 and 9). At each of the two levels of the situational moderator, we simulated each cognitive agent once with the process statement that states the mediator and once with its denial, so that the induced mediator was the same at both levels. In four tests, one of the general rules of MVCA applied to the induced mediator (Table 1): the motive–means rule in the powerlessness test, the discounting rule in the decision-difficulty and affect-reliance tests, and the standard-comparison rule in the pennies-a-day test. In these tests, the predicted moderation also follows from a principle of a broader psychological theory, such as discounting a feeling attributed to an unrelated source, and MVCA applies this principle in the same way in every experiment.

Before conducting the cognitive simulations of Study 3, we grouped the 11 tests of Study 3 into classes by how, according to the theory, the situational moderator makes the mediator relevant or irrelevant to the decision (Supplementary Table 9), and we fixed the plan to report the tests in these classes. In three tests, the moderator is whether the option itself is linked to the mediator: identity salience, identity threat, and group membership in the social-exclusion experiment. In four, the option can serve a desire only through what others see or infer from the choice: going green, identity divergence, social exclusion with the payment seen, and powerlessness. The remaining moderators alter the source of a feeling (decision difficulty), what a feeling indicates about the product (affect reliance), the normative message, or the direction of comparison (pennies a day). Study 3, Fig. 8A, and Supplementary Table 9 report the tests in these classes.

Supplementary Table 9 also states the function of each situational moderator in the theory. The functions come from a classification of situational moderators by function, which we derived from research on social cognition and consumer judgment, and the 11 tests fall into six of its categories. In one category, the description of the option itself states whether the option bears on the mediator, as when a product is framed in terms of the cognitive agent's family or by its features. In another, choosing the option serves the motive through what others see or infer from the choice, as when a green product is chosen in public or in private, or when a product's exclusivity signals status. The remaining four categories are the source of a feeling, what a feeling indicates about the product, a competing normative message, and the direction of comparison.

Two documents, both prepared on 6 October 2026, before the cognitive simulations of Study 3 were conducted, classify these situational moderators: a classification of the literature, which codes each situational moderator by the mechanism that the published article proposes, and the design report for the tests, which groups the tested conditions by where the relevance of the option to the mediator is represented, either in the description of the option (identity salience, identity threat, and group membership) or in what others see or infer from the choice (going green, identity divergence, social exclusion with the payment seen, and powerlessness). The classification of the literature, by contrast, assigns identity salience to two categories, the one in which the description of the option states whether the option bears on the mediator and the one in which choosing the option serves the motive through what others see or infer from the choice, but it assigns the identity-threat and group-membership conditions only to the latter. The two documents agree on the other eight tests: both assign going green, identity divergence, and social exclusion with the payment seen (in which choosing the option serves the motive only if others see the choice), together with powerlessness, to this category, and both assign decision difficulty, affect reliance, the normative message, and pennies a day to the remaining categories. The function column of Supplementary Table 9 follows the design report, whose grouping matches the classes of situational moderator described above. The design report also notes that, in cognitive agents created from other agent profiles, the induced mediator did not interact with any of the four conditions in which the option serves the motive through what others see or infer from the choice.

Supplementary Note 11. Joint sign-flip permutation test (Study 3)

The interactions differed widely between the 11 tests (with Gemma 4 26B, $I^2 = 84\%$). This variation is the main reason why the pooled interaction with each LLM alone was positive but its 95% CI included zero (with Gemma 4 26B, $d_z = 0.23$, 95% CI [−0.03, 0.48]; Supplementary Table 4).

We therefore also tested whether five interactions significant in the predicted direction could arise if no situational moderator altered the mediator's effect. Because the 11 tests of Study 3 share cognitive agents and an LLM, their interactions are not independent, so we used a sign-flip permutation test that preserves this dependence, applying it to the data from the first time each test was conducted with Gemma 4 26B (50 cognitive agents). If no moderator alters the mediator's effect, each cognitive agent's interaction is as likely to be negative as positive. In each of 20,000 permutations, we therefore multiplied all interactions of a cognitive agent, in all 11 tests, by the same randomly chosen sign (+1 or −1), recomputed the d_z of each test with its 95% CI, and counted the tests whose CI lay above zero. We observed five such tests; no permutation produced five or more, so $P < 0.0001$ (at most 1 in 20,000).

Supplementary Note 12. Ways of introducing an individual characteristic (Study 4)

In cognitive agents created from other agent profiles, we compared five ways of introducing an individual characteristic, including trait activation; Fig. 8C shows four of them. The three ways of describing a characteristic that Fig. 8C compares with trait activation do not depend on whether the cognitive agent has noticed a cue to which the characteristic responds: a description in the agent profile, as researchers usually describe a participant's individual characteristics in silicon samples and digital twins[47,48]; a process statement presented in both conditions; and an if–then rule ("When …, I …"), which describes a characteristic as a stable relation between situations and behavior[19]. With none of these descriptions did the manipulation's effect differ significantly between the high and the low level of the characteristic (pooled interactions between −0.09 and 0.18; Fig. 8C), whereas it did with trait activation, in which MVCA introduces the description of the characteristic only after the cognitive agent has noticed such a cue. In the remaining way, we described how a cognitive agent with the characteristic would react; with this description, interactions appeared even in the two null tests of Study 4 (d_z = 0.48 and 0.27, both significant), because the description itself stated the response. The sentences that describe a characteristic in Study 4, in contrast, state what the cognitive agent is like or believes, and none of them states how the cognitive agent responds to the experimental scenario.

In Study 4, the heterogeneity also depended on what the sentences said about the cognitive agent. The null test in which the sentences about self-monitoring were introduced only in the fancy-hotel condition differed from the beer-price test only in what the sentences said about the cognitive agent: in both, the cue was that the beer comes from a fancy hotel rather than a run-down store. With Gemma 4 26B, the sentences about self-monitoring produced no interaction (d_z = 0.01), whereas the sentences stating the belief that higher costs justify higher prices did (d_z = 0.63; Supplementary Table 3).

Supplementary Note 13. Levels and activating cues of the individual characteristics (Study 4)

Each cognitive agent took part in every test at both levels of the individual characteristic; the assigned level applied to that test only, and the agent profile of the cognitive agent was the same at both levels. At either level, MVCA introduced the sentences describing the characteristic only when the cognitive agent's encoding contained the activating cue of the test, the cue to which the characteristic responds (Supplementary Table 3). In five of the six tests, this cue is the feature that the experiment manipulates. In the effort-and-indulgence test, by contrast, both conditions require the cognitive agent to earn the reward and differ in the effort required, so both contain the cue, having to earn the reward. In one null test, both conditions likewise contain the cue for self-monitoring, that someone the cognitive agent knows is present or will see the choice; in the other, the cue for self-monitoring is the cue of the beer-price test, that the beer comes from a fancy hotel rather than a run-down store, which only the fancy-hotel condition contains.

Supplementary Note 14. Repeated tests (Studies 3 and 4)

Because an LLM does not produce exactly the same text twice from the same information, the result of a test can vary by chance between repetitions. To assess this variation, we conducted the tests that had not reached significance the first time, six in Study 3 and three in Study 4, two more times with Gemma 4 26B and the same cognitive agents and materials (3,521 and 1,192 further cognitive simulations, respectively), so that each was conducted three times in all

(Supplementary Table 10). We decided on these repetitions after inspecting the first results of Studies 3 and 4.

Supplementary Note 15. Comparison across LLMs (Studies 1 to 4)

Because the CIs of the main text describe variation across cognitive agents, they do not capture variation across LLMs, which we examined by conducting each study with three LLMs. We used Qwen 3.7 Flash and Ling 3.0 Flash in all four studies with the first 30 of the 50 cognitive agents of the four studies and the same materials, instructions, and rules. We conducted each test once with each of them, except that we conducted Study 2 twice with Ling 3.0 Flash; because the second set of cognitive simulations covered only some of the experiments, the results with Ling 3.0 Flash are based on the first of the two sets (7,858 cognitive simulations; Supplementary Table 13). We accessed all three LLMs through OpenRouter, as released by their developers, and simulated the first 30 cognitive agents in the same numbering with each. Gemma 4 26B (Google) has 26 billion parameters in total; the size of Qwen 3.7 Flash (Alibaba) has not been published. The temperature of 0.2 determines how much the text produced by an LLM varies by chance.

To compare the LLMs on the same cognitive agents, we recomputed the results with Gemma 4 26B for these 30 cognitive agents (Supplementary Tables 7, 9 and 3; for Study 1, the values below). For these 30 and for all 50 cognitive agents, respectively, the pooled effects with Gemma 4 26B were d_z = 0.22 and 0.21 for support for the mediator in Study 1, 0.32 and 0.37 for the rating of the mediator in Study 1, 0.83 and 0.86 in Study 2, 0.22 and 0.23 in Study 3, and 0.39 and 0.40 in Study 4. We computed the estimates averaged over LLMs as described in the Results (Supplementary Table 4, note b), by averaging each cognitive agent's difference (Studies 1 and 2) or interaction (Studies 3 and 4) over the three LLMs before computing d_z. Agreement between two LLMs is the Spearman correlation of their d_z values across experiments or tests.

In Study 1, support for the mediator increased with each LLM (Gemma 4 26B, d_z = 0.21, 95% CI [0.08, 0.34]; Qwen 3.7 Flash, d_z = 0.24, 95% CI [0.12, 0.37]; Ling 3.0 Flash, d_z = 0.28, 95% CI [0.16, 0.41]; Supplementary Table 5), and so did the rating of the mediator (d_z = 0.37, 95% CI [0.17, 0.58]; d_z = 0.52, 95% CI [0.28, 0.76]; and d_z = 0.51, 95% CI [0.29, 0.72], respectively). For the 30 cognitive agents simulated with all three LLMs, the ratings of the mediator ranked the 60 experiments similarly across LLMs (Spearman ρ = 0.75 between Gemma 4 26B and Qwen 3.7 Flash, 0.49 between Gemma 4 26B and Ling 3.0 Flash, and 0.59 between Qwen 3.7 Flash and Ling 3.0 Flash). For the same 30 cognitive agents, support for the mediator ranked the experiments less similarly than the ratings did between Gemma 4 26B and the other two LLMs (ρ = 0.35 with Qwen 3.7 Flash and 0.45 with Ling 3.0 Flash) and about as similarly between Qwen 3.7 Flash and Ling 3.0 Flash (ρ = 0.58). Within each LLM, support and rating ranked the experiments similarly (ρ = 0.39 with Gemma 4 26B, 0.54 with Qwen 3.7 Flash, and 0.51 with Ling 3.0 Flash; 0.46 averaged over the three LLMs). The manipulation did not significantly increase the rating of a favorable evaluation of the target of the judgment with any LLM (d_z = 0.07 with Gemma 4 26B, 0.13 with Qwen 3.7 Flash, and 0.08 with Ling 3.0 Flash), and with each LLM, it increased the rating of the mediator more than this rating (difference d_z = 0.25, 95% CI [0.07, 0.43]; d_z = 0.34, 95% CI [0.15, 0.53]; and d_z = 0.38, 95% CI [0.20, 0.57], respectively).

In Study 2, the induced mediator shifted the response further in the predicted direction than its denial did with each LLM (Gemma 4 26B, d_z = 0.86, 95% CI [0.64, 1.09]; Qwen 3.7 Flash, d_z = 0.63, 95% CI [0.48, 0.79]; Ling 3.0 Flash, d_z = 0.77, 95% CI [0.56, 0.98]; Supplementary Table

7). Averaged over the three LLMs, this effect was d_z = 1.17 (95% CI [0.91, 1.43]); it was positive in 48 of the 52 experiments and significant in 44, and in no experiment was it significant in the opposite direction. The equally favorable process statement shifted the response about as much as the mediator did (averaged over the three LLMs, difference d_z = 0.10, 95% CI [−0.12, 0.32]), and the difference was not significant with any LLM. The unrelated process statement and the neutral process statement did not differ significantly in their effect on the response with any LLM (Supplementary Table 7). With Ling 3.0 Flash, however, the induced mediator did not shift responses more than an induced pleasant mood did (d_z = 0.04, 95% CI [−0.25, 0.33]; five experiments), whereas it did with Gemma 4 26B and Qwen 3.7 Flash and when averaged over the three LLMs (d_z = 0.50, 95% CI [0.23, 0.78]; Supplementary Table 7).

In Study 4, the interactions with Ling 3.0 Flash were smaller (pooled d_z = 0.06, 95% CI [−0.22, 0.35]; significant in the test of identity threat × collective self-esteem, d_z = 0.40), although tests with larger interactions with Gemma 4 26B also had larger interactions with Ling 3.0 Flash (Spearman ρ = 0.83). Because MVCA introduces the description of a characteristic only when the encoding mentions the cue to which the characteristic responds, part of this difference arose because the encodings that Ling 3.0 Flash produced mentioned this cue less often. In the time-versus-money and beer-price tests, these encodings mentioned the cue in 45 to 77% of the cognitive simulations in the treatment condition, compared with 100% with Gemma 4 26B, and in the test of identity salience × importance of family identity, in 33 to 53%, compared with 64%. In these three tests, the description of the characteristic was therefore introduced less often (Supplementary Table 3), and the interactions were smaller than with Gemma 4 26B for the same 30 cognitive agents (time versus money, −0.08 and 0.20; beer price, 0.29 and 0.60; identity salience, −0.17 and 0.13). In the foot-in-the-door test, the interaction was also smaller (0.22 and 0.53), although the encodings mentioned the cue about as often with both LLMs, so how often the cue was noticed does not account for this difference. With Qwen 3.7 Flash, by contrast, the interaction was in the predicted direction in all six tests and significant in the tests of identity threat × collective self-esteem (d_z = 0.64) and foot in the door × preference for consistency (d_z = 0.50) (pooled d_z = 0.27, 95% CI [0.03, 0.51]).

Supplementary Note 16. Stability of the manipulation's effect in individual cognitive agents (Study 4)

The manipulation's effect in an individual cognitive agent was not stable from one cognitive simulation to the next. In the null test of Study 4 in which the cue for self-monitoring was that someone the cognitive agent knows is present or will see the choice, the sentences about self-monitoring were introduced in both conditions, and the effect of the fancy hotel on the price of the beer was similar at the high and the low level of self-monitoring (d_z = 1.34 and 1.43). If each cognitive agent's effect were stable, its effects at the two levels should therefore have correlated across the 50 cognitive agents; they did not (r = −0.11). This instability matters less for tests of process theories, which in consumer psychology usually concern the average effect of a manipulation, than for digital twins, which are evaluated against the responses of particular human participants[48]. We decided on this analysis after inspecting the first results of Study 4.

Supplementary Note 17. Identification test and search for names of published findings

For each experiment, the identification test used the Study 1 cognitive simulations that Gemma 4 26B carried out for the first three cognitive agents that took part in the experiment. For both conditions, we extracted the exact input that the LLM had received for each of the 12 processes

(the system prompt, any sentences about the cognitive agent's individual characteristics, the earlier process statements, and, for some processes, the scenario, the names of drives, or the question) and, for comparison, the scenario followed by the question. In total, we presented 4,680 inputs to the separate LLM, DeepSeek V4.1 Flash: 360 (60 experiments × three cognitive agents × two conditions) for each of the 12 processes and 360 for the scenario followed by the question. With each input, we instructed DeepSeek V4.1 Flash to decide whether the input came from a published experiment in psychology or consumer research and, if so, to name the experiment and state its result, or otherwise to answer that it could not tell. In a second LLM call, DeepSeek V4.1 Flash received the answer together with the citation of the actual experiment, the statement of its theory, and its predicted direction, and judged whether the named experiment was the actual one and whether the stated result matched the published result. We counted an answer as identifying the experiment only when the named experiment was judged to be the actual one, so answers that stated the published result but named another experiment did not count; these were 25 of the 360 answers to the scenario and the question (Supplementary Table 11). As the measure of how identifiable each experiment was, we computed the proportion of the six answers to the scenario and the question, three in each condition, that identified it. Except in the identity-threat experiment, whose scenario describes a film tied to the cognitive agent's gender, the scenario and the question of a condition are the same for every cognitive agent, so the three answers in a condition came from the same input.

When the LLM carried out the ten processes between encoding and judgment, it did not receive the scenario, only the cognitive agent's encoding and the process statements derived from it; for the appraisal of goal relevance, it also received the names of some drives that MVCA had selected from words in the scenario. From the information for these processes, the separate LLM identified the experiment in a median of 18% of its answers, ranging from 0% for affective response to 25% for categorization; for the comparison of alternatives, the first-person account, and the coherence review, which immediately precede judgment, it did so in at most 7% (Supplementary Table 11). For judgment, the LLM received the question together with the earlier process statements, and from this information, the separate LLM identified the experiment in 16% of its answers, less than half as often as from the scenario and the question (38%).

With each LLM and on either measure of Study 1, the increase in the mediator did not grow with how often an experiment was identified (details in Supplementary Table 11). For support for the mediator, the correlation between how often an experiment was identified from the scenario and the question and the increase in support was $\rho = 0.14$ with Gemma 4 26B, 0.09 with Qwen 3.7 Flash, and 0.12 with Ling 3.0 Flash (all $P \geq 0.30$); for the second measure of Study 1, the rating of the mediator by the separate LLM, this correlation was likewise not significant with any LLM (Spearman $\rho = -0.01$ for Gemma 4 26B, 0.16 for Qwen 3.7 Flash, and 0.03 for Ling 3.0 Flash; all $P \geq 0.22$). In the 23 experiments that the separate LLM never identified from the scenario and the question, support for the mediator, averaged over the three LLMs, increased significantly in 9 (Supplementary Table 11), and its pooled increase was 0.37, compared with 0.49 in the 37 experiments identified at least once. With each LLM, the pooled increase in support in these 23 experiments was positive, and it was significant with two of the three (Gemma 4 26B, $d_z = 0.15$, 95% CI [−0.06, 0.37]; Qwen 3.7 Flash, $d_z = 0.18$, 95% CI [0.01, 0.34]; Ling 3.0 Flash, $d_z = 0.30$, 95% CI [0.05, 0.55]); in the experiments identified at least once, it was $d_z = 0.25$, 0.29, and 0.27, respectively. For the rating of the mediator, the pooled increase in the never-identified experiments was significant with each LLM and about as large as in the experiments identified at least once (Gemma 4 26B, $d_z = 0.40$ and 0.36; Qwen 3.7 Flash, 0.43 and 0.57; Ling 3.0 Flash,

0.55 and 0.49; Supplementary Table 11). When an experiment was instead counted as identified if the separate LLM named it from the information that the LLM received for judgment, the two groups did not differ significantly in the increase in support (averaged over the LLMs, $d_z = 0.37$ for the 41 experiments never identified and 0.60 for the 19 identified at least once) or in the effect of the induced mediator on the response in Study 2 (1.14 for 35 experiments and 1.24 for 17; Supplementary Table 11).

Conversely, identification did not ensure a predicted effect. The separate LLM identified the normative-message experiment in all six of its answers to the scenario and the question and the identity-divergence experiment in three of the six, yet neither experiment showed a significant interaction in the predicted direction in Study 3 with any of the three LLMs (Supplementary Tables 9 and 11). Identification was thus neither necessary for the increase in the mediator nor sufficient for a predicted effect.

In the search for names of published findings, we looked for the following names in the process statements of the 47 experiments whose finding has an established name: endowment effect (in the endowment, emotional-attachment, possession-and-valuation, and buying-and-selling-prices experiments; also "mere ownership" in the possession-and-valuation experiment); IKEA effect; foot in the door; door in the face or reciprocal concessions; transaction utility; pennies a day; decision quicksand; feature fatigue; regulatory fit; goal, moral, or self-licensing; boomerang effect; red sneakers effect; labeling effect; competitive altruism; mere touch; payment transparency; side-effect effect, also under its name formed from the original author's surname (side effects and intent); moral typecasting; tempting fate; gambler's fallacy (retrospective gambler's fallacy); consumerism (consumer label and trust); overhead aversion; affect-rich, probability neglect, or probability weighting (affect-rich prizes); verb aspect, grammatical aspect, or imperfective (verb aspect and intent); reason-based choice (choosing vs rejecting); asymmetry of similarity or similarity asymmetry; framing effect or Asian disease (disease framing); anchoring, anchoring effect, or anchoring bias (in the anchoring and incidental-anchors experiments); prestige suggestion (quote attribution); norm of reciprocity; forbid/allow asymmetry; less is better or less is more; response-scale effect or scale range effect (response-scale range); trolley problem, footbridge dilemma, or footbridge problem (personal force); flag priming; money priming, currency priming, or system justification (currency priming); attribute substitution (heat priming); power is up (vertical position and power); embodied cognition (in the weight-and-importance and physical-warmth experiments; also weight as embodiment in the weight-and-importance experiment); analytic thinking (analytic thinking and belief); compensatory control (structure and goal pursuit); and mortality salience or terror management. A match to a name counted only if the matched words did not occur in the scenario text of the same cognitive simulation. The words that refer to research in general were "experiment(s)," "participant(s)," "researcher(s)," "hypothesis" or "hypotheses," "the study," "this study," "studies show," "control condition" or "control group," "treatment condition" or "treatment group," "manipulation," "psychology," "psychologist(s)," "published," and "et al." In the search for authors, we looked for the surnames of all authors of the 60 original articles.

A process statement named the experiment's finding in 20 of the 4,648 cognitive simulations of these 47 experiments with Gemma 4 26B (0.4%), in 177 of 2,802 with Qwen 3.7 Flash (6.3%), and in 110 of 2,801 with Ling 3.0 Flash (3.9%; Supplementary Table 12). Most of these cognitive simulations came from the experiments on the endowment effect, on the framing of a disease program, and on the trolley dilemma. The authors of an original article were named in 2

of the 12,983 cognitive simulations of Study 1, both carried out with Qwen 3.7 Flash in the disease-framing experiment. Words that refer to research in general appeared in 6.12% of the process statements with Gemma 4 26B (3,606 of 58,920), in 10.83% with Qwen 3.7 Flash (3,842 of 35,460), and in 4.55% with Ling 3.0 Flash (1,612 of 35,446). In the 45 experiments whose scenarios contain none of these words, they appeared in 4.72%, 8.80%, and 3.67% of the process statements, respectively. The most frequent of these words were “researchers,” “manipulation,” “experiment,” “the study,” and “participant(s).”

Supplementary Note 18. Experiments with replication outcomes (Studies 1 and 2)

We selected 28 published experiments whose original effects had been re-examined in large preregistered replications: the Many Labs projects[49–52], a Registered Replication Report[53], the Social Sciences Replication Project[54], a registered report of the Psychological Science Accelerator[55], and a preregistered replication at three sites[56]. An experiment qualified if its manipulation could be conveyed in a scenario in the second person, if its response was a judgment, a choice, a rating, or an amount, and if a mediator that explains the effect had been proposed for it. Following the conclusions of the replication reports, we selected 14 experiments whose effects replicated and 14 whose effects did not; these 28 experiments test effects other than those of the other 32 experiments. We selected and prepared them in two sets: first 10 experiments (6 whose effects replicated and 4 whose effects did not), and then 18 (8 and 10). Supplementary Table 1 lists the experiments, their replications, and whether each effect replicated.

We prepared the materials of these experiments by the procedure used for the other 32 experiments (Methods); where the replication projects had published their materials, the scenario and the question follow these materials. As in the other experiments, no process agent received the cognitive agent's current mood or its levels on the characteristics of the agent profile that correspond to constructs of the theory under test (Supplementary Note 4). Before conducting any cognitive simulation in an experiment, we assigned its mediator to the appraisal process in which, according to the theory, the mediator arises, following the assignment rule in the deposited specification of MVCA. We deposited the materials and these assignments in time-stamped files (13 October 2026) and conducted the cognitive simulations with MVCA as deposited, with the same three LLMs and the same cognitive agents as in the other experiments.

To check that each mediator statement stated the mediator and each denial statement denied it, the separate LLM (DeepSeek V4.1 Flash, temperature 0) rated each statement alone, three times, on the 1-to-7 scale of the mediator. A mediator statement met its target when its mean rating was 6.0 or higher (2.0 or lower for a statement at the low end of the scale), and a denial statement met its target when its mean rating was 2.0 or lower (6.0 or higher for a statement at the low end). For the first 10 experiments, we conducted this check after their cognitive simulations, as a record, and, as the analysis plan specifies, did not alter their materials. In these experiments, 39 of the 40 statements met their targets. The exception was one of the two denial statements of the tempting-fate experiment (“I can't really picture him calling on me while I sit there without an answer”), which received a mean rating of 5.33 and remained in the materials. For the other 18 experiments, we conducted the check before fixing their materials, and a statement that missed its target was to be revised once. All 72 statements met their targets, so no statement was revised.

We fixed the analysis comparing the experiments whose effects replicated with those whose effects did not after conducting the cognitive simulations in the first 10 experiments and before

conducting any in the other 18, and we deposited it in the same way. For each experiment, we computed the effect of the manipulation on the response (treatment minus control, as in Study 1), the two Study 1 measures of the mediator in the process statements, and the effect of the induced mediator rather than its denial on the response (as in Study 2), each averaged over the three LLMs as in the main text and signed so that positive values indicate the published direction. We counted the cognitive agents as showing an experiment's original effect when the 95% CI of d_z for the treatment-minus-control difference in the response, averaged over the three LLMs, lay above zero, and we tested whether showing the effect was associated with the replication outcome with a one-sided Fisher's exact test, because, in the deposited plan, we had predicted that the cognitive agents would show the effect where it had replicated rather than where it had not. Within each of the two groups, we pooled each effect with a random-effects model (restricted maximum likelihood, Hartung–Knapp–Sidik–Jonkman CI), and we tested the difference between the two pooled estimates with $z = (\hat{d}_1 - \hat{d}_2)/\sqrt{(SE_1^2 + SE_2^2)}$. In an experiment, we counted the proposition that the manipulation affects the mediator as supported when support for the mediator or the ratings of the mediator increased significantly (Study 1), and the proposition that the mediator affects the response as supported when the induced mediator shifted the response significantly more than its denial (Study 2; Supplementary Table 8). The cognitive agents showed the original effect in 9 of the 14 experiments whose effects replicated and in 3 of the 14 whose effects did not (Study 2 and Supplementary Table 8). Support for the mediator, the first measure of Study 1, ranked the two groups of experiments in the same order as the ratings of the mediator did: the pooled increase in support was $d_z = 0.51$ (95% CI [−0.06, 1.09]) in the experiments whose effects replicated and $d_z = 0.27$ (95% CI [0.00, 0.54]) in those whose effects did not (difference, P = 0.41; Supplementary Table 8).

Supplementary Note 19. Effect of the manipulation on the response (Study 1 cognitive simulations)

The four propositions concern the steps from the manipulation to the response, so each study tests one of these steps rather than the manipulation's effect on the response alone. In the cognitive simulations of Study 1, however, the cognitive agents also responded to the question of each experiment, so this effect could be computed for every experiment. Averaged over the three LLMs, the manipulation shifted the response in the published direction across the 60 experiments (pooled $d_z = 0.33$, 95% CI [0.13, 0.54]). The main text and Supplementary Table 8 report this effect in the 28 experiments whose effects large preregistered replications re-examined. In the other 32 experiments, the pooled effect was $d_z = 0.24$ (95% CI [0.001, 0.47]), and the effect was significant in 11 of them; with each LLM separately, the pooled effect was positive (Gemma 4 26B, $d_z = 0.10$; Qwen 3.7 Flash, 0.12; Ling 3.0 Flash, 0.12).

Supplementary Note 20. Results of the individual tests of Study 3

In the social-exclusion test in which the product did or did not signal group membership[57], the interaction was significant with Gemma 4 26B but not when each cognitive agent's interaction was averaged over the three LLMs (Supplementary Table 9).

In the powerlessness test, the product was exclusive or mass-produced, and according to the theory, an exclusive product signals status and thus serves the induced desire to feel more powerful[58]. Under the motive–means rule, the desire shaped the cognitive agents' later process statements only when the LLM determined, solely from their encoding and the statement of the desire, that the product would serve it. Averaged over the three LLMs, the interaction was $d_z =$

0.72; it was significant with Qwen 3.7 Flash and Ling 3.0 Flash but not with Gemma 4 26B (Supplementary Table 9). With Qwen 3.7 Flash and Ling 3.0 Flash, the cognitive agents thus inferred from what they had noticed that an exclusive product, but not a mass-produced one, would serve a desire to feel more powerful.

In the decision-difficulty test, we induced not the belief that the choice was important, which Study 1 measured, but the feeling that the choice was difficult, which according to the theory makes a choice seem important[45]. When a warning stated that the website's small font would make the flight details hard to make out, the cognitive agents could attribute this feeling to the font. As predicted, the induced feeling increased the time that the cognitive agents would spend on the choice more without the warning than with it (interaction $d_z = 0.61$, averaged over the three LLMs). Because we had formulated the discounting rule for this experiment, after the cognitive agents failed to discount this feeling (Supplementary Note 9), this interaction is not an independent test of the discounting rule.

The affect-reliance test, by contrast, examined the discounting rule in an experiment for which we had not designed it. In this test, the induced process statement was that the feeling evoked by an advertisement was a good guide to the dictionary that it advertised[59]. According to the theory, relying on this feeling should raise evaluations of the dictionary when the advertisement is attractive with weak arguments but not when it is plain with strong arguments. This interaction was also significant ($d_z = 0.58$). Through the discounting rule, which applies a general principle of affect-as-information theory[13], the cognitive agents thus showed the predicted moderation in both tests, including the one for which we had not designed the rule.

In the pennies-a-day test[60], the requested daily donation was low or high, and the standard-comparison rule applied (interaction $d_z = -0.10$). In the normative-message test[61], feedback comparing a household's energy use with that of its neighbors did or did not include an approving smiling face, and no rule applied (interaction $d_z = 0.28$). With Gemma 4 26B, the induced process statement shifted responses at both levels of the situational moderator in both tests (Supplementary Table 9), whereas the theories predict that the moderator weakens the mediator's effect at one of them.

Supplementary Note 21. The 60 experiments and their materials

The 60 experiments vary in the kind of mediator, such as a view of oneself, a desire, a feeling, or an inference about others, and in the domain of the response, such as prices, compliance with requests, judgments of intent and blame, or products linked to one's identity (Supplementary Table 1). Of the 60 experiments, 22 were published in the Journal of Consumer Research, the Journal of Marketing Research, or the Journal of Consumer Psychology, seven in the Journal of Personality and Social Psychology, six in Psychological Science, four in Science, and the other 21 in 15 other journals and two edited volumes (Supplementary Table 1).

Where the original authors did not measure the mediator, we formulated the two mediator statements from their description of it; Supplementary Table 1 shows one of the two. When a scenario presupposed an attribute that only some individuals have, such as a particular gender, only the cognitive agents whose agent profile has this attribute took part (numbers of cognitive agents per experiment in Supplementary Table 5). When the original measure of the response comprised several items, the response was the average across these items.

Supplementary Note 22. Confidence intervals and variation between tests

Cognitive agents are the units of analysis; for an interaction, each cognitive agent contributes the difference between two within-agent differences. The 95% CI of d_z, $d_z \pm t(0.975, n - 1)/\sqrt{n}$, is the CI of the mean difference, expressed in units of the standard deviation of the differences, and it therefore excludes zero exactly when a paired t test of the differences yields $P < 0.05$. We computed I^2 for the interactions of Study 3 from the restricted-maximum-likelihood estimate of the between-test variance. In the joint sign-flip permutation test of Study 3, all interactions of a cognitive agent receive the same random sign, so that the dependence between tests that share cognitive agents is preserved; the test statistic was the number of tests whose 95% CI lay above zero (Supplementary Note 11).

Supplementary Note 23. Deposited specification and criteria

The specification of MVCA that we deposited in a time-stamped repository on 7 October 2026 (repository version f1ae3f2), before conducting any cognitive simulation of the four studies, included the appraisal process in which, according to the theory, each mediator arises and the general rule, if any, that applies to it; it also defined the cognitive agents of the four studies, the designs, the primary analyses, and the criteria for each study. For Study 3, it further stated that the tests would be reported in groups defined by how, according to the theory, the situational moderator alters the mediator's effect (Supplementary Note 10). The criteria concerned the results with Gemma 4 26B. In every study, one criterion was a pooled effect whose 95% CI excludes zero. Study 1 had three additional criteria: agreement across experiments between support for the mediator and the ratings of the mediator by the separate LLM, measured by their Spearman correlation (we set no minimum value in advance); a larger increase in the mediator-minus-denial score for the experiment's own statements than for the statements of other experiments; and no increase in pleasantness. Study 2 had two additional criteria: a larger effect of the induced mediator than of a pleasant mood, and no difference in the response between a neutral process statement about the target of the judgment and a process statement unrelated to the decision. For Study 4, the additional criterion was a significant interaction in at least half of the six tests and in neither null test.

With Gemma 4 26B, Studies 1, 2, and 4 met these criteria in the 32 experiments that the deposited specification comprised. In these experiments, the Spearman correlation between support for the mediator and the ratings was 0.46, and the mediator-minus-denial score, averaged over the ten process statements, increased more for the experiment's own statements than for the statements of other experiments (difference $d_z = 0.13$, 95% CI [0.01, 0.24]). Study 3, however, did not meet its criterion: the pooled interaction was positive, but its CI included zero ($d_z = 0.23$, 95% CI [−0.03, 0.48]).

Supplementary Table 1.

Published experiments simulated in the four studies. For each experiment, the table lists the label used in the text, figures, and tables, with the identifier used in the deposited data; the original experiment and its source and, for the 28 experiments that large preregistered replications re-examined, the replication and whether the effect replicated; one of the two mediator statements; the dependent measure; the appraisal process in which, according to the theory, the mediator arises; and the studies that included the experiment. Studies: 1, manipulation → mediator; 2, induced mediator → response; 3, situational moderator × mediator; 4, individual characteristic × manipulation. Placeholders in braces are filled from the agent profile (for example, the cognitive agent's gender).

Label (identifier)	Experiment (source)	Mediator statement (one of two)	Dependent measure	Appraisal process in which, according to the theory, the mediator arises	Studies
Brand extension (aaker_keller_brand_extension)	A brand's new product in a close vs a distant category[40]	To me, this new product is a natural fit for what this brand makes.	What do you think of this new product? Answer two questions on scales from 1 to 7: How would you rate its overall quality (1 = inferior, 7 = superior)? How likely would you be to try it (1 = extremely unlikely, 7 = extremely likely)?	categorization	1, 2
Anthropomorphized products (aggarwal_mcgill_anthropomorphism)	Smiling vs frowning car under a first-person (human) presentation[62]	Looking at this car, I feel like I'm looking at a person, not a machine.	How much do you agree with each of the following statements about this car?	categorization	1, 2
Consumer label and trust (bauer_cuing_consumerism_trust)	Being called a consumer vs an individual in a shared water shortage[63], replicated[50]	To me, the other three are rivals for the same water, each out for themselves.	How much do you trust the other parties involved to use less water?	social inference	1, 2
Nonconforming dress (bellezza_red_sneaker)	A red vs a black bow tie at a black-tie party and inferred standing[64]	It seems to me Daniel can afford to do whatever pleases him.	What is your impression of Daniel? Answer two questions on scales from 1 to 7: How likely is Daniel to be one of the top members of the golf club (1 = not likely at all, 7 = extremely likely)? Do you think Daniel has won golf competitions or prizes in the past (1 = not likely at all, 7 = extremely likely)?	social inference	1, 2
Identity divergence (berger_heath_identity_divergence)	Abandoning a taste when dissimilar others adopt it, by identity relevance of the domain[65]	I don't want anyone taking me for one of those new buyers.	How would these people buying it change your own use of it over the next few months? Answer from 1 to 7: 1 = use it much less or stop using it, 4 = no change, 7 = use it much more.	social inference	1, 3

Label (identifier)	Experiment (source)	Mediator statement (one of two)	Dependent measure	Appraisal process in which, according to the theory, the mediator arises	Studies
Emotion and sharing (berger_milkman_virality)	More vs less anger-evoking version of the same story and sharing[66]	Reading this has left me feeling keyed up and full of energy.	How likely would you be to share this story with others?	affective response	1, 2
Labeling (burger_caldwell_labeling)	Labeling after a small act[67]	I'm the kind of person who pitches in for people in need.	How likely are you to agree to give two hours to the food drive this weekend?	self-relevance	1, 2
Buying and selling prices (carmon_ariely_forgone)	Buyers' and sellers' prices for a championship ticket[68]	What I keep thinking about is the game itself and being there in person.	What price do you give the caller? Give a dollar amount from $0 to $10,000.	goal relevance	1, 2
Flag priming (carter_flag_priming)	Photographs with vs without American flags and political attitudes[69], not replicated[49]	To me, those American flags stand for the Republican Party and conservative values.	How much do you agree or disagree with each of the six statements on the page, from 1 = Completely disagree to 7 = Completely agree?	categorization	1, 2
Currency priming (caruso_currency_priming)	A faint image of banknotes vs a blurred image behind a survey page and endorsement of the social system[70], not replicated[49]	I believe the free market is fair: what people earn in it, they deserve.	How much do you agree or disagree with each of the six statements on the page, from 1 = Strongly Disagree to 7 = Strongly Agree?	causal inference	1, 2
Door in the face (cialdini_door_in_the_face)	Door-in-the-face request sequence[71]	He came down from his first request, so I feel I owe him a yes.	How likely are you to agree to chaperone the zoo trip?	social inference	1, 2
Incidental anchors (critcher_gilovich_incidental_anchors)	A high vs a low number in a phone's model name and the estimated share of its sales in the United States[72], not replicated[50]	The phone's number, 97, is the figure my percentage estimate starts from.	If you had to estimate, what percentage of the phones of this model sold during the initial introduction period will be purchased in the United States, as opposed to Western Europe? Give a percentage from 0 to 100.	categorization	1, 2
Goal progress (fishbach_dhar_licensing)	Perceived goal progress and an indulgent parting gift[73]	I feel like I've come a long way toward my ideal weight.	Which would you take as your parting gift, the apple or the chocolate bar?	goal relevance	1, 2
Foot in the door (freedman_fraser_foot_in_the_door)	A small petition first, then a large lawn sign requested by another group[74]	I'm the kind of person who does my part for good causes around here.	How likely are you to agree to have the sign put up in your front yard?	self-relevance	1, 2, 4

Label (identifier)	Experiment (source)	Mediator statement (one of two)	Dependent measure	Appraisal process in which, according to the theory, the mediator arises	Studies
Shaken self-view (gao_wheeler_shaken_self)	Shaken confidence in being competent and the choice of competent products[75]	Right now I'm honestly not sure I'm as competent as I process statement.	For each of the three pairs, which prize would you prefer if you won it? Answer each on a scale from 1 to 7, where 1 = definitely the first option and 7 = definitely the second option.	self-relevance	1, 2
Analytic thinking and belief (gervais_norenzayan_analytic_religion)	Viewing photographs of Rodin's The Thinker vs a discus thrower and belief in God[76], not replicated[54]	Right now I am thinking slowly and analytically, questioning my gut reactions before accepting them.	We are interested in your belief in God, from 0 to 100. To clarify, if you are certain that God does not exist, you would put "0." If you are certain that God does exist, you would put "100." So, how strongly do you believe in God?	affective response	1, 2
Vertical position and power (giessner_schubert_verticality_power)	A long vs a short vertical line between a manager and his employees in an organization chart and judgments of his power[77], not replicated[50]	Sitting so far above his employees on the chart, Manager A seems very powerful to me.	Please answer the following questions about the manager you just read about, on a scale from 1 (totally disagree) to 7 (totally agree). (1) I think that Manager A is dominant. (2) I think that Manager A has a strong leader personality. (3) I think that Manager A is self-confident. (4) I think that Manager A has a lot of control in the company. (5) I think that Manager A holds a very high status within the company.	social inference	1, 2
Overhead aversion (gneezy_keenan_gneezy_overhead_aversion)	A charity whose overhead is covered by another donor vs paid from donations[78], replicated[54]	All of the money I give charity: water would go to the cause itself.	How likely is it that you would give the $100 to "charity: water" rather than to "Kids Korps"?	goal relevance	1, 2
Pennies a day (gourville_pennies_a_day)	Pennies-a-day framing of a donation[60]	This donation feels to me like a small everyday expense, like a coffee.	What is the likelihood you would agree to donate the amount requested?	categorization	1, 2, 3
Moral typecasting (gray_wegner_moral_typecasting)	An adult vs a baby causing the same harm[79], replicated[50]	I see Sam as someone who can plan and control what he does.	How responsible is Sam for his behavior?	social inference	1, 2

Label (identifier)	Experiment (source)	Mediator statement (one of two)	Dependent measure	Appraisal process in which, according to the theory, the mediator arises	Studies
Mortality salience (greenberg_mortality_salience_worldview)	Writing about one's own death vs watching television and the evaluation of authors who praise vs criticize the United States[80], not replicated[52]	I feel a strong need to defend my country and the values I stand for.	Taking together how intelligent, how knowledgeable and how likable you judge each author to be, how much more favorably do you evaluate the author of Essay 1 than the author of Essay 2, from -8 (the author of Essay 2 far more favorably) through 0 (both equally) to +8 (the author of Essay 1 far more favorably)?	self-relevance	1, 2
Personal force (greene_personal_force)	Pushing a man off a footbridge with a pole vs dropping him through a trap door with a switch to save five others[81], replicated[55]	Picturing Joe pushing the workman off with that pole fills me with horror.	To what extent is this action morally acceptable?	affective response	1, 2
Going green (griskevicius_going_green)	Status motive and environmentally friendly products when shopping in public or in private[82]	Right now I really want to get ahead of the people I know.	For each pair, which of the two products is more attractive to you? Rate each pair from 1 to 9, where 1 = definitely Product A and 9 = definitely Product B.	social inference	1, 3
Verb aspect and intent (hart_albarracin_verb_aspect_intent)	A shooting described as ongoing ("was firing") vs completed ("fired")[83], not replicated[53]	I picture the shooting as it unfolds, moment by moment, in clear detail.	To what extent did Westmoreland intentionally cause harm to McElroy?	categorization	1, 2
Future self and saving (hershfield_future_self_saving)	Seeing one's aged self and saving[84]	The person I'll be at 70 feels like the same me, just older.	How would you divide the $1,000 among the four uses? Give a dollar amount for each use; the four amounts must add up to $1,000.	self-relevance	1, 2
Less is better (hsee_less_is_better)	A $90 scarf vs a $110 coat as a gift, each judged on its own, and the giver's generosity[85], replicated[50]	To me, $90 is close to the most that store charges for a scarf.	How generous do you think your friend was?	categorization	1, 2
Norm of reciprocity (hyman_sheatsley_norm_of_reciprocity)	Admitting a rival country's reporters, asked after vs before admitting one's own country's reporters there[86], replicated[49]	Since I said North Korea should admit our reporters, I feel we must admit theirs.	Do you think the United States should let newspaper reporters from North Korea come in here and send back to their papers the news as they see it? Give the likelihood, from 0 to 100, that you answer Yes.	social inference	1, 2

Label (identifier)	Experiment (source)	Mediator statement (one of two)	Dependent measure	Appraisal process in which, according to the theory, the mediator arises	Studies
Anchoring (jacowitz_kahneman_anchoring)	A high vs a low stated bound and the estimated number of babies born per day in the United States[87], replicated[49]	I start from the 50,000 figure and come down only a little from it.	How many babies do you think are born in the U.S. each day? Give a whole number.	categorization	1, 2
Weight and importance (jostmann_weight_importance)	Holding a heavy vs a light clipboard and the judged importance of a committee listening to students[88], not replicated[51]	How this committee decides about my grant strikes me as a weighty, serious matter.	How important is it for you that the committee would listen to the opinion of the students? Answer from 1 to 7: 1 = not at all important, 2 = unimportant, 3 = somewhat unimportant, 4 = neutral, 5 = somewhat important, 6 = important, 7 = very important.	goal relevance	1, 2
Endowment effect (kahneman_knetsch_endowment)	Owners' and buyers' prices for the same mug[89]	To me, selling this mug would feel like losing something, not like making money.	What price do you write on the form? Give an amount from $0.00 to $9.50.	goal relevance	1, 2
Structure and goal pursuit (kay_structure_goal_pursuit)	Reading that leaves grow in orderly vs random patterns and interest in pursuing a long-term goal[90], not replicated[50]	I am confident that the effort I put in leads to predictable results.	How interested are you in each of the two things the page asks about, from 1 = Not very interested to 7 = Extremely interested?	causal inference	1, 2
Effort and indulgence (kivetz_simonson_indulgence)	Program requirements and choice between a luxury and a necessity reward[91]	After that many rentals, I'd feel I had truly earned a treat.	Which of the two programs would you prefer to join?	causal inference	1, 2, 4
Side effects and intent (knobe_side_effect_intentionality)	A harmful vs a helpful side effect of a profit-seeking decision[92], replicated[50]	To me, the chairman deserves real blame for what his program did here.	How much do you agree with the statement, "The chairman brought about what happened to the environment intentionally."?	causal inference	1, 2
Goal fluency (labroo_lee_regulatory_fluency)	A related prime ad and purchase intent for a goal-compatible target brand[93]	This spray was easy for me to take in; it made sense right away.	How likely would you be to buy Raid Ant & Roach Killer?	affective response	1, 2
Quote attribution (lorge_curtiss_quote_attribution)	A quotation attributed to a liked vs a disliked public figure[94], replicated[49]	Since George Washington said these words, they carry real weight with me.	How much do you agree with this quote? Answer on the scale from 1 = strongly agree to 9 = strongly disagree, where 5 = neither agree nor disagree.	social inference	1, 2

Label (identifier)	Experiment (source)	Mediator statement (one of two)	Dependent measure	Appraisal process in which, according to the theory, the mediator arises	Studies
Social exclusion (mead_social_exclusion)	Social exclusion and paying for a prospective partner's favorite food, seen or unseen by the partner[57]	I really want Lin to like me when we meet today.	How much would you be willing to pay for a serving of each dish at the tasting? Rate each dish from 1 to 9, where 1 = $0 to $1, 2 = $2, 3 = $3, and so on up to 9 = $9.	social inference	1, 3
Time versus money (mogilner_aaker_time_money)	A lemonade-stand sign mentioning time vs money[95]	Having this lemonade feels personal to me, like it reflects who I am.	What is your attitude toward C & D's lemonade? Rate it on each of the three scales below.	self-relevance	1, 2, 4
IKEA effect (norton_ikea_effect)	Assembling a product oneself and willingness to pay for it[96]	Building this box myself, and getting it right, makes me feel really capable.	What is the most you would pay for this box, in dollars?	self-relevance	1, 2
Retrospective gambler's fallacy (oppenheimer_monin_retrospective_gamblers_fallacy)	Seeing three sixes vs two sixes and a three rolled at a casino[97], replicated[49]	I think a roll like three sixes only turns up after many tries.	How many times do you think the man had rolled the dice before you walked by? Give a whole number.	causal inference	1, 2
Mere touch (peck_shu_mere_touch)	Mere touch and perceived ownership[32]	Somehow this mug already feels like it's mine, as if I owned it.	What is the most you would be willing to pay for this mug? Choose an amount from $0.00 to $9.00, in steps of 25 cents.	self-relevance	1, 2
Affect reliance (pham_avnet_regulatory_fit)	Ideals vs oughts and reliance on feelings about an ad[59]	How this ad makes me feel is a good guide to this dictionary.	How do you evaluate the Wordwise Desk Dictionary? Answer on three scales from 1 to 7: 1 = bad, 7 = good; 1 = unfavorable, 7 = favorable; 1 = dislike, 7 = like.	affective response	1, 2, 3
Possession and valuation (reb_connolly_ownership)	Possession without ownership and valuation[98]	To me this chocolate bar already feels like it's mine, even though it isn't.	What is the smallest amount of money on the sheet that you would take instead of the chocolate bar? Give an amount from $0.10 to $4.00; if you would rather have the chocolate bar at every amount on the sheet, give 4.00.	self-relevance	1, 2
Identity salience (reed_identity_salience)	Self-importance of family identity and an identity-framed product[99]	Being part of my family is central to who I am right now.	How likely would you be to buy this product?	self-relevance	1, 3, 4

Label (identifier)	Experiment (source)	Mediator statement (one of two)	Dependent measure	Appraisal process in which, according to the theory, the mediator arises	Studies
Tempting fate (risen_gilovich_tempting_fate)	Not having vs having done the reading for class[100], replicated[50]	I can easily picture him calling on me while I sit there with no answer.	How likely do you believe it is that the professor will call on you?	affective response	1, 2
Affect-rich prizes (rottenstreich_hsee_affect_rich_risk)	A 1% chance vs certainty of a kiss from one's favorite movie star or $50[101], not replicated[50]	Even a 1% chance of kissing my favorite star gives me a real thrill.	How likely are you to choose the option involving your favorite movie star rather than the option involving the $50 in cash? Give a number from 0 to 100.	affective response	1, 2
Powerlessness (rucker_galinsky_powerlessness)	Powerlessness and willingness to pay for an item described as exclusive or mass-produced[58]	Right now I'd really like to feel more powerful and in control.	What is the most you would bid for this photograph, in US dollars? Give a whole number.	goal relevance	1, 3
Forbid vs allow (rugg_forbid_allow)	Asking whether speeches against democracy should be forbidden vs allowed[102], replicated[49]	To me, forbidding these speeches outright would be a harsh, heavy-handed step.	Think of the answer, Yes or No, that you give to this survey question. How likely is it that your answer is the one under which public speeches against democracy could still be given in the United States? Give a number from 0 to 100.	categorization	1, 2
Normative message (schultz_boomerang_norms)	Neighborhood-average feedback and a smiling face for a below-average household[61]	Since my household uses less than the neighbors, I have room to use more.	Over the next three weeks, how much electricity do you expect your household to use each day, compared with the past two weeks? Answer from 1 to 7: 1 = much less, 4 = about the same, 7 = much more.	social inference	1, 3
Response-scale range (schwarz_category_range_scales)	Answer options covering a high vs a low range of hours and reported daily television viewing[103], replicated[49]	These answer options tell me that people usually watch three or four hours a day.	How much television do you watch per day? How likely are you to choose an option that starts at two and a half hours or more, rather than one that ends at two and a half hours or less? Give a number from 0 to 100.	social inference	1, 2
Decision difficulty (sela_berger_decision_quicksand)	Hard-to-read options and time spent on a trivial choice[45]	Getting this flight choice right seems more important than I first process statement.	Which flight would you book, and about how many seconds would you spend looking over the two flights before choosing?	causal inference	1, 3

Label (identifier)	Experiment (source)	Mediator statement (one of two)	Dependent measure	Appraisal process in which, according to the theory, the mediator arises	Studies
Choosing vs rejecting (shafir_choosing_rejecting)	Awarding vs denying sole custody to one of two parents[104], not replicated[50]	To decide who should get custody, I'm weighing each parent's strengths most heavily.	As a result of your decision, how likely is it that Parent B ends up with sole custody of the child? Give a number from 0 to 100.	goal relevance	1, 2
Emotional attachment (shu_peck_emotional_attachment)	Psychological ownership, affective reaction and the endowment effect[105]	I really feel like this pen is mine, my own personal pen.	What is the smallest amount of money you would take instead of the pen? Give an amount from $0.00 to $6.00; if you would rather have the pen at every amount on the form, give 6.00.	self-relevance	1, 2
Payment transparency (soman_payment_transparency)	Paying a month's expenses by credit card vs by check, then a discretionary purchase[106]	I feel like I haven't spent much this month and have money to spare.	How likely are you to buy the boxed set?	goal relevance	1, 2
Beer price (thaler_transaction_utility)	Beer on the beach: the same beer from a fancy hotel or a run-down store[44]	Beer at a place like that costs more, so I think a high price is fair.	What price do you tell your friend? Give a dollar amount from $0 to $50.	causal inference	1, 2, 4
Feature fatigue (thompson_feature_fatigue)	Number of product features and evaluation before use[107]	This player can do just about everything I could want it to do.	How likely would you be to subscribe to this video player?	categorization	1, 2
Similarity asymmetry (tversky_gati_similarity_asymmetry)	The less vs the more prominent country named first in similarity judgments[108], not replicated[50]	In each pair, what I know of the first country mostly holds for the second.	What similarity rating do you give each of the five items, from 1 = no similarity to 20 = maximal similarity?	categorization	1, 2
Disease framing (tversky_kahneman_disease_framing)	Outcomes of two disease programs described as lives saved vs deaths[109], replicated[49]	To me, Program A means a sure gain: 200 people will certainly be saved.	How likely are you to choose Program A rather than Program B? Give a number from 0 to 100.	goal relevance	1, 2
Identity threat (white_argo_identity_threat)	Social identity threat and products linked to the threatened identity[21]	Right now I'd rather not be lumped in with other {men/women}.	How do you evaluate this film? Answer on three scales from 1 to 9: 1 = unfavorable, 9 = favorable; 1 = dislike, 9 = like; 1 = bad, 9 = good.	self-relevance	1, 3, 4

Label (identifier)	Experiment (source)	Mediator statement (one of two)	Dependent measure	Appraisal process in which, according to the theory, the mediator arises	Studies
Physical warmth (williams_bargh_physical_warmth)	Holding a hot vs a cold therapeutic pack and choosing a reward for a friend rather than for oneself[110], not replicated[56]	Right now I feel warm and generous toward the people I care about.	How likely are you to tick the gift for a friend, the bottle of Snapple, rather than the ice-cream voucher for yourself? Give a whole number from 0 (certain to choose the ice-cream voucher for yourself) to 100 (certain to choose the Snapple for a friend).	affective response	1, 2
Heat priming (zaval_heat_priming_climate)	Unscrambling sentences with heat vs cold words and worry about global warming[111], not replicated[50]	Right now my process statements are full of heat: sweat, sunburn and hot water.	How much do you personally worry about global warming?	categorization	1, 2

Sources of the experiments:[21,32,40,44,45,57–111].

Replication sources:[49–56].

Supplementary Table 2.

Individual characteristics of the agent profiles. The 139 characteristics served as constructs in the construct-activation analysis (Study 1); the last two columns show how each characteristic enters MVCA. In the analysis, each characteristic was represented by its name and a short description, except seven characteristics (the generational cohort, the five formative memories, and brand authenticity perception), which were represented by their name alone; the 60 mediators proposed by the theories were represented as worded in Supplementary Table 1. "Described to" lists the processes for which the LLM receives the cognitive agent's score, percentile, and the meaning of its level; the LLM receives this description only if published research links the characteristic to that kind of process and the level is clearly high or low (41 characteristics). "Shifts the elaboration of" lists the processes whose process statement the characteristic, without being described to the LLM, makes fuller, briefer, or absent (17 characteristics). The other 81 characteristics are never described to the LLM. The names of 18 characteristics (7 of the described characteristics, 6 of those that shift elaboration, and 5 of the other 81) can also appear, without the cognitive agent's level, among the candidate drives that the LLM receives for the appraisal of goal relevance (Supplementary Note 4). For judgment, the LLM receives none of the characteristics.

#	Individual characteristic	Described to	Shifts the elaboration of
1	Awe proneness	affective response	–
2	Emotion differentiation (granularity)	affective response	–
3	Fading affect bias	–	–
4	Flow proneness	–	–
5	Interoceptive awareness	–	–
6	Life satisfaction	–	–
7	Processing fluency	–	–
8	Misinformation susceptibility	–	–
9	Nostalgia proneness	self-relevance; affective response	–
10	Peak-end evaluation	–	–
11	Savoring	–	–
12	Current arousal	–	every process governed by knowledge activation
13	Current mood (valence)	alternative readings (withheld in every experiment; Supplementary Note 4)	–
14	Trait anxiety	–	goal relevance; alternative readings
15	Trait negative affect	affective response; causal inference	–
16	Trait positive affect	affective response	–
17	Achievement motive	goal relevance	–
18	Affiliation motive	–	–
19	Reliance on anticipated feelings	–	–
20	Autonomy (need satisfaction)	–	–
21	Brand attachment	–	–
22	Competence (need satisfaction)	–	–
23	Fatigue	–	–

#	Individual characteristic	Described to	Shifts the elaboration of
24	Goal commitment	–	–
25	Hunger	goal relevance; comparison of alternatives	–
26	Impression motivation	social inference; comparison of alternatives	–
27	Materialism	goal relevance; social inference	–
28	Mortality-salience susceptibility	–	–
29	Need for cognitive closure	–	categorization; alternative readings
30	Need for cognition	–	affective response; alternative readings; comparison of alternatives
31	Need for uniqueness	social inference; comparison of alternatives	–
32	Need to belong	–	social inference
33	Regret anticipation	comparison of alternatives	–
34	Rejection sensitivity	affective response; social inference	–
35	Relatedness (need satisfaction)	–	–
36	Social comparison orientation	social inference; comparison of alternatives	–
37	Status motive	–	goal relevance; social inference
38	Anchoring susceptibility	–	–
39	Assessment orientation	comparison of alternatives	–
40	Accountability / audience salience	–	–
41	Availability heuristic	–	–
42	Reward sensitivity (BAS)	goal relevance; affective response	–
43	Base-rate neglect	–	–
44	Belief bias	–	–
45	Punishment sensitivity (BIS)	goal relevance; affective response	–
46	Choice-overload sensitivity	–	–
47	Cognitive reflection	–	coherence review
48	Confirmation bias	–	–
49	Construal level	–	encoding
50	Decoy (attraction) effect	–	–
51	Diffusion of responsibility	–	–
52	Distinction bias	–	–

#	Individual characteristic	Described to	Shifts the elaboration of
53	Empathic concern	–	affective response; social inference
54	Hindsight bias	–	–
55	Identifiable-victim effect	–	–
56	Illusory correlation	–	–
57	Implementation-intention disposition	–	–
58	Locomotion orientation	–	–
59	Mental accounting	–	–
60	Trait mindfulness	–	–
61	Normalcy bias	–	–
62	Omission bias	–	–
63	Information avoidance (ostrich effect)	–	–
64	Outcome bias	–	–
65	Overconfidence	–	–
66	Perspective taking	social inference	–
67	Planning fallacy	–	–
68	Prevention focus	–	goal relevance
69	Private self-consciousness	–	–
70	Promotion focus	–	goal relevance
71	Public self-consciousness	–	–
72	Reactance proneness	–	alternative readings
73	Habitual reappraisal	affective response	–
74	Reference dependence	–	–
75	Risk compensation	–	–
76	General schema accessibility	–	–
77	Trait self-control	comparison of alternatives	–
78	Self-handicapping	–	–
79	Self-monitoring	–	social inference
80	Subadditivity	–	–
81	Sunk-cost sensitivity	–	–
82	Habitual expressive suppression	–	–
83	Survivorship bias	–	–
84	Zero-risk bias	–	–
85	Ambivalence tolerance	–	–
86	Trust in automated advice	–	–
87	Brand authenticity perception	–	–
88	Brand loyalty	–	–

#	Individual characteristic	Described to	Shifts the elaboration of
89	Tendency to see brands as persons	–	–
90	Continued influence of corrected information	–	–
91	Defensive attribution	–	–
92	Susceptibility to descriptive norms	–	social inference
93	Dispositional attribution tendency	causal inference; social inference	–
94	Psychological essentialism (groups)	categorization; social inference	–
95	Implicit theory of ability	goal relevance; causal inference	–
96	Implicit theory of personality	causal inference; social inference	–
97	Susceptibility to injunctive norms	social inference; comparison of alternatives	–
98	Belief in a just world	causal inference; social inference	–
99	Internal locus of control	goal relevance; causal inference	–
100	Loss aversion	comparison of alternatives	–
101	Willpower theory (named mind–body dualism in the agent profile and the instructions)	–	–
102	Moral meaning of money	–	–
103	Naive realism / bias blind spot	–	–
104	Persuasion knowledge	–	social inference; alternative readings
105	Price-quality belief	categorization; comparison of alternatives	–
106	Reactive devaluation	–	–
107	Signal sophistication	–	–
108	Belief that consumption signals status	–	–
109	Theory-of-mind capacity	–	–
110	Zero-sum belief	–	–
111	Generational cohort	–	–
112	Individualism-collectivism	goal relevance; comparison of alternatives	–
113	Power distance belief	–	–
114	Cultural tightness	–	–
115	Long-term orientation	comparison of alternatives	–
116	Formative memory: belonging	–	–
117	Formative memory: failure	–	–
118	Formative memory: loss	–	–

#	Individual characteristic	Described to	Shifts the elaboration of
119	Formative memory: rejection	–	–
120	Formative memory: success	–	–
121	Actual self (self-description)	self-relevance	–
122	Hoped-for possible self	self-relevance; goal relevance	–
123	Communal brand-relationship norm	–	–
124	Self-brand connection	self-relevance	–
125	Self-concept clarity	–	–
126	Independent self-construal	–	self-relevance
127	Interdependent self-construal	–	social inference
128	Self-esteem	self-relevance; affective response	–
129	Feared possible self	self-relevance; goal relevance	–
130	Ideal self	self-relevance; affective response	–
131	Salient identities	self-relevance	–
132	Meaning in life	–	–
133	Moral identity	self-relevance; goal relevance	–
134	Possessions as extended self	–	–
135	Ought self	self-relevance; affective response	–
136	Psychological ownership	–	–
137	Role-identity centrality	self-relevance	–
138	Self-perception (inferring attitudes from behavior)	–	–
139	Third-person perception	–	–

Supplementary Table 3.

Tests of individual characteristics in Study 4. Entries are d_z [95% CI] (asterisk, 95% CI excludes zero); positive values are in the predicted direction. The interaction is the effect of the manipulation (treatment vs control condition) at the level with the larger predicted effect minus that effect at the other level. Cue found is the share of cognitive simulations in which the LLM determined that the encoding mentioned the activating cue and supported this answer with a passage quoted verbatim from the encoding, in the treatment and the control condition (range over the two levels of the characteristic). (**A**) Gemma 4 26B, all 50 cognitive agents of the four studies. (**B**) Interaction with each LLM and averaged over the three LLMs, for the 30 cognitive agents simulated with all three LLMs, and cue found with Ling 3.0 Flash.

(**A**) Gemma 4 26B

Test	Individual characteristic	Level at which the theory predicts the larger effect	Activating cue (looked for in the cognitive agent's encoding of the situation)	High level: sentences describing the characteristic	Low level: sentences describing the characteristic	Cue found (treatment; control)	Effect of the manipulation at the high level, d_z [95% CI]	Effect of the manipulation at the low level, d_z [95% CI]	Interaction, d_z [95% CI]
Time versus money × materialism	materialism[112]	low	the scenario asks you to think about spending time	For me, the things I own and what they cost say a lot about who I am; spending money on possessions feels meaningful.	For me, what I own matters little; time with people and experiences are what really count.	100%; 0%	−0.26 [−0.54, +0.03] (n 50)	+0.13 [−0.16, +0.41] (n 50)	+0.27 [−0.02, +0.55] (n 50)
Identity threat × collective self-esteem	collective self-esteem	low	the scenario reports negative findings about {own_group}	I feel good about being a {member}; being a {member} is a source of pride for me.	Being a {member} is not something I feel especially good about.	98–100%; 0%	−0.04 [−0.34, +0.26] (n 44)	+1.21* [+0.91, +1.51] (n 45)	+0.70* [+0.40, +1.01] (n 44)
Effort and indulgence × guilt about luxuries	guilt about luxuries	high	you would have to do something to earn the reward	Spending on luxuries usually makes me feel guilty unless I feel I have earned it.	I rarely feel guilty about treating myself to something nice.	98–100%; 100%	+0.24 [−0.04, +0.53] (n 49)	−0.18 [−0.46, +0.11] (n 50)	+0.28 [−0.01, +0.57] (n 49)
Beer price × belief that higher costs justify higher prices	the belief that higher costs justify higher prices	high	the beer would come from a fancy or upscale place	Fancy places have higher costs, so it is fair for them to charge more for the same thing.	A beer is a beer; where you buy it should not change what it is worth.	100%; 0%	+1.56* [+1.27, +1.84] (n 50)	+1.34* [+1.05, +1.62] (n 50)	+0.63* [+0.34, +0.91] (n 50)
Null: beer price × self-monitoring (other cue)	self-monitoring	high (null test)	someone you know is with you or will see what you choose	I often adjust how I act depending on who I am with and what they expect.	I act the same way no matter who I am with.	100%; 98–100%	+1.34* [+1.06, +1.63] (n 50)	+1.43* [+1.14, +1.71] (n 50)	−0.08 [−0.36, +0.21] (n 50)
Null: beer price × self-monitoring (same cue)	self-monitoring (same cue as the beer-price test)	high (null test)	the beer would come from a fancy or upscale place	I often adjust how I act depending on who I am with and what they expect.	I act the same way no matter who I am with.	100%; 0%	+1.97* [+1.69, +2.25] (n 50)	+1.46* [+1.18, +1.74] (n 50)	+0.01 [−0.28, +0.29] (n 50)
Foot in the door × preference for consistency	preference for consistency[113]	high	you earlier agreed to a smaller request for a similar cause	I like my actions to be consistent with what I have done and said before.	I do not mind acting differently from how I acted before.	98–100%; 0%	+0.62* [+0.34, +0.90] (n 50)	+0.06 [−0.23, +0.34] (n 50)	+0.42* [+0.14, +0.71] (n 50)
Identity salience × importance of family identity	importance of family identity	high	you recalled moments when you and your family grew closer	My family is one of the most important parts of who I am.	My family is not a big part of how I see myself.	64%; 0%	+0.65* [+0.37, +0.93] (n 50)	+0.41* [+0.13, +0.70] (n 50)	+0.15 [−0.13, +0.44] (n 50)
Pooled (random effects), six tests (null tests excluded)									+0.40 [+0.17, +0.62], positive 6/6, significant 3; without the beer-price test, +0.35 [+0.10, +0.61]

(**B**) Interaction with each LLM, 30 cognitive agents, d_z (asterisk: 95% CI excludes zero), and cue found with Ling 3.0 Flash

Test	Gemma 4 26B	Qwen 3.7 Flash	Ling 3.0 Flash	Averaged over LLMs	Cue found, Ling 3.0 Flash (treatment; control)
Time versus money × materialism	+0.20	+0.11	−0.08	+0.22	45–52%; 3–6%
Identity threat × collective self-esteem	+0.64*	+0.64*	+0.40*	+0.97*	89–93%; 0%
Effort and indulgence × guilt about luxuries	+0.30	+0.08	−0.25	+0.05	100%; 100%
Beer price × belief that higher costs justify higher prices	+0.60*	+0.09	+0.29	+0.58*	57–77%; 0%
Null: beer price × self-monitoring (other cue)	−0.23	+0.12	+0.14	+0.10	70–83%; 57–63%
Null: beer price × self-monitoring (same cue)	−0.03	−0.29	+0.23	−0.05	67–73%; 0%
Foot in the door × preference for consistency	+0.53*	+0.50*	+0.22	+0.52*	93–97%; 0%
Identity salience × importance of family identity	+0.13	+0.28	−0.17	+0.15	33–53%; 0%
Pooled (random effects), six tests	+0.39 [+0.16, +0.61], positive 6/6, significant 3	+0.27 [+0.03, +0.51], positive 6/6, significant 2	+0.06 [−0.22, +0.35], positive 3/6, significant 1	+0.40 [+0.05, +0.74], positive 6/6, significant 3	
Agreement across the six tests, Spearman ρ (same sign)	Gemma–Qwen 0.31 (6/6); Gemma–Ling 0.83 (3/6); Qwen–Ling 0.60 (3/6)				

All entries are from the first time each test was conducted; Supplementary Table 10 reports two repetitions of the three tests that did not reach significance in panel A.

Because d_z divides each mean by its standard deviation, the difference between the two effects of the manipulation need not equal the interaction.

In panel A, the last row reports the pooled random-effects estimate of the interaction in the six tests, the numbers of tests with a positive and with a significant interaction, and the pooled estimate without the beer-price test.

The Supplementary Information describes the levels and the activating cues (Supplementary Note 13) and the average over LLMs (Supplementary Note 15).

Placeholders in braces are filled from the agent profile (for example, the cognitive agent's gender).

Supplementary Table 4.

Pooled effects by study and LLM. Pooled d_z [95% CI] from random-effects models with Hartung–Knapp–Sidik–Jonkman CIs; positive values are in the predicted direction; bold, CI excludes zero. Gemma 4 26B, all 50 cognitive agents; Qwen 3.7 Flash and Ling 3.0 Flash, the first 30 of these cognitive agents. In Studies 3 and 4, the estimates are based on the first time each test was conducted; the repetitions with Gemma 4 26B are in Supplementary Table 10.

Study and effect*	Gemma 4 26B	Qwen 3.7 Flash	Ling 3.0 Flash	Averaged over the three LLMs†
Study 1: effect of the manipulation on support for the mediator (60 experiments)	**0.21 [0.08, 0.34]**	**0.24 [0.12, 0.37]**	**0.28 [0.16, 0.41]**	**0.44 [0.25, 0.63]**
Study 1: effect of the manipulation on ratings of the mediator by a separate LLM (60 experiments)	**0.37 [0.17, 0.58]**	**0.52 [0.28, 0.76]**	**0.51 [0.29, 0.72]**	**0.77 [0.45, 1.09]**
Study 2: effect of the induced mediator, rather than its denial, on the response (52 experiments)	**0.86 [0.64, 1.09]**	**0.63 [0.48, 0.79]**	**0.77 [0.56, 0.98]**	**1.17 [0.91, 1.43]**
Study 3: interaction of the situational moderator and the induced mediator (11 tests)	0.23 [−0.03, 0.48]	0.15 [−0.06, 0.36]	0.15 [−0.04, 0.34]	**0.31 [0.04, 0.57]**
Study 4: interaction of the individual characteristic and the manipulation (6 tests)	**0.40 [0.17, 0.62]**	**0.27 [0.03, 0.51]**	0.06 [−0.22, 0.35]	**0.40 [0.05, 0.74]**

[a] Numbers of cognitive simulations with Gemma 4 26B, Qwen 3.7 Flash, and Ling 3.0 Flash, respectively: Study 1, 5,892, 3,546, and 3,545; Study 2, 13,054, 7,860, and 7,858; Study 3, 3,224, 1,944, and 1,928; Study 4, 1,578, 948, and 946.

[b] For each cognitive agent that all three LLMs simulated, the difference (in Studies 3 and 4, the interaction) was averaged over the three LLMs before d_z was computed; because averaging reduces chance variation specific to one LLM, these values can exceed those of every single LLM.

Supplementary Table 5.

Effects of the manipulation on the process statements in each experiment (Study 1). Rows are the 60 experiments, ordered by the effect on support for the mediator averaged over LLMs. Entries are d_z (asterisk, 95% CI excludes zero); positive values are in the predicted direction and, for pleasantness, indicate more pleasant process statements in the treatment than in the control condition. Gemma 4 26B, all 50 cognitive agents (27 to 50 per experiment); Qwen 3.7 Flash, Ling 3.0 Flash, and the average over LLMs, the first 30 (15 to 30 per experiment). (**A**) Natural language inference model (Supplementary Note 5). Support for the mediator is the probability that a process statement implied the mediator statements minus the probability that it contradicted them, averaged over the process statements of the ten processes between encoding and judgment or measured in the process statement of the appraisal process in which, according to the theory, the mediator arises (Supplementary Table 1). In the process statement of that appraisal process, the mediator-minus-denial score is the probability that the process statement implied the mediator statements minus the probability that it implied the denial statements; the column "Beyond other experiments' mediators" reports the treatment-minus-control difference in this score minus the same difference for the statements of the next two experiments in the order of Supplementary Table 1. Pleasantness is the probability that the process statements implied "I feel good and pleasant right now" minus the probability that they implied "I feel bad and unpleasant right now" (Gemma 4 26B). (**B**) Rating of the mediator by the separate LLM (DeepSeek V4.1 Flash, 1 to 7): how strongly the process statements reflected the mediator, alone and beyond a favorable evaluation of the target of the judgment (the rating of the mediator minus the rating of how favorably the process statements evaluated that target).

(**A**) Inference model

Experiment	n (Gemma 4 26B)	Support, Gemma 4 26B	Support, Qwen 3.7 Flash	Support, Ling 3.0 Flash	Support, averaged over LLMs	Support in the appraisal process in which the mediator arises, averaged over LLMs	Mediator minus denial in the appraisal process in which the mediator arises, averaged over LLMs	Beyond other experiments' mediators, averaged over LLMs	Pleasantness, Gemma 4 26B
Nonconforming dress	50	+2.55*	+1.71*	+1.70*	+3.83*	+2.83*	+1.06*	+1.32*	+0.39*
Brand extension	50	+1.68*	+2.33*	+1.12*	+2.98*	+1.71*	+1.84*	+1.67*	+0.22
Less is better	50	+1.46*	+1.62*	+2.19*	+2.79*	+2.44*	−0.22	−0.14	+0.76*
Disease framing	50	+1.23*	+1.27*	+0.92*	+2.24*	+0.83*	+0.44*	+0.39*	+0.47*
Quote attribution	50	+0.98*	+0.57*	+1.51*	+1.88*	+1.41*	+0.40*	+0.36	+1.91*
Physical warmth	50	+0.62*	+0.86*	+0.90*	+1.59*	+0.67*	+0.47*	+0.34	+0.57*
IKEA effect	50	+0.09	+1.35*	+0.88*	+1.26*	+0.60*	+0.41*	+0.44*	−0.16
Side effects and intent	50	+0.83*	+0.22	+0.94*	+1.08*	+0.97*	+0.29	−0.05	−1.85*
Labeling	50	+0.67*	+0.28	+1.29*	+1.02*	+0.29	+0.24	+0.19	+0.10
Pennies a day	47	+0.47*	+1.09*	+0.22	+1.00*	+0.72*	−0.08	−0.18	+0.07
Vertical position and power	50	+0.42*	+0.36	+0.51*	+0.90*	+0.26	+0.74*	+0.54*	−0.16
Emotional attachment	50	+0.44*	+0.38*	+0.64*	+0.81*	+0.46*	+0.06	+0.14	−0.14
Shaken self-view	50	+0.17	+0.68*	+0.34	+0.80*	+0.44*	+0.40*	+0.15	−0.12
Identity salience	50	+0.18	+0.72*	+0.33	+0.74*	+1.15*	+1.07*	+1.01*	+0.03
Mere touch	50	+0.80*	−0.08	+0.60*	+0.72*	+0.28	+0.36	+0.28	+0.55*
Possession and valuation	50	+0.00	+0.30	+0.76*	+0.62*	+0.70*	+0.26	+0.37	+0.02
Endowment effect	50	−0.10	+0.40*	+0.47*	+0.61*	+0.30	+0.26	+0.18	+0.38*
Structure and goal pursuit	47	+0.99*	+0.27	−0.22	+0.59*	+0.30	+0.31	+0.26	−0.16

Experiment	n (Gemma 4 26B)	Support, Gemma 4 26B	Support, Qwen 3.7 Flash	Support, Ling 3.0 Flash	Support, averaged over LLMs	Support in the appraisal process in which the mediator arises, averaged over LLMs	Mediator minus denial in the appraisal process in which the mediator arises, averaged over LLMs	Beyond other experiments' mediators, averaged over LLMs	Pleasantness, Gemma 4 26B
Identity divergence	50	+0.15	+0.44*	+0.44*	+0.51*	+0.17	−0.40*	−0.34	−0.20
Emotion and sharing	50	+0.91*	−0.11	+0.23	+0.51*	+0.54*	+0.20	+0.20	−0.87*
Future self and saving	27	+0.30	+0.35	+0.70*	+0.50	+0.77*	+0.76*	+0.78*	−0.27
Goal progress	50	+0.11	+0.49*	+0.30	+0.46*	+0.43*	+0.14	+0.09	−0.08
Mortality salience	46	+0.08	+0.09	+0.41*	+0.43*	+0.17	+0.21	+0.33	+0.06
Effort and indulgence	50	+0.07	+0.23	+0.47*	+0.40*	+0.17	+0.05	−0.03	−0.21
Incidental anchors	50	+0.10	+0.12	+0.18	+0.34	+0.49*	+0.29	+0.09	+0.01
Anthropomorphized products	50	−0.10	+0.04	+0.33	+0.31	+0.01	+0.06	+0.18	+0.12
Flag priming	46	+0.12	−0.03	+0.27	+0.25	−0.27	+0.11	+0.09	+0.01
Similarity asymmetry	50	+0.22	+0.02	+0.00	+0.23	+0.15	+0.22	−0.22	−0.17
Going green	50	+0.23	+0.36	−0.39*	+0.22	−0.17	+0.06	+0.04	+0.06
Analytic thinking and belief	50	+0.08	−0.02	+0.22	+0.20	−0.00	+0.03	+0.10	+0.11
Forbid vs allow	46	−0.16	+0.36	+0.18	+0.18	+0.51*	−0.04	−0.02	+0.11
Anchoring	50	+0.07	−0.14	+0.41*	+0.17	−0.01	+0.12	+0.19	+0.07
Beer price	50	+0.02	+0.03	+0.21	+0.17	−0.00	+0.18	+0.13	+0.26
Choosing vs rejecting	50	+0.02	+0.12	+0.02	+0.15	−0.11	−0.19	−0.38*	+0.07
Overhead aversion	50	+0.20	−0.03	+0.17	+0.15	+0.04	−0.16	−0.18	+0.13
Time versus money	50	−0.03	+0.12	+0.07	+0.14	−0.08	−0.12	−0.11	−0.32*
Feature fatigue	50	−0.03	+0.15	−0.07	+0.14	+0.40*	−0.06	+0.14	+0.01
Powerlessness	50	+0.07	+0.01	+0.16	+0.09	+0.32	−0.01	−0.07	−0.42*
Decision difficulty	50	−0.27	+0.04	+0.17	+0.09	−0.02	−0.24	−0.15	−1.06*
Norm of reciprocity	46	+0.04	−0.01	−0.01	+0.08	+0.13	+0.26	+0.44*	−0.34*
Door in the face	50	+0.19	+0.08	−0.25	+0.07	+0.12	+0.56*	+0.68*	+0.02
Normative message	50	+0.19	−0.07	−0.16	+0.07	+0.33	+0.06	+0.10	+0.73*
Moral typecasting	50	−0.72*	+0.26	+0.39*	+0.07	+1.03*	+0.63*	+0.24	−0.80*
Tempting fate	50	−0.04	+0.10	+0.06	+0.06	+0.07	−0.23	−0.09	−1.18*
Heat priming	50	−0.25	+0.21	−0.02	+0.03	+0.01	−0.28	−0.16	+0.05
Buying and selling prices	50	+0.06	+0.05	−0.20	+0.01	−0.26	−0.13	−0.22	−0.09
Payment transparency	50	−0.13	+0.01	−0.02	−0.01	+0.02	−0.27	−0.26	−0.38*
Affect reliance	50	+0.02	−0.10	−0.05	−0.04	−0.19	−0.20	−0.31	−0.02
Affect-rich prizes	50	−0.03	−0.21	+0.12	−0.05	+0.22	+0.31	+0.39*	−0.35*

Experiment	n (Gemma 4 26B)	Support, Gemma 4 26B	Support, Qwen 3.7 Flash	Support, Ling 3.0 Flash	Support, averaged over LLMs	Support in the appraisal process in which the mediator arises, averaged over LLMs	Mediator minus denial in the appraisal process in which the mediator arises, averaged over LLMs	Beyond other experiments' mediators, averaged over LLMs	Pleasantness, Gemma 4 26B
Identity threat	45	+0.06	−0.16	+0.10	−0.05	−0.10	−0.10	−0.25	−0.47*
Social exclusion	50	+0.01	+0.10	−0.08	−0.06	−0.16	−0.01	+0.10	−0.88*
Personal force	50	−0.11	+0.11	−0.05	−0.07	+0.00	−0.06	−0.31	+0.06
Verb aspect and intent	50	−0.29*	−0.11	+0.06	−0.13	−0.15	+0.02	−0.51*	−0.12
Foot in the door	50	−0.02	−0.18	+0.01	−0.14	+0.40*	+0.37	+0.41*	−0.17
Weight and importance	50	−0.17	−0.20	+0.03	−0.14	+0.01	+0.18	+0.09	−0.51*
Goal fluency	50	+0.07	−0.29	−0.16	−0.20	−0.00	+0.29	+0.29	+0.05
Response-scale range	50	−0.34*	+0.05	−0.10	−0.28	+0.09	+0.05	−0.01	+0.13
Currency priming	46	−0.14	−0.08	−0.23	−0.30	−0.19	−0.35	−0.28	−0.23
Retrospective gambler's fallacy	50	−0.14	−0.07	−0.39*	−0.30	−0.13	+0.39*	+0.18	+0.05
Consumer label and trust	50	−0.18	−0.12	−0.30	−0.33	−0.17	−0.03	+0.01	+0.11
Pooled (random effects), 60 experiments		+0.21 [+0.08, +0.34], positive 41/60, significant 14; significantly negative 3	+0.24 [+0.12, +0.37], positive 42/60, significant 14; significantly negative 0	+0.28 [+0.16, +0.41], positive 43/60, significant 20; significantly negative 2	+0.44 [+0.25, +0.63], positive 46/60, significant 23; significantly negative 0	+0.31 [+0.17, +0.45], positive 43/60, significant 21; significantly negative 0	+0.16 [+0.07, +0.26], positive 40/60, significant 13; significantly negative 1	+0.13 [+0.03, +0.22], positive 38/60, significant 11; significantly negative 2	−0.07 [−0.19, +0.06], positive 32/60, significant 8; significantly negative 13
Support in the appraisal process in which the mediator arises, each LLM		+0.16 [+0.05, +0.27], positive 38/60, significant 13; significantly negative 2	+0.18 [+0.08, +0.27], positive 40/60, significant 13; significantly negative 0	+0.18 [+0.09, +0.26], positive 42/60, significant 14; significantly negative 0					
Mediator minus denial in the appraisal process in which the mediator arises, each LLM		+0.06 [−0.01, +0.13], positive 31/60, significant 10; significantly negative 1	+0.07 [+0.01, +0.14], positive 37/60, significant 3; significantly negative 2	+0.15 [+0.09, +0.22], positive 44/60, significant 15; significantly negative 0					
Beyond other experiments' mediators, each LLM		+0.06 [−0.01, +0.13], positive 31/60, significant 7; significantly negative 2	+0.06 [−0.00, +0.12], positive 41/60, significant 4; significantly negative 2	+0.13 [+0.06, +0.20], positive 40/60, significant 9; significantly negative 0					

Experiment	n (Gemma 4 26B)	Support, Gemma 4 26B	Support, Qwen 3.7 Flash	Support, Ling 3.0 Flash	Support, averaged over LLMs	Support in the appraisal process in which the mediator arises, averaged over LLMs	Mediator minus denial in the appraisal process in which the mediator arises, averaged over LLMs	Beyond other experiments' mediators, averaged over LLMs	Pleasantness, Gemma 4 26B
Agreement between LLMs, support (Spearman ρ across the 60 experiments, the 30 cognitive agents simulated with all three LLMs): Gemma–Qwen 0.35; Gemma–Ling 0.45; Qwen–Ling 0.58									

(**B**) Rating of the mediator by the separate LLM

Experiment	Rating, Gemma 4 26B	Rating, Qwen 3.7 Flash	Rating, Ling 3.0 Flash	Rating, averaged over LLMs	Beyond a favorable evaluation, Gemma 4 26B	Beyond a favorable evaluation, Qwen 3.7 Flash	Beyond a favorable evaluation, Ling 3.0 Flash	Beyond a favorable evaluation, averaged over LLMs
Nonconforming dress	+1.46*	+1.73*	+1.20*	+2.42*	+1.54*	+1.35*	+1.16*	+2.26*
Brand extension	+1.83*	+2.58*	+2.15*	+3.58*	+1.19*	+1.64*	+1.59*	+2.74*
Less is better	+1.82*	+2.27*	+2.05*	+3.42*	+0.97*	+1.69*	+2.12*	+2.60*
Disease framing	+2.81*	+2.22*	+1.31*	+3.58*	+1.26*	+1.37*	+0.93*	+1.79*
Quote attribution	+1.28*	+1.46*	+1.53*	+2.47*	−0.40*	+0.56*	+0.74*	+0.36
Physical warmth	+0.26	+0.34	+0.04	+0.37	+0.20	+0.38*	+0.09	+0.36
IKEA effect	−0.07	+1.54*	+1.46*	+1.57*	+0.05	+0.34	+0.93*	+0.71*
Side effects and intent	+2.23*	+1.31*	+1.55*	+2.78*	+2.26*	+1.09*	+1.82*	+2.49*
Labeling	−0.32*	−0.52*	+0.58*	−0.13	−0.09	−0.28	+0.52*	+0.03
Pennies a day	+0.67*	+1.20*	+0.51*	+1.57*	+0.63*	+0.70*	+0.49*	+1.13*
Vertical position and power	+2.19*	+1.39*	−0.06	+1.59*	+1.74*	+1.30*	+0.05	+1.74*
Emotional attachment	+0.46*	+0.85*	+1.41*	+1.72*	+0.62*	+0.64*	+1.38*	+1.67*
Shaken self-view	+0.15	+1.14*	+0.48*	+0.98*	+0.00	+1.05*	+0.07	+0.80*
Identity salience	+1.22*	+1.67*	+0.65*	+2.21*	+0.66*	+0.80*	+0.58*	+1.25*
Mere touch	+0.24	+0.59*	+1.07*	+1.11*	−0.36*	+0.55*	+0.94*	+0.80*
Possession and valuation	+0.13	+0.12	+0.48*	+0.29	−0.08	+0.35	+0.12	+0.22
Endowment effect	−0.25	+0.06	+0.73*	+0.37	−0.38*	−0.05	+0.19	−0.19
Structure and goal pursuit	+0.69*	+0.18	+0.00	+0.42*	+0.73*	+0.03	+0.24	+0.57*
Identity divergence	+0.74*	+1.35*	+3.75*	+2.32*	+0.42*	+0.93*	+2.93*	+1.76*
Emotion and sharing	+0.75*	+0.22	+0.29	+0.69*	+0.75*	+0.18	+0.26	+0.64*
Future self and saving	−0.41*	−0.46	+0.68*	−0.32	−0.31	−0.31	+0.51	−0.25

Experiment	Rating, Gemma 4 26B	Rating, Qwen 3.7 Flash	Rating, Ling 3.0 Flash	Rating, averaged over LLMs	Beyond a favorable evaluation, Gemma 4 26B	Beyond a favorable evaluation, Qwen 3.7 Flash	Beyond a favorable evaluation, Ling 3.0 Flash	Beyond a favorable evaluation, averaged over LLMs
Goal progress	−0.42*	+0.61*	−0.30	−0.10	−0.27	+0.41*	−0.40*	−0.13
Mortality salience	−0.20	+0.36	+0.13	+0.34	−0.15	+0.48*	+0.17	+0.34
Effort and indulgence	+0.06	−0.10	−0.19	−0.19	+0.13	−0.04	−0.18	−0.12
Incidental anchors	+0.00	+0.00	+0.00	+0.00	−0.12	+0.02	+0.21	+0.00
Anthropomorphized products	−0.49*	+0.27	+0.42*	+0.23	−0.59*	−0.15	+0.18	−0.25
Flag priming	+0.23	+0.17	+0.00	+0.26	+0.23	+0.17	+0.00	+0.26
Similarity asymmetry	+0.00	+0.00	+0.00	+0.00	+0.00	+0.00	+0.00	+0.00
Going green	+0.74*	+3.31*	+0.15	+2.04*	+0.67*	+1.57*	+0.26	+1.70*
Analytic thinking and belief	−0.15	+0.00	+0.08	+0.03	−0.03	−0.08	+0.08	+0.05
Forbid vs allow	+0.20	+0.95*	+0.78*	+1.14*	−0.21	−0.45*	−0.29	−0.48*
Anchoring	−0.14	−0.18	+0.00	−0.26	+0.00	−0.05	+0.00	−0.13
Beer price	+1.11*	+0.53*	+0.95*	+1.31*	+1.01*	+0.53*	+0.59*	+1.18*
Choosing vs rejecting	+0.18	+0.38*	+0.21	+0.38*	−0.11	−0.03	+0.21	+0.02
Overhead aversion	+1.16*	+0.98*	+0.53*	+1.25*	+0.97*	+0.72*	+0.58*	+1.05*
Time versus money	+0.11	−0.09	+0.32	+0.11	+0.02	−0.26	+0.42*	+0.05
Feature fatigue	+0.02	+0.52*	+1.47*	+0.96*	+0.66*	+1.11*	+1.04*	+1.49*
Powerlessness	+0.26	+0.60*	+0.30	+0.68*	+0.62*	+0.45*	+0.48*	+1.17*
Decision difficulty	−0.62*	−1.26*	+0.20	−0.59*	−0.58*	−0.58*	+0.26	−0.37
Norm of reciprocity	−0.15	−0.39*	−0.19	−0.50*	−0.19	−0.41*	−0.19	−0.52*
Door in the face	−0.28	−0.10	−0.05	−0.19	−0.04	−0.03	+0.00	+0.06
Normative message	−0.78*	−1.78*	−0.49*	−1.59*	−0.95*	−1.21*	−0.49*	−1.53*
Moral typecasting	+3.24*	+3.25*	+4.92*	+6.82*	+3.04*	+4.11*	+3.78*	+6.68*
Tempting fate	+0.87*	+1.14*	+1.31*	+1.77*	+0.92*	+0.94*	+1.31*	+1.74*
Heat priming	+0.00	+0.18	+0.00	+0.18	−0.20	−0.31	+0.26	−0.36
Buying and selling prices	−0.11	+0.69*	+0.00	+0.33	−0.15	+0.22	+0.04	+0.03
Payment transparency	−0.31*	−0.60*	−0.35	−0.86*	+0.02	−0.60*	−0.24	−0.62*
Affect reliance	+0.03	+0.14	+0.70*	+0.45*	−0.16	+0.18	+0.43*	+0.12
Affect-rich prizes	+0.28	+0.67*	+0.09	+0.58*	−0.03	+0.39*	+0.18	+0.53*
Identity threat	−0.02	−0.38	−0.38	−0.54*	−0.72*	−0.64*	−0.49*	−1.22*
Social exclusion	−0.05	−0.10	+0.61*	+0.38*	+0.19	−0.05	+0.63*	+0.48*
Personal force	+0.26	+0.42*	+0.32	+0.69*	+0.23	+0.75*	+0.33	+0.92*
Verb aspect and intent	+0.00	−0.19	−0.03	−0.10	+0.03	−0.24	+0.35	+0.04
Foot in the door	+0.14	+0.79*	+0.50*	+0.93*	−0.60*	+0.23	−0.49*	−0.35
Weight and importance	−0.06	−0.12	+0.21	+0.08	+0.12	+0.03	+0.17	+0.16

Experiment	Rating, Gemma 4 26B	Rating, Qwen 3.7 Flash	Rating, Ling 3.0 Flash	Rating, averaged over LLMs	Beyond a favorable evaluation, Gemma 4 26B	Beyond a favorable evaluation, Qwen 3.7 Flash	Beyond a favorable evaluation, Ling 3.0 Flash	Beyond a favorable evaluation, averaged over LLMs
Goal fluency	+0.15	+0.22	−0.29	+0.03	+0.05	+0.13	−0.36	−0.15
Response-scale range	−0.14	+0.31	−0.18	+0.27	−0.02	+0.14	−0.28	−0.09
Currency priming	+0.04	−0.36	+0.17	−0.17	+0.11	−0.21	−0.05	−0.05
Retrospective gambler's fallacy	+0.81*	+0.59*	+0.18	+1.19*	+0.56*	+0.89*	−0.38*	+0.82*
Consumer label and trust	−0.20	+0.20	+0.02	−0.02	+0.13	+0.00	−0.19	+0.03
Pooled (random effects), 60 experiments	+0.37 [+0.17, +0.58], positive 36/60, significant 19; significantly negative 7	+0.52 [+0.28, +0.76], positive 42/60, significant 29; significantly negative 5	+0.51 [+0.29, +0.72], positive 42/60, significant 27; significantly negative 1	+0.77 [+0.45, +1.09], positive 44/60, significant 31; significantly negative 5	+0.25 [+0.07, +0.43], positive 33/60, significant 20; significantly negative 8	+0.34 [+0.15, +0.53], positive 38/60, significant 26; significantly negative 6	+0.38 [+0.20, +0.57], positive 43/60, significant 22; significantly negative 5	+0.54 [+0.26, +0.82], positive 41/60, significant 26; significantly negative 5
Agreement between support and rating (Spearman ρ across the 60 experiments)	Gemma 0.39	Qwen 0.54	Ling 0.51	averaged 0.46				

In both panels, the last rows report pooled random-effects estimates with Hartung–Knapp–Sidik–Jonkman 95% CIs and the agreement between measures and LLMs.

Supplementary Table 6.

Activation of constructs in the process statements (Study 1). Results for each LLM and averaged over LLMs. Activation is the similarity in meaning between a construct's description and the process statements of the ten processes between encoding and judgment, after the part of each process statement's meaning that it shares with the scenario is removed; closeness is the similarity in meaning between a construct and the experiment's mediator, and ρ is their Spearman correlation, over the 69 characteristics of the agent profile that the LLM never received in any form (the main analysis) or over all 139 characteristics. For comparison, ρ is also computed with the activations in the process statements of each of the other 59 experiments and averaged over them. The rank of the experiment's own mediator is its position among the 60 mediators by activation in the experiment's process statements, after the activations are standardized within each experiment across the 60 mediators and then within each mediator across the 60 experiments. Level is the activation of the experiment's own mediator minus the mean activation of the other 59 mediators. The effect of the manipulation is the treatment-minus-control difference in the activation of the own mediator, in the process statements of all ten processes or in the process statement of the appraisal process in which, according to the theory, the mediator arises (Supplementary Table 1), and this difference minus the same difference for the other 59 mediators; level and effects are reported as d_z (asterisk, 95% CI excludes zero). (**A**) Summary over the 60 experiments; numbers of cognitive agents in parentheses; pooled random-effects estimates with Hartung–Knapp–Sidik–Jonkman 95% CIs. (**B**) Per experiment, averaged over LLMs; rows in the order of Supplementary Table 1.

(**A**) Summary

Measure	Gemma 4 26B (50)	Qwen 3.7 Flash (30)	Ling 3.0 Flash (30)	Averaged over LLMs (30)
Closeness–activation ρ, 69 characteristics: mean (positive of 60)	0.32 (57)	0.32 (55)	0.29 (51)	0.32 (54)
Same with other experiments' process statements, 69 characteristics: mean	0.17	0.19	0.20	0.19
Own process statements higher, 69 characteristics (of 60); Wilcoxon signed-rank test	53; P < 0.001	47; P < 0.001	44; P < 0.001	52; P < 0.001
Closeness–activation ρ, all 139 characteristics: mean (positive of 60)	0.30 (55)	0.31 (58)	0.27 (54)	0.30 (56)
Same with other experiments' process statements, all 139 characteristics: mean	0.15	0.17	0.18	0.17
Own process statements higher, all 139 characteristics (of 60); Wilcoxon signed-rank test	52; P < 0.001	52; P < 0.001	46; P < 0.001	53; P < 0.001
Own mediator ranked first / in the first three among 60 (median rank)	27 / 32 (2)	24 / 33 (3)	20 / 31 (3)	24 / 38 (2)
Own mediator minus other mediators, level (d_z)	+3.28 [+2.47, +4.09], positive 53/60, significant 53; significantly negative 7	+2.77 [+2.05, +3.49], positive 52/60, significant 49; significantly negative 6	+1.76 [+1.13, +2.38], positive 50/60, significant 46; significantly negative 7	+3.91 [+2.93, +4.88], positive 54/60, significant 52; significantly negative 5
Treatment minus control, own mediator, all ten process statements (d_z)	+0.32 [+0.14, +0.50], positive 34/60, significant 25; significantly negative 7	+0.35 [+0.16, +0.54], positive 43/60, significant 20; significantly negative 6	+0.09 [−0.04, +0.22], positive 32/60, significant 15; significantly negative 7	+0.37 [+0.15, +0.60], positive 36/60, significant 22; significantly negative 5
Treatment minus control, own mediator, process statement of the appraisal process in which the mediator arises (d_z)	+0.18 [+0.06, +0.30], positive 40/60, significant 16; significantly negative 5	+0.23 [+0.11, +0.35], positive 42/60, significant 20; significantly negative 2	+0.04 [−0.06, +0.13], positive 32/60, significant 9; significantly negative 4	+0.22 [+0.07, +0.38], positive 37/60, significant 17; significantly negative 6
Treatment minus control, own mediator minus other mediators (d_z)	+0.30 [+0.12, +0.48], positive 35/60, significant 25; significantly negative 8	+0.35 [+0.15, +0.54], positive 37/60, significant 25; significantly negative 6	+0.11 [−0.02, +0.24], positive 34/60, significant 17; significantly negative 9	+0.37 [+0.14, +0.59], positive 35/60, significant 21; significantly negative 4

(**B**) Per experiment, averaged over LLMs

Experiment	ρ, own process statements, 69 characteristics	ρ, other experiments' process statements, 69 characteristics	ρ, own process statements, 139 characteristics	ρ, other experiments' process statements, 139 characteristics	Rank of own mediator among 60	Treatment minus control, own mediator (d_z)	Treatment minus control, own minus other mediators (d_z)
Brand extension	+0.51	+0.47	+0.45	+0.39	1	−0.60*	−0.34
Anthropomorphized products	+0.55	+0.31	+0.44	+0.24	5	+0.45*	+0.52*
Consumer label and trust	+0.35	+0.30	+0.34	+0.29	2	−0.02	−0.10
Nonconforming dress	+0.41	+0.35	+0.39	+0.36	1	−0.96*	−1.14*
Identity divergence	+0.47	+0.35	+0.48	+0.33	1	+0.72*	+1.73*
Emotion and sharing	−0.08	+0.30	−0.15	+0.21	38	+0.22	+0.19
Labeling	+0.64	+0.50	+0.64	+0.50	1	+2.71*	+2.68*
Buying and selling prices	+0.21	−0.09	+0.25	−0.03	2	+0.21	−0.35
Flag priming	+0.46	+0.21	+0.44	+0.21	2	+0.19	+0.30
Currency priming	+0.59	+0.25	+0.54	+0.23	1	+0.03	−0.31
Door in the face	−0.02	−0.09	−0.02	−0.06	1	+0.40*	+0.94*
Incidental anchors	−0.07	−0.16	+0.02	−0.13	13	+0.00	+0.06
Goal progress	+0.52	+0.40	+0.48	+0.35	1	−0.07	−0.19
Foot in the door	+0.41	+0.41	+0.32	+0.41	47	−1.14*	−1.03*
Shaken self-view	+0.50	+0.48	+0.43	+0.43	5	+0.02	+0.12
Analytic thinking and belief	+0.48	+0.26	+0.29	+0.16	10	−0.17	−0.25
Vertical position and power	+0.04	+0.04	+0.10	+0.09	1	−0.21	−0.19
Overhead aversion	+0.22	−0.02	+0.29	+0.05	2	−0.29	−0.18
Pennies a day	+0.19	+0.18	+0.12	+0.10	10	+0.08	+0.18
Moral typecasting	+0.02	−0.02	+0.17	−0.05	3	+2.24*	+2.28*
Mortality salience	+0.70	+0.45	+0.66	+0.43	1	−0.20	−0.21
Personal force	+0.15	+0.28	+0.16	+0.16	8	+0.64*	+0.65*
Going green	+0.44	+0.43	+0.40	+0.39	2	−0.17	−0.15
Verb aspect and intent	+0.47	+0.26	+0.33	+0.21	9	−0.36	−0.18
Future self and saving	+0.44	+0.35	+0.56	+0.39	1	+1.72*	+1.68*
Less is better	+0.27	+0.00	+0.26	−0.01	1	+0.13	+0.02
Norm of reciprocity	+0.30	+0.20	+0.24	+0.17	60	+2.28*	+2.75*
Anchoring	+0.04	−0.28	+0.09	−0.23	2	+0.49*	+0.54*
Weight and importance	−0.01	−0.12	−0.14	−0.13	13	−0.19	−0.12
Endowment effect	−0.14	−0.20	−0.11	−0.22	4	−0.10	−0.56*
Structure and goal pursuit	+0.31	+0.08	+0.35	+0.09	1	−0.04	−0.11
Effort and indulgence	+0.33	+0.32	+0.41	+0.26	3	+0.14	+0.21
Side effects and intent	+0.18	−0.20	+0.15	−0.17	2	−0.06	+0.23
Goal fluency	+0.36	+0.36	+0.29	+0.26	17	−0.56*	−0.31

Experiment	ρ, own process statements, 69 characteristics	ρ, other experiments' process statements, 69 characteristics	ρ, own process statements, 139 characteristics	ρ, other experiments' process statements, 139 characteristics	Rank of own mediator among 60	Treatment minus control, own mediator (d_z)	Treatment minus control, own minus other mediators (d_z)
Quote attribution	+0.31	+0.17	+0.30	+0.14	1	+1.16*	+0.86*
Social exclusion	+0.28	+0.21	+0.29	+0.27	40	+2.23*	+1.72*
Time versus money	+0.49	+0.30	+0.47	+0.32	17	−0.20	−0.35
IKEA effect	+0.66	+0.55	+0.54	+0.44	1	+3.15*	+2.69*
Retrospective gambler's fallacy	+0.09	−0.24	+0.12	−0.22	2	+1.08*	+1.66*
Mere touch	+0.42	+0.47	+0.42	+0.45	42	+0.27	+0.25
Affect reliance	+0.53	+0.54	+0.40	+0.45	8	+0.39*	+0.35
Possession and valuation	+0.49	+0.48	+0.49	+0.45	7	+0.80*	+0.85*
Identity salience	+0.56	+0.42	+0.45	+0.41	27	+0.57*	+0.64*
Tempting fate	+0.50	+0.31	+0.40	+0.20	1	+1.40*	+1.49*
Affect-rich prizes	+0.59	+0.25	+0.52	+0.19	1	+0.61*	+0.61*
Powerlessness	+0.16	+0.22	+0.03	+0.12	49	+0.45*	+0.20
Forbid vs allow	+0.27	+0.01	+0.06	−0.11	1	+0.23	+0.31
Normative message	+0.04	−0.14	+0.03	−0.12	5	+2.53*	+0.88*
Response-scale range	+0.26	+0.08	+0.18	+0.03	2	−0.49*	−0.46*
Decision difficulty	+0.06	−0.29	+0.19	−0.18	1	+1.68*	+2.32*
Choosing vs rejecting	+0.22	+0.19	+0.26	+0.14	1	−0.16	−0.09
Emotional attachment	+0.53	+0.51	+0.50	+0.47	2	+1.21*	+1.25*
Payment transparency	+0.21	+0.20	+0.19	+0.07	33	−0.13	−0.26
Beer price	−0.20	−0.37	+0.07	−0.19	1	−0.12	−0.23
Feature fatigue	+0.40	+0.19	+0.41	+0.14	1	+0.14	+0.09
Similarity asymmetry	+0.18	+0.17	+0.26	+0.20	1	+0.12	+0.22
Disease framing	+0.46	−0.11	+0.42	−0.06	1	−0.07	−0.24
Identity threat	+0.59	+0.39	+0.58	+0.39	2	−0.10	−0.36
Physical warmth	+0.65	+0.52	+0.64	+0.47	1	+0.37	+0.48*
Heat priming	+0.35	+0.21	+0.38	+0.16	2	−0.18	−0.24

Supplementary Table 7.

Effects of induced process statements on the response in each experiment (Study 2). Process statements were induced, in the treatment condition, in the appraisal process in which, according to the theory, the mediator arises. The 52 experiments of Study 2 are the 28 experiments whose effects large preregistered replications re-examined and 24 of the other 32 experiments. Entries are d_z for the difference in the response between two induced process statements (asterisk, 95% CI excludes zero); positive values are in the direction that the theory predicts for the mediator. A pleasant mood was induced in five of these 24 experiments; the Supplementary Information describes the comparison process statements (Supplementary Note 8) and the average over LLMs (Supplementary Note 15). (**A**) Gemma 4 26B, all 50 cognitive agents of the four studies; experiments in the order of the effect of the mediator vs its denial. (**B**) Effect of the mediator vs its denial in each experiment and pooled effects of the other comparisons, with each LLM and averaged over Gemma 4 26B, Qwen 3.7 Flash, and Ling 3.0 Flash, for the 30 cognitive agents simulated with all three LLMs; the Gemma 4 26B value for the comparison with the equally favorable process statement is based on all 50 cognitive agents.

(**A**) Gemma 4 26B, 50 cognitive agents

Experiment	n	Mediator vs denial	Mediator vs pleasant mood	Unrelated vs neutral process statement	Mediator vs equally favorable process statement	Mediator vs neutral process statement
Overhead aversion	50	+3.43*	–	−0.12	+0.94*	+1.33*
Brand extension	50	+3.40*	+0.76*	−0.06	−0.38*	+1.62*
Disease framing	50	+2.96*	–	−0.06	+0.08	+0.90*
Currency priming	46	+2.60*	–	−0.02	+1.67*	+2.05*
Mere touch	50	+1.94*	–	−0.01	+0.39*	+0.89*
Consumer label and trust	50	+1.80*	–	−0.06	+0.05	+0.67*
Feature fatigue	50	+1.65*	–	−0.19	−0.43*	+0.88*
Labeling	50	+1.54*	+0.14	+0.10	−0.12	+0.23
Norm of reciprocity	46	+1.46*	–	−0.03	−0.56*	+1.78*
Emotional attachment	50	+1.44*	–	+0.08	+0.88*	+0.86*
Affect-rich prizes	49	+1.44*	–	−0.14	−0.23	+0.65*
Vertical position and power	50	+1.33*	–	−0.15	+0.85*	+0.63*
Door in the face	50	+1.32*	–	−0.13	+0.56*	+1.02*
Payment transparency	50	+1.30*	–	+0.07	+0.71*	+1.05*
Anchoring	49	+1.26*	–	−0.39*	+0.95*	+0.50*
Nonconforming dress	50	+1.10*	–	−0.24	+0.34*	+0.39*
Similarity asymmetry	50	+1.08*	–	+0.07	+0.24	+0.58*
Beer price	50	+1.02*	+0.73*	+0.27	+0.53*	+0.68*
Anthropomorphized products	50	+0.99*	–	+0.19	−1.65*	−0.05
Forbid vs allow	46	+0.96*	–	−0.47*	−0.11	−0.14
Retrospective gambler's fallacy	50	+0.94*	–	−0.07	+1.31*	+1.25*
Time versus money	50	+0.83*	–	+0.07	−0.50*	+0.41*
Side effects and intent	50	+0.83*	–	+0.04	+0.13	0.00
Weight and importance	50	+0.72*	–	−0.20	+0.08	−0.14
Physical warmth	50	+0.70*	–	−0.09	+0.53*	+0.18
Personal force	50	+0.69*	–	0.00	+0.53*	+0.84*

Experiment	n	Mediator vs denial	Mediator vs pleasant mood	Unrelated vs neutral process statement	Mediator vs equally favorable process statement	Mediator vs neutral process statement
Incidental anchors	50	+0.66*	–	+0.33*	+0.75*	+0.84*
Less is better	50	+0.65*	–	0.00	−0.24	−0.09
IKEA effect	50	+0.63*	+0.24	+0.42*	−0.53*	+0.31*
Flag priming	46	+0.61*	–	−0.15	0.00	+0.10
Response-scale range	50	+0.51*	–	+0.20	−0.93*	+0.48*
Pennies a day	47	+0.50*	–	+0.14	−0.42*	+0.45*
Foot in the door	50	+0.49*	+0.19	+0.05	−0.10	+0.38*
Tempting fate	50	+0.46*	–	+0.10	−0.08	−0.04
Heat priming	50	+0.42*	–	+0.11	+0.43*	+0.47*
Emotion and sharing	50	+0.39*	–	+0.14	+0.14	+0.20
Shaken self-view	49	+0.38*	–	+0.28	+0.26	+0.53*
Verb aspect and intent	50	+0.37*	–	+0.04	+0.08	+0.31*
Moral typecasting	50	+0.35*	–	+0.16	+0.86*	+0.60*
Buying and selling prices	50	+0.28*	–	+0.12	+0.18	+0.14
Future self and saving	27	+0.28	–	−0.58*	+0.42*	+0.09
Structure and goal pursuit	47	+0.28	–	+0.16	−0.01	+0.29
Quote attribution	50	+0.26	–	−0.16	−0.25	+0.11
Effort and indulgence	45	+0.26	–	+0.01	0.00	−0.04
Possession and valuation	50	+0.22	–	+0.13	+0.17	+0.55*
Choosing vs rejecting	50	+0.18	–	−0.08	−0.77*	−0.07
Goal progress	50	+0.12	–	−0.10	−0.11	−0.11
Affect reliance	50	+0.11	–	+0.01	−1.15*	−0.24
Endowment effect	50	+0.07	–	+0.04	−0.22	−0.03
Analytic thinking and belief	50	+0.05	–	+0.04	+0.09	+0.06
Goal fluency	50	+0.03	–	+0.23	−1.26*	+0.06
Mortality salience	46	−0.87*	–	−0.01	−2.40*	−0.86*
Pooled (random effects)		+0.86 [+0.64, +1.09], positive 51/52, significant 40	+0.40 [+0.03, +0.78], positive 5/5	+0.01 [−0.04, +0.06]	+0.04 [−0.16, +0.23]	+0.44 [+0.29, +0.59]

(**B**) Induced process statements with each LLM and averaged over LLMs, 30 cognitive agents

Experiment	Gemma 4 26B	Qwen 3.7 Flash	Ling 3.0 Flash	Averaged over LLMs
Brand extension	+3.06*	+2.59*	+2.93*	+4.92*
Consumer label and trust	+1.93*	+1.60*	+3.82*	+3.50*
Affect-rich prizes	+1.39*	+1.29*	+1.89*	+3.47*
Retrospective gambler's fallacy	+1.12*	+2.06*	+2.71*	+3.05*
Currency priming	+2.65*	+1.08*	+1.65*	+2.34*

Experiment	Gemma 4 26B	Qwen 3.7 Flash	Ling 3.0 Flash	Averaged over LLMs
Labeling	+1.66*	+1.40*	+1.23*	+2.22*
Similarity asymmetry	+0.91*	+1.59*	+0.46*	+2.18*
Disease framing	+3.05*	+0.67*	+1.19*	+2.07*
Nonconforming dress	+0.99*	+1.47*	+0.51*	+2.06*
Norm of reciprocity	+1.58*	+1.00*	+1.40*	+2.05*
Feature fatigue	+1.59*	+1.34*	+1.60*	+2.00*
Overhead aversion	+2.97*	+0.57*	+1.28*	+1.79*
Mere touch	+2.01*	+0.56*	+1.55*	+1.78*
Anthropomorphized products	+0.99*	+0.91*	+1.44*	+1.75*
Vertical position and power	+1.52*	+0.88*	+0.87*	+1.67*
Emotion and sharing	+0.32	+0.50*	+2.39*	+1.54*
Emotional attachment	+1.35*	+0.63*	+0.94*	+1.52*
Physical warmth	+0.78*	+1.17*	+0.68*	+1.49*
Forbid vs allow	+0.84*	+1.44*	+0.44*	+1.40*
Door in the face	+1.51*	+0.89*	+0.18	+1.39*
Side effects and intent	+0.67*	+0.49	+1.06*	+1.36*
Beer price	+1.12*	+1.19*	+0.62*	+1.33*
Incidental anchors	+0.76*	+0.79*	+0.78*	+1.21*
IKEA effect	+0.76*	+0.38*	+1.13*	+1.19*
Weight and importance	+0.71*	+0.68*	+0.87*	+1.15*
Anchoring	+1.07*	−0.14	+0.73*	+1.14*
Heat priming	+0.70*	+0.73*	+0.22	+1.13*
Foot in the door	+0.58*	+0.58*	+0.43*	+1.06*
Payment transparency	+1.35*	+0.44*	+0.54*	+1.02*
Quote attribution	+0.32	+0.77*	+0.76*	+0.99*
Personal force	+0.60*	+0.67*	+0.50*	+0.83*
Structure and goal pursuit	+0.21	+0.52*	+0.79*	+0.74*
Flag priming	+0.81*	+0.15	+0.86*	+0.72*
Less is better	+0.83*	+0.21	+0.64*	+0.66*
Time versus money	+0.61*	+0.49*	+0.48*	+0.65*
Affect reliance	−0.07	+0.45*	+0.64*	+0.58*
Pennies a day	+0.54*	+0.32	+0.35	+0.58*
Shaken self-view	+0.24	+0.36	+0.45*	+0.56*
Moral typecasting	+0.17	+0.37	+0.37	+0.52*
Choosing vs rejecting	+0.36	+0.29	+0.22	+0.51*
Goal fluency	+0.07	+0.10	+0.65*	+0.48*
Response-scale range	+0.60*	+0.17	+0.15	+0.47*

Experiment	Gemma 4 26B	Qwen 3.7 Flash	Ling 3.0 Flash	Averaged over LLMs
Tempting fate	+0.46*	+0.15	+0.17	+0.43*
Endowment effect	+0.17	+0.29	+0.29	+0.41*
Future self and saving	−0.03	+0.19	+0.23	+0.19
Buying and selling prices	+0.14	+0.26	−0.09	+0.17
Possession and valuation	+0.26	−0.39*	+0.22	+0.16
Analytic thinking and belief	+0.05	−0.02	+0.06	+0.03
Goal progress	+0.04	+0.16	−0.37	0.00
Verb aspect and intent	+0.23	+0.04	−0.25	−0.01
Effort and indulgence	+0.02	−0.03	−0.05	−0.03
Mortality salience	−0.97*	+0.27	+0.31	−0.18
Pooled (random effects), mediator vs denial	+0.83 [+0.61, +1.06], positive 49/52, significant 35	+0.63 [+0.48, +0.79], positive 48/52, significant 32	+0.77 [+0.56, +0.98], positive 48/52, significant 36	+1.17 [+0.91, +1.43], positive 48/52, significant 44
Pooled, unrelated vs neutral process statement	−0.02 [−0.07, +0.04]	−0.05 [−0.10, +0.00]	−0.02 [−0.08, +0.03]	−0.05 [−0.11, −0.002]
Pooled, mediator vs pleasant mood (5 experiments)	+0.42 [+0.12, +0.72], positive 5/5, significant 3	+0.50 [+0.15, +0.85], positive 5/5, significant 4	+0.04 [−0.25, +0.33], positive 3/5, significant 0	+0.50 [+0.23, +0.78], positive 5/5, significant 3
Pooled, mediator vs equally favorable process statement (52 experiments)	+0.04 [−0.16, +0.23] (50 cognitive agents)	+0.05 [−0.07, +0.18]	+0.08 [−0.07, +0.24]	+0.10 [−0.12, +0.32]
Agreement between LLMs across the 52 experiments (mediator vs denial)	Gemma–Qwen $\rho = 0.68$, same sign 45/52; Gemma–Ling $\rho = 0.69$, same sign 45/52; Qwen–Ling $\rho = 0.61$, same sign 46/52			

Supplementary Table 8.

Results in the experiments that large preregistered replications re-examined. The 28 experiments whose original effects large preregistered replications re-examined; in 14 the original effect replicated, and in 14 it did not. The replication column identifies the replication study. The analysis followed the plan deposited after the cognitive simulations in the first 10 of these experiments and before those in the other 18. Entries are d_z averaged over the three LLMs: each cognitive agent's difference was averaged over Gemma 4 26B, Qwen 3.7 Flash, and Ling 3.0 Flash before d_z was computed (30 cognitive agents simulated with all three LLMs); asterisk, 95% CI excludes zero. Response: treatment minus control condition (Study 1 design), signed so that positive values are in the published direction; the cognitive agents showed the original effect when the 95% CI lay above zero, and they agreed with the verdict of the replication when they showed the effect where it replicated or did not show it where it did not. Construal (Study 1 design; the first step of the mechanism, the manipulation's effect on the mediator): treatment minus control in support for the mediator (natural language inference, NLI, over the process statements of the ten processes between encoding and judgment) and in the rating of the mediator by the separate LLM; construal was supported when either measure increased significantly. Reliance (Study 2 design; the second step of the mechanism, the mediator's effect on the response): induced mediator vs its denial; reliance was supported when the 95% CI lay above zero. Experiments within each group are in the order of the response effect. Pooled rows: random-effects estimates with 95% CI within each group, and the number of experiments that met each criterion. Difference between groups: $z = (\hat{d}_{\text{replicated}} - \hat{d}_{\text{not replicated}})/\sqrt{(SE^2_{\text{replicated}} + SE^2_{\text{not replicated}})}$. Agreement with the verdict: one-sided Fisher's exact test of the association between showing the effect and the replication outcome.

Experiment	Replication	Response, d_z [95% CI]	Agrees with the verdict	Construal: support for the mediator (NLI)	Construal: rating of the mediator	Construal supported	Reliance: mediator vs denial	Reliance supported	At least one step not supported
Original effect replicated (14 experiments)									
Moral typecasting	Many Labs 2	+5.54* [+5.17, +5.92]	yes	+0.07	+6.82*	yes	+0.52*	yes	no
Side effects and intent	Many Labs 2	+2.51* [+2.08, +2.93]	yes	+1.08*	+2.78*	yes	+1.36*	yes	no
Less is better	Many Labs 2	+2.01* [+1.62, +2.40]	yes	+2.79*	+3.42*	yes	+0.66*	yes	no
Quote attribution	Many Labs 1	+1.02* [+0.63, +1.40]	yes	+1.88*	+2.47*	yes	+0.99*	yes	no
Disease framing	Many Labs 1	+0.81* [+0.43, +1.19]	yes	+2.24*	+3.58*	yes	+2.07*	yes	no
Tempting fate	Many Labs 2	+0.71* [+0.33, +1.08]	yes	+0.06	+1.77*	yes	+0.43*	yes	no
Retrospective gambler's fallacy	Many Labs 1	+0.71* [+0.33, +1.08]	yes	−0.30	+1.19*	yes	+3.05*	yes	no
Norm of reciprocity	Many Labs 1	+0.66* [+0.28, +1.04]	yes	+0.08	−0.50*	no	+2.05*	yes	yes
Personal force	Registered report (Psychological Science Accelerator)	+0.43* [+0.06, +0.80]	yes	−0.07	+0.69*	yes	+0.83*	yes	no
Overhead aversion	Social Sciences Replication Project	+0.34 [−0.06, +0.75]	no	+0.15	+1.25*	yes	+1.79*	yes	no
Forbid vs allow	Many Labs 1	+0.18 [−0.21, +0.56]	no	+0.18	+1.14*	yes	+1.40*	yes	no

Experiment	Replication	Response, d_z [95% CI]	Agrees with the verdict	Construal: support for the mediator (NLI)	Construal: rating of the mediator	Construal supported	Reliance: mediator vs denial	Reliance supported	At least one step not supported
Anchoring	Many Labs 1	+0.10 [−0.27, +0.48]	no	+0.17	−0.26	no	+1.14*	yes	yes
Response-scale range	Many Labs 1	−0.08 [−0.45, +0.29]	no	−0.28	+0.27	no	+0.47*	yes	yes
Consumer label and trust	Many Labs 2	−0.27 [−0.65, +0.11]	no	−0.33	−0.02	no	+3.50*	yes	yes
Pooled (random effects)		+0.95 [+0.16, +1.75]; effect shown in 9/14	9/14	+0.51 [−0.06, +1.09]	+1.66 [+0.58, +2.75]	10/14	+1.39 [+0.86, +1.92]	14/14	4/14
Original effect not replicated (14 experiments)									
Vertical position and power †	Many Labs 2	+0.62* [+0.22, +1.03]	no	+0.90*	+1.59*	yes	+1.67*	yes	no
Affect-rich prizes	Many Labs 2	+0.47* [+0.08, +0.85]	no	−0.05	+0.58*	yes	+3.47*	yes	no
Physical warmth †	Preregistered three-site replication	+0.39* [+0.02, +0.76]	no	+1.59*	+0.37	yes	+1.49*	yes	no
Weight and importance †	Many Labs 3	+0.26 [−0.11, +0.64]	yes	−0.14	+0.08	no	+1.15*	yes	yes
Verb aspect and intent	Registered replication report	+0.23 [−0.14, +0.61]	yes	−0.13	−0.10	no	−0.01	no	yes
Structure and goal pursuit	Many Labs 2	+0.13 [−0.26, +0.51]	yes	+0.59*	+0.42*	yes	+0.74*	yes	no
Analytic thinking and belief	Social Sciences Replication Project	+0.10 [−0.28, +0.47]	yes	+0.20	+0.03	no	+0.03	no	yes
Choosing vs rejecting	Many Labs 2	+0.04 [−0.35, +0.44]	yes	+0.15	+0.38*	yes	+0.51*	yes	no
Incidental anchors	Many Labs 2	−0.02 [−0.39, +0.36]	yes	+0.34	+0.00	no	+1.21*	yes	yes
Flag priming †	Many Labs 1	−0.12 [−0.52, +0.27]	yes	+0.25	+0.26	no	+0.72*	yes	yes
Mortality salience	Many Labs 4	−0.17 [−0.55, +0.21]	yes	+0.43*	+0.34	yes	−0.18	no	yes
Currency priming †	Many Labs 1	−0.24 [−0.62, +0.14]	yes	−0.30	−0.17	no	+2.34*	yes	yes
Similarity asymmetry	Many Labs 2	−0.25 [−0.64, +0.13]	yes	+0.23	+0.00	no	+2.18*	yes	yes
Heat priming †	Many Labs 2	−0.36 [−0.74, +0.01]	yes	+0.03	+0.18	no	+1.13*	yes	yes
Pooled (random effects)		+0.07 [−0.10, +0.24]; effect shown in 3/14	11/14	+0.27 [+0.00, +0.54]	+0.26 [+0.03, +0.49]	6/14	+1.12 [+0.56, +1.69]	11/14	9/14

Experiment	Replication	Response, d_z [95% CI]	Agrees with the verdict	Construal: support for the mediator (NLI)	Construal: rating of the mediator	Construal supported	Reliance: mediator vs denial	Reliance supported	At least one step not supported
Difference between groups (replicated minus not replicated)		z = 2.35, P = 0.019		z = 0.82, P = 0.41	z = 2.74, P = 0.006		z = 0.74, P = 0.46		
Agreement with the verdict			20/28; one-sided Fisher's exact test, P = 0.027, odds ratio 6.6						

† The manipulation is an incidental cue, which, according to the theory, activates the mediator without any role in the decision (flag, currency, heat, weight, warmth, vertical position). In 4 of these 6 experiments, construal was not supported; none of the six original effects replicated.

Supplementary Table 9.

Tests of situational moderators in Study 3. Rows are grouped, as in Fig. 8A, by the classes of situational moderator specified before testing. Entries are d_z [95% CI] (asterisk, 95% CI excludes zero); positive values are in the predicted direction. The interaction is the effect of the induced mediator (vs its denial) at the level with the larger predicted effect minus that effect at the other level. (**A**) Gemma 4 26B, 35 to 50 cognitive agents per test. (**B**) Interaction with each LLM and averaged over the three LLMs, for the 30 cognitive agents simulated with all three LLMs.

(**A**) Gemma 4 26B

Test	Situational moderator (level with the larger predicted effect vs other level)	Induced mediator (one of two wordings)	Function of the situational moderator	General rule that applies	Effect of the induced mediator at the level with the larger predicted effect, d_z [95% CI]	Effect of the induced mediator at the other level, d_z [95% CI]	Interaction, d_z [95% CI]
The moderator is whether the option itself is linked to the mediator							
Identity salience: product for family vs features	product framed for family vs by its features	Being part of my family is central to who I am right now.	the description of the option states whether the option bears on the mediator	none	+0.77* [+0.49, +1.06]	−0.09 [−0.38, +0.19]	+0.72* [+0.44, +1.01]
Identity threat: film tied to own gender vs documentary	film tied to the cognitive agent's own gender vs documentary	Right now I'd rather not be lumped in with other {men/women}.	the description of the option states whether the option bears on the mediator	none	+1.16* [+0.86, +1.47]	+0.27 [−0.03, +0.57]	+0.84* [+0.54, +1.14]
Social exclusion: product signals group membership vs not	product shows group membership vs not	I really want other people to like me and accept me.	the description of the option states whether the option bears on the mediator	none	+0.50* [+0.21, +0.78]	−0.10 [−0.39, +0.18]	+0.45* [+0.17, +0.73]
The option can serve a desire only through what others see or infer from the choice							
Powerlessness: exclusive vs mass-produced product	exclusive vs mass-produced product	Right now I'd really like to feel more powerful and in control.	choosing the option serves the motive through what others see or infer from the choice	motive–means rule	+0.16 [−0.18, +0.51]	−0.21 [−0.50, +0.07]	+0.24 [−0.10, +0.58]
Going green: public vs private choice	choice in public vs in private	Right now I really want to get ahead of the people I know.	choosing the option serves the motive through what others see or infer from the choice	none	−0.36* [−0.64, −0.07]	+0.02 [−0.27, +0.30]	−0.24 [−0.52, +0.04]
Identity divergence: jacket vs dish soap	jacket vs dish soap	I don't want anyone taking me for one of those new buyers.	choosing the option serves the motive through what others see or infer from the choice	none	+0.14 [−0.14, +0.43]	+0.24 [−0.04, +0.52]	−0.18 [−0.46, +0.10]
Social exclusion: partner sees payment vs not	payment seen vs not seen by the other person	I really want Lin to like me when we meet today.	choosing the option serves the motive through what others see or infer from the choice	none	+0.56* [+0.28, +0.85]	+0.54* [+0.26, +0.83]	+0.04 [−0.24, +0.33]

Test	Situational moderator (level with the larger predicted effect vs other level)	Induced mediator (one of two wordings)	Function of the situational moderator	General rule that applies	Effect of the induced mediator at the level with the larger predicted effect, d_z [95% CI]	Effect of the induced mediator at the other level, d_z [95% CI]	Interaction, d_z [95% CI]
Remaining tests: the moderator alters the source of a feeling, what a feeling indicates about the product, the normative message, or the direction of comparison							
Decision difficulty: no print warning vs warning	no print warning vs print warning	Choosing between these two flights is harder than I expected.	source of a feeling	discounting rule	+0.64* [+0.36, +0.93]	−0.10 [−0.39, +0.20]	+0.49* [+0.19, +0.79]
Affect reliance: attractive vs plain advertisement	attractive advertisement with weak arguments vs plain advertisement with strong arguments	How this ad makes me feel is a good guide to this dictionary.	what a feeling indicates about the product	discounting rule	+0.38* [+0.10, +0.67]	−0.21 [−0.50, +0.07]	+0.40* [+0.12, +0.69]
Normative message: no approval vs approval	no approval vs approval (smiling face) added to the comparison with the neighbors	Since my household uses less than the neighbors, I have room to use more.	competing normative message	none	+0.95* [+0.67, +1.24]	+1.10* [+0.81, +1.38]	−0.16 [−0.45, +0.12]
Pennies a day: low vs high daily amount	low vs high daily amount	This donation feels to me like a small everyday expense, like a coffee.	direction of comparison	standard-comparison rule	+0.47* [+0.18, +0.77]	+0.81* [+0.52, +1.10]	−0.04 [−0.34, +0.25]
Pooled (random effects), 11 tests							+0.23 [−0.03, +0.48], positive 7/11, significant 5

(**B**) Interaction with each LLM, 30 cognitive agents, d_z (asterisk: 95% CI excludes zero)

Test	Gemma 4 26B	Qwen 3.7 Flash	Ling 3.0 Flash	Averaged over LLMs
The moderator is whether the option itself is linked to the mediator				
Identity salience: product for family vs features	+0.66*	+0.46*	+0.57*	+0.93*
Identity threat: film tied to own gender vs documentary	+0.79*	+0.33	+0.08	+0.59*
Social exclusion: product signals group membership vs not	+0.50*	−0.16	+0.08	+0.21
The option can serve a desire only through what others see or infer from the choice				
Powerlessness: exclusive vs mass-produced product	+0.03	+0.81*	+0.47*	+0.72*
Going green: public vs private choice	−0.17	+0.24	−0.06	+0.08
Identity divergence: jacket vs dish soap	−0.08	−0.30	−0.13	−0.34

Test	Gemma 4 26B	Qwen 3.7 Flash	Ling 3.0 Flash	Averaged over LLMs
Social exclusion: partner sees payment vs not	+0.18	−0.13	−0.24	−0.05
Remaining tests: the moderator alters the source of a feeling, what a feeling indicates about the product, the normative message, or the direction of comparison				
Decision difficulty: no print warning vs warning	+0.55*	+0.07	+0.58*	+0.61*
Affect reliance: attractive vs plain advertisement	+0.45*	+0.30	+0.20	+0.58*
Normative message: no approval vs approval	−0.34	+0.17	+0.29	+0.28
Pennies a day: low vs high daily amount	−0.07	−0.02	−0.09	−0.10
Pooled (random effects), 11 tests	+0.22 [−0.03, +0.48], positive 7/11, significant 5	+0.15 [−0.06, +0.36], positive 7/11, significant 2	+0.15 [−0.04, +0.34], positive 7/11, significant 3	+0.31 [+0.04, +0.57], positive 8/11, significant 5
Agreement across the 11 tests, Spearman ρ (same sign)	Gemma–Qwen 0.29 (7/11); Gemma–Ling 0.39 (9/11); Qwen–Ling 0.65 (9/11)			

All entries are from the first time each test was conducted; Supplementary Table 10 reports two repetitions of the six tests that did not reach significance in panel A.

In panel A, the last row reports the pooled random-effects estimate of the interaction and the numbers of tests with a positive and with a significant interaction.

The Supplementary Information describes the functions of the situational moderators (Supplementary Note 10) and the average over LLMs (Supplementary Note 15).

Placeholders in braces are filled from the agent profile.

Supplementary Table 10.

Repeated tests of Studies 3 and 4. Tests that did not reach significance with Gemma 4 26B the first time they were conducted, with two repetitions; d_z [95% CI]. After inspecting the first results, we conducted each of these tests twice more with Gemma 4 26B, the same 50 cognitive agents, and the same materials, so that each was conducted three times in all. The last column reports d_z for each cognitive agent's interaction averaged over the three times. Because this average uses every cognitive agent with a valid interaction in at least one of the three times, its n can exceed that of each time (powerlessness; effort and indulgence).

Test	First time	First repetition	Second repetition	Average over the three times for each cognitive agent
Study 3: Powerlessness: exclusive vs mass-produced product	+0.24 [−0.10, +0.58] (n 35)	+0.20 [−0.13, +0.52] (n 38)	+0.39 [+0.05, +0.72] (n 37)	+0.43 [+0.14, +0.72] (n 48)
Study 3: Social exclusion: partner sees payment vs not	+0.04 [−0.24, +0.33] (n 50)	−0.09 [−0.37, +0.20] (n 50)	+0.01 [−0.28, +0.29] (n 50)	−0.03 [−0.31, +0.26] (n 50)
Study 3: Normative message: no approval vs approval	−0.16 [−0.45, +0.12] (n 50)	+0.21 [−0.08, +0.49] (n 50)	+0.22 [−0.07, +0.50] (n 50)	+0.15 [−0.13, +0.44] (n 50)
Study 3: Identity divergence: jacket vs dish soap	−0.18 [−0.46, +0.10] (n 50)	−0.22 [−0.50, +0.07] (n 50)	−0.12 [−0.40, +0.17] (n 50)	−0.27 [−0.55, +0.01] (n 50)
Study 3: Going green: public vs private choice	−0.24 [−0.52, +0.04] (n 50)	−0.12 [−0.40, +0.16] (n 50)	−0.21 [−0.49, +0.08] (n 50)	−0.31 [−0.59, −0.02] (n 50)
Study 3: Pennies a day: low vs high daily amount	−0.04 [−0.34, +0.25] (n 47)	−0.13 [−0.43, +0.16] (n 47)	+0.08 [−0.21, +0.37] (n 47)	−0.04 [−0.33, +0.25] (n 47)
Study 4: Time versus money × materialism	+0.27 [−0.02, +0.55] (n 50)	+0.04 [−0.25, +0.32] (n 50)	+0.11 [−0.17, +0.40] (n 50)	+0.23 [−0.05, +0.52] (n 50)
Study 4: Effort and indulgence × guilt about luxuries	+0.28 [−0.01, +0.57] (n 49)	+0.10 [−0.19, +0.39] (n 48)	+0.13 [−0.17, +0.44] (n 44)	+0.28 [−0.002, +0.57] (n 50)
Study 4: Identity salience × importance of family identity	+0.15 [−0.13, +0.44] (n 50)	+0.05 [−0.24, +0.33] (n 50)	+0.22 [−0.06, +0.51] (n 50)	+0.27 [−0.01, +0.56] (n 50)

Pooled interaction for Study 3 with the averages over the three times: d_z = 0.25 [−0.003, 0.51], 11 tests, positive in 7. Pooled interaction for Study 4 with the averages: d_z = 0.41 [0.21, 0.62], 6 tests, positive in 6.

Supplementary Table 11.

Identification of the published experiments by a separate LLM. (**A**) Share of inputs from which DeepSeek V4.1 Flash named the correct published experiment (same experiment) and share from which it stated the published result, whichever experiment it named (same published result), for the input of each process; inputs of the first three cognitive agents that took part in each of the 60 experiments, in both conditions (Gemma 4 26B, Study 1; 360 inputs per row, 4,680 in all). (**B**) For each experiment, share of the six answers to the scenario and the question (three for each condition) that named the correct experiment. (**C**) This share and the effect of the manipulation on support for the mediator in Study 1: Spearman ρ across the 60 experiments, and pooled d_z in the 23 experiments never identified and the 37 identified at least once. (**D**) The same with the rating of the mediator by the separate LLM. (**E**) Pooled effects in experiments never identified and identified at least once, for three definitions of identification: from the scenario and the question; from the information that the LLM received for the ten processes between encoding and judgment; and from the information that it received for judgment. Study 2 rows cover the 52 experiments of Study 2. Difference: identified minus never identified, with an approximate 95% CI from the standard errors of the two pooled estimates.

(**A**) Inputs from which the experiment was identified

Input	n	same experiment	same published result
Scenario and question (what an LLM receives when it answers in one step)	360	38%	39%
Encoding	360	37%	39%
Categorization	360	25%	28%
Self-relevance	360	18%	20%
Goal relevance	360	23%	25%
Affective response	360	0%	0%
Causal inference	360	21%	24%
Social inference	360	23%	24%
Alternative readings	360	19%	20%
Comparison of alternatives	360	4%	4%
First-person account	360	7%	8%
Coherence review	360	4%	5%
Judgment	360	16%	17%

In 25 of the 360 answers to the scenario and the question, the stated result matched the published result although the named experiment was not the actual one.

(**B**) Share of answers that identified the experiment

Experiment	share of answers that identified the experiment
Buying and selling prices	100%
Foot in the door	100%
Endowment effect	100%
IKEA effect	100%
Normative message	100%
Beer price	100%
Side effects and intent	100%
Overhead aversion	100%
Disease framing	100%

Experiment	share of answers that identified the experiment
Anchoring	100%
Personal force	100%
Vertical position and power	100%
Physical warmth	100%
Brand extension	83%
Response-scale range	83%
Emotional attachment	67%
Payment transparency	67%
Anthropomorphized products	50%
Identity divergence	50%
Door in the face	50%
Future self and saving	50%
Affect-rich prizes	50%
Incidental anchors	50%
Norm of reciprocity	50%
Forbid vs allow	50%
Mortality salience	50%
Nonconforming dress	33%
Powerlessness	33%
Weight and importance	33%
Emotion and sharing	17%
Goal progress	17%
Pennies a day	17%
Going green	17%
Feature fatigue	17%
Choosing vs rejecting	17%
Similarity asymmetry	17%
Currency priming	17%
Labeling	0%
Shaken self-view	0%
Effort and indulgence	0%
Goal fluency	0%
Social exclusion	0%
Time versus money	0%
Mere touch	0%
Affect reliance	0%
Possession and valuation	0%

Experiment	share of answers that identified the experiment
Identity salience	0%
Decision difficulty	0%
Identity threat	0%
Moral typecasting	0%
Tempting fate	0%
Retrospective gambler's fallacy	0%
Consumer label and trust	0%
Verb aspect and intent	0%
Quote attribution	0%
Less is better	0%
Flag priming	0%
Heat priming	0%
Analytic thinking and belief	0%
Structure and goal pursuit	0%

(**C**) Support for the mediator

LLM	Spearman ρ (share of answers that identified the experiment, d_z)	P	never identified: significant increase	pooled d_z [95% CI]	identified at least once: significant	pooled d_z [95% CI]
Gemma 4 26B (50)	+0.14	0.30	5/23	+0.15 [−0.06, +0.37]	9/37	+0.25 [+0.07, +0.43]
Qwen 3.7 Flash (30)	+0.09	0.48	4/23	+0.18 [+0.01, +0.34]	10/37	+0.29 [+0.11, +0.46]
Ling 3.0 Flash (30)	+0.12	0.37	7/23	+0.30 [+0.05, +0.55]	13/37	+0.27 [+0.13, +0.41]
Average over LLMs (30)	+0.13	0.33	9/23	+0.37 [+0.07, +0.66]	14/37	+0.49 [+0.22, +0.75]

(**D**) Rating of the mediator by the separate LLM

LLM	k	Spearman ρ (share of answers that identified the experiment, mediator d_z)	P
Gemma 4 26B	60	−0.01	0.93
Qwen 3.7 Flash	60	+0.16	0.22
Ling 3.0 Flash	60	+0.03	0.82
Average over LLMs	60	+0.11	0.41

LLM		k	Spearman ρ (share of answers that identified the experiment, mediator d_z)	P
LLM	**experiments**	**k**	**pooled d_z [95% CI]**	**experiments with CI excluding 0**
Gemma 4 26B	never identified (0%)	23	+0.40 [+0.05, +0.75]	9
Gemma 4 26B	identified at least once	37	+0.36 [+0.09, +0.63]	17
Qwen 3.7 Flash	never identified (0%)	23	+0.43 [+0.02, +0.84]	10
Qwen 3.7 Flash	identified at least once	37	+0.57 [+0.27, +0.88]	24
Ling 3.0 Flash	never identified (0%)	23	+0.55 [+0.12, +0.97]	11
Ling 3.0 Flash	identified at least once	37	+0.49 [+0.23, +0.74]	17
Average over LLMs	never identified (0%)	23	+0.80 [+0.15, +1.44]	13
Average over LLMs	identified at least once	37	+0.77 [+0.39, +1.14]	23

(**E**) Pooled effects for three definitions of identification

From the scenario and the question: 23 experiments never identified, 37 identified at least once

Effect	LLM	never identified: k, pooled d_z [95% CI]	identified: k, pooled d_z [95% CI]	difference [approx. 95% CI]	Spearman ρ (share, d_z), P
Study 1, support	Gemma 4 26B (50)	23, +0.15 [−0.06, +0.37]	37, +0.25 [+0.07, +0.43]	+0.10 [−0.17, +0.36]	+0.14, 0.30
Study 1, support	Qwen 3.7 Flash (30)	23, +0.18 [+0.01, +0.34]	37, +0.29 [+0.11, +0.46]	+0.11 [−0.12, +0.34]	+0.09, 0.48
Study 1, support	Ling 3.0 Flash (30)	23, +0.30 [+0.05, +0.55]	37, +0.27 [+0.13, +0.41]	−0.03 [−0.31, +0.24]	+0.12, 0.37
Study 1, support	Average over LLMs (30)	23, +0.37 [+0.07, +0.66]	37, +0.49 [+0.22, +0.75]	+0.12 [−0.26, +0.50]	+0.13, 0.33
Study 1, LLM rating	Gemma 4 26B (50)	23, +0.40 [+0.05, +0.75]	37, +0.36 [+0.09, +0.63]	−0.04 [−0.46, +0.38]	−0.01, 0.93
Study 1, LLM rating	Qwen 3.7 Flash (30)	23, +0.43 [+0.02, +0.84]	37, +0.57 [+0.27, +0.88]	+0.14 [−0.34, +0.63]	+0.16, 0.22
Study 1, LLM rating	Ling 3.0 Flash (30)	23, +0.55 [+0.12, +0.97]	37, +0.49 [+0.23, +0.74]	−0.06 [−0.53, +0.41]	+0.03, 0.82
Study 1, LLM rating	Average over LLMs (30)	23, +0.80 [+0.15, +1.44]	37, +0.77 [+0.39, +1.14]	−0.03 [−0.74, +0.68]	+0.11, 0.41
Study 2, mediator vs denial (52 experiments)	Gemma 4 26B (50)	19, +0.59 [+0.32, +0.85]	33, +1.03 [+0.71, +1.35]	+0.44 [+0.04, +0.84]	+0.32, 0.02
Study 2, mediator vs denial (52 experiments)	Qwen 3.7 Flash (30)	19, +0.47 [+0.19, +0.76]	33, +0.73 [+0.54, +0.91]	+0.25 [−0.06, +0.57]	+0.19, 0.19
Study 2, mediator vs denial (52 experiments)	Ling 3.0 Flash (30)	19, +0.75 [+0.31, +1.18]	33, +0.80 [+0.56, +1.04]	+0.05 [−0.42, +0.52]	+0.11, 0.46
Study 2, mediator vs denial (52 experiments)	Average over LLMs (30)	19, +0.90 [+0.45, +1.36]	33, +1.33 [+1.00, +1.65]	+0.43 [−0.10, +0.95]	+0.22, 0.11

From the information that the LLM received for the ten processes between encoding and judgment: 26 experiments never identified, 34 identified at least once

Effect	LLM	never identified: k, pooled d_z [95% CI]	identified: k, pooled d_z [95% CI]	difference [approx. 95% CI]	Spearman ρ (share, d_z), P
Study 1, support	Gemma 4 26B (50)	26, +0.13 [−0.06, +0.33]	34, +0.27 [+0.09, +0.46]	+0.14 [−0.12, +0.40]	+0.20, 0.12
Study 1, support	Qwen 3.7 Flash (30)	26, +0.14 [−0.01, +0.29]	34, +0.33 [+0.14, +0.51]	+0.19 [−0.04, +0.42]	+0.18, 0.17
Study 1, support	Ling 3.0 Flash (30)	26, +0.21 [−0.00, +0.43]	34, +0.33 [+0.18, +0.49]	+0.12 [−0.13, +0.38]	+0.25, 0.06
Study 1, support	Average over LLMs (30)	26, +0.29 [+0.04, +0.55]	34, +0.56 [+0.27, +0.84]	+0.26 [−0.10, +0.63]	+0.23, 0.07
Study 1, LLM rating	Gemma 4 26B (50)	26, +0.34 [+0.03, +0.66]	34, +0.40 [+0.11, +0.68]	+0.05 [−0.36, +0.46]	+0.01, 0.94
Study 1, LLM rating	Qwen 3.7 Flash (30)	26, +0.36 [−0.01, +0.74]	34, +0.64 [+0.32, +0.95]	+0.28 [−0.19, +0.75]	+0.19, 0.16
Study 1, LLM rating	Ling 3.0 Flash (30)	26, +0.44 [+0.07, +0.81]	34, +0.56 [+0.29, +0.83]	+0.12 [−0.32, +0.56]	+0.15, 0.25
Study 1, LLM rating	Average over LLMs (30)	26, +0.67 [+0.09, +1.25]	34, +0.86 [+0.47, +1.25]	+0.19 [−0.48, +0.86]	+0.13, 0.32
Study 2, mediator vs denial (52 experiments)	Gemma 4 26B (50)	22, +0.66 [+0.44, +0.87]	30, +1.02 [+0.66, +1.39]	+0.37 [−0.04, +0.77]	+0.21, 0.13
Study 2, mediator vs denial (52 experiments)	Qwen 3.7 Flash (30)	22, +0.67 [+0.42, +0.92]	30, +0.61 [+0.40, +0.82]	−0.06 [−0.36, +0.25]	+0.03, 0.86
Study 2, mediator vs denial (52 experiments)	Ling 3.0 Flash (30)	22, +0.85 [+0.46, +1.23]	30, +0.73 [+0.47, +0.98]	−0.12 [−0.56, +0.32]	+0.02, 0.90
Study 2, mediator vs denial (52 experiments)	Average over LLMs (30)	22, +1.13 [+0.75, +1.52]	30, +1.20 [+0.82, +1.58]	+0.07 [−0.45, +0.58]	+0.08, 0.56

From the information that the LLM received for judgment: 41 experiments never identified, 19 identified at least once

Effect	LLM	never identified: k, pooled d_z [95% CI]	identified: k, pooled d_z [95% CI]	difference [approx. 95% CI]	Spearman ρ (share, d_z), P
Study 1, support	Gemma 4 26B (50)	41, +0.17 [+0.02, +0.31]	19, +0.32 [+0.02, +0.62]	+0.15 [−0.16, +0.46]	+0.16, 0.23
Study 1, support	Qwen 3.7 Flash (30)	41, +0.22 [+0.07, +0.37]	19, +0.30 [+0.08, +0.52]	+0.08 [−0.17, +0.34]	+0.13, 0.33
Study 1, support	Ling 3.0 Flash (30)	41, +0.24 [+0.08, +0.39]	19, +0.38 [+0.14, +0.61]	+0.14 [−0.12, +0.41]	+0.17, 0.20
Study 1, support	Average over LLMs (30)	41, +0.37 [+0.16, +0.58]	19, +0.60 [+0.15, +1.05]	+0.23 [−0.23, +0.70]	+0.12, 0.35
Study 1, LLM rating	Gemma 4 26B (50)	41, +0.30 [+0.08, +0.52]	19, +0.54 [+0.05, +1.02]	+0.24 [−0.26, +0.74]	+0.13, 0.32
Study 1, LLM rating	Qwen 3.7 Flash (30)	41, +0.53 [+0.23, +0.82]	19, +0.50 [+0.07, +0.94]	−0.02 [−0.52, +0.47]	+0.11, 0.39
Study 1, LLM rating	Ling 3.0 Flash (30)	41, +0.46 [+0.20, +0.71]	19, +0.61 [+0.19, +1.04]	+0.15 [−0.32, +0.62]	+0.15, 0.26
Study 1, LLM rating	Average over LLMs (30)	41, +0.72 [+0.32, +1.12]	19, +0.88 [+0.28, +1.48]	+0.16 [−0.52, +0.84]	+0.15, 0.25
Study 2, mediator vs denial (52 experiments)	Gemma 4 26B (50)	35, +0.73 [+0.48, +0.98]	17, +1.14 [+0.66, +1.63]	+0.41 [−0.09, +0.92]	+0.24, 0.09
Study 2, mediator vs denial (52 experiments)	Qwen 3.7 Flash (30)	35, +0.62 [+0.40, +0.83]	17, +0.68 [+0.47, +0.89]	+0.06 [−0.22, +0.35]	+0.12, 0.40
Study 2, mediator vs denial (52 experiments)	Ling 3.0 Flash (30)	35, +0.83 [+0.53, +1.13]	17, +0.68 [+0.43, +0.93]	−0.15 [−0.52, +0.22]	−0.00, 0.98
Study 2, mediator vs denial (52 experiments)	Average over LLMs (30)	35, +1.14 [+0.78, +1.51]	17, +1.24 [+0.89, +1.60]	+0.10 [−0.39, +0.58]	+0.14, 0.32

Supplementary Table 12.

Names of published findings in the process statements (Study 1). We searched the process statements of the ten processes between encoding and judgment in all Study 1 cognitive simulations of the 60 experiments (12,983 cognitive simulations: Gemma 4 26B, 50 cognitive agents; Qwen 3.7 Flash and Ling 3.0 Flash, 30 cognitive agents) for the names and words listed in the Supplementary Information (Supplementary Note 17). A match counted only if the matched words did not occur in the scenario of the same cognitive simulation. (**A**) The name under which the experiment's own finding is known, in the 47 experiments whose finding has an established name. (**B**) Experiments in which the experiment's own finding was named. (**C**) Words that refer to research in general, in all 60 experiments and in the 45 experiments whose scenarios contain none of these words.

(**A**) Own finding named

LLM	Cognitive simulations in the 47 experiments	Cognitive simulations in which a process statement named the finding	Process statements	Process statements that named the finding
Gemma 4 26B	4,648	20 (0.4%)	46,480	29 (0.06%)
Qwen 3.7 Flash	2,802	177 (6.3%)	28,020	417 (1.49%)
Ling 3.0 Flash	2,801	110 (3.9%)	28,007	245 (0.87%)

(**B**) Experiments in which the own finding was named (cognitive simulations with the name / cognitive simulations)

Experiment	Gemma 4 26B	Qwen 3.7 Flash	Ling 3.0 Flash
Disease framing	1 / 100	31 / 60	39 / 60
Personal force	10 / 100	31 / 60	25 / 60
Endowment effect	0 / 100	33 / 60	16 / 60
Emotional attachment	4 / 100	33 / 60	5 / 60
Anchoring	5 / 100	4 / 60	13 / 60
IKEA effect	0 / 100	21 / 60	0 / 60
Buying and selling prices	0 / 100	6 / 60	4 / 60
Possession and valuation	0 / 100	6 / 60	0 / 60
Incidental anchors	0 / 100	3 / 60	3 / 60
Door in the face	0 / 100	3 / 60	2 / 60
Currency priming	0 / 92	1 / 58	1 / 58
Physical warmth	0 / 100	2 / 60	0 / 60
Goal progress	0 / 100	0 / 60	1 / 60
Pennies a day	0 / 94	1 / 56	0 / 56
Affect-rich prizes	0 / 100	0 / 60	1 / 60
Weight and importance	0 / 100	1 / 60	0 / 60
Mortality salience	0 / 92	1 / 58	0 / 58

(**C**) Words that refer to research in general

LLM	Process statements	Process statements with a research word	Process statements in experiments whose scenarios contain no research word	Process statements with a research word in these experiments	Most frequent words in these experiments
Gemma 4 26B	58,920	3,606 (6.12%)	44,560	2,102 (4.72%)	researchers 776, manipulation 352, experiment 341, the study 313, participants 307
Qwen 3.7 Flash	35,460	3,842 (10.83%)	26,840	2,363 (8.80%)	experiment 783, manipulation 712, researchers 654, the study 382, participant 317
Ling 3.0 Flash	35,446	1,612 (4.55%)	26,826	984 (3.67%)	researchers 442, the study 426, researcher 183, participant 153, experiment 105

The surnames of the authors of the 60 original articles matched in 3 of the 12,983 cognitive simulations, all with Qwen 3.7 Flash. In 2 of these, both from the disease-framing experiment, a process statement named the authors of the original article; the third match referred to the government of North Korea, not to an author.

Supplementary Table 13.

Language models used in the four studies. All LLMs were accessed through the OpenRouter API. Gemma 4 26B A4B was served by five providers, all with the same bf16 weights; each other LLM was served by one fixed provider, with no fallback to other providers. The text calls the LLMs Gemma 4 26B, Qwen 3.7 Flash, Ling 3.0 Flash, and DeepSeek V4.1 Flash. Every LLM that carried out the cognitive processes received the same materials, agent profiles, system prompts, and rules. Numbers of cognitive simulations are those conducted in Studies 1 and 2 and those with a valid response in Studies 3 and 4.

Role	Model (OpenRouter identifier)	Provider	Settings	Studies	Cognitive agents; cognitive simulations	Dates of data collection (Studies 1 to 4)
Cognitive processes and rules	Gemma 4 26B A4B (google/gemma-4-26b-a4b-it)	five providers serving the same bf16 weights	temperature 0.2; extended reasoning off; JSON output	1–4	all 50 cognitive agents; Study 1, 5,892; Study 2, 13,054; Study 3, 3,224 (repetitions, 3,521); Study 4, 1,578 (repetitions, 1,192)	7–13 October 2026
Cognitive processes and rules	Qwen 3.7 Flash (qwen/qwen3.7-flash)	Alibaba	temperature 0.2; reasoning disabled; JSON output	1–4	the 30 cognitive agents simulated with all three LLMs; Study 1, 3,546; Study 2, 7,860; Study 3, 1,944; Study 4, 948	8–13 October 2026
Cognitive processes and rules	Ling 3.0 Flash (inclusionai/ling-3.0-flash)	Novita	temperature 0.2; reasoning disabled; JSON requested in the instructions (the provider does not offer a JSON output mode)	1–4	the 30 cognitive agents simulated with all three LLMs; Study 1, 3,545; Study 2, 7,858 (first of two sets); Study 3, 1,928; Study 4, 946	8–13 October 2026
Rater of process statements (Study 1)	DeepSeek V4.1 Flash (deepseek/deepseek-v4.1-flash)	DeepInfra	temperature 0; reasoning off; JSON output	1	process statements produced with each of the three LLMs	7–13 October 2026
Identification test	DeepSeek V4.1 Flash (deepseek/deepseek-v4.1-flash)	DeepInfra	temperature 0; JSON output	identification test	inputs from the Study 1 cognitive simulations with Gemma 4 26B of the first three cognitive agents in each experiment (4,680 inputs)	9 and 13 October 2026
Natural language inference model (Study 1)	cross-encoder/nli-deberta-v3-large	computed locally on a GPU (Google Colab)	half precision; maximum input 384 tokens	1	process statements produced with each of the three LLMs	7–13 October 2026

Supplementary References

1. Lakatos, I. Falsification and the methodology of scientific research programmes. In *Criticism and the Growth of Knowledge* (eds Lakatos, I. & Musgrave, A.) 91–196 (Cambridge Univ. Press, Cambridge, 1970).

2. Meehl, P. E. Appraising and amending theories: the strategy of Lakatosian defense and two principles that warrant it. *Psychol. Inq.* **1**, 108–141 (1990).

3. Earp, B. D. & Trafimow, D. Replication, falsification, and the crisis of confidence in social psychology. *Front. Psychol.* **6**, 621 (2015).

4. Scheel, A. M., Tiokhin, L., Isager, P. M. & Lakens, D. Why hypothesis testers should spend less time testing hypotheses. *Perspect. Psychol. Sci.* **16**, 744–755 (2021).

5. Spencer, S. J., Zanna, M. P. & Fong, G. T. Establishing a causal chain: why experiments are often more effective than mediational analyses in examining psychological processes. *J. Pers. Soc. Psychol.* **89**, 845–851 (2005).

6. Bullock, J. G., Green, D. P. & Ha, S. E. Yes, but what's the mechanism? (Don't expect an easy answer). *J. Pers. Soc. Psychol.* **98**, 550–558 (2010).

7. Pirlott, A. G. & MacKinnon, D. P. Design approaches to experimental mediation. *J. Exp. Soc. Psychol.* **66**, 29–38 (2016).

8. Wyer Jr., R. S. & Srull, T. K. Human cognition in its social context. *Psychol. Rev.* **93**, 322–359 (1986).

9. Ross, L. & Nisbett, R. E. *The Person and the Situation: Perspectives of Social Psychology* (McGraw-Hill, New York, 1991).

10. Griffin, D. W. & Ross, L. Subjective construal, social inference, and human misunderstanding. In *Advances in Experimental Social Psychology* Vol. 24 (ed. Zanna, M. P.) 319–359 (Academic Press, San Diego, 1991).

11. Cronbach, L. J. & Meehl, P. E. Construct validity in psychological tests. *Psychol. Bull.* **52**, 281–302 (1955).

12. Shadish, W. R., Cook, T. D. & Campbell, D. T. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (Houghton Mifflin, Boston, 2002).

13. Schwarz, N. & Clore, G. L. Mood, misattribution, and judgments of well-being: informative and directive functions of affective states. *J. Pers. Soc. Psychol.* **45**, 513–523 (1983).

14. Schwarz, N. Attitude construction: evaluation in context. *Soc. Cogn.* **25**, 638–656 (2007).

15. Higgins, E. T. Knowledge activation: accessibility, applicability, and salience. In *Social Psychology: Handbook of Basic Principles* (eds Higgins, E. T. & Kruglanski, A. W.) 133–168 (Guilford Press, New York, 1996).

16. Feldman, J. M. & Lynch Jr., J. G. Self-generated validity and other effects of measurement on belief, attitude, intention, and behavior. *J. Appl. Psychol.* **73**, 421–435 (1988).

17. Jacoby, J. & Sassenberg, K. Interactions do not only tell us when, but can also tell us how: testing process hypotheses by interaction. *Eur. J. Soc. Psychol.* **41**, 180–190 (2011).

18. Lewin, K. *Principles of Topological Psychology* (McGraw-Hill, New York, 1936).

19. Mischel, W. & Shoda, Y. A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychol. Rev.* **102**, 246–268 (1995).

20. Tett, R. P. & Guterman, H. A. Situation trait relevance, trait expression, and cross-situational consistency: testing a principle of trait activation. *J. Res. Pers.* **34**, 397–423 (2000).

21. White, K. & Argo, J. J. Social identity threat and consumer preferences. *J. Consum. Psychol.* **19**, 313–325 (2009).

22. Bolger, N., Zee, K. S., Rossignac-Milon, M. & Hassin, R. R. Causal processes in psychology are heterogeneous. *J. Exp. Psychol. Gen.* **148**, 601–618 (2019).

23. Bryan, C. J., Tipton, E. & Yeager, D. S. Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nat. Hum. Behav.* **5**, 980–989 (2021).

24. Ludwig, J., Mullainathan, S. & Rambachan, A. Large language models: an applied econometric framework. Preprint at https://arxiv.org/abs/2412.07031 (2024).

25. Orne, M. T. On the social psychology of the psychological experiment: with particular reference to demand characteristics and their implications. *Am. Psychol.* **17**, 776–783 (1962).

26. Barrett, L. F. The future of psychology: connecting mind to brain. *Perspect. Psychol. Sci.* **4**, 326–339 (2009).

27. McClelland, J. L. et al. Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends Cogn. Sci.* **14**, 348–356 (2010).

28. Lindquist, K. A. Emotions emerge from more basic psychological ingredients: a modern psychological constructionist model. *Emot. Rev.* **5**, 356–368 (2013).

29. Moors, A., Ellsworth, P. C., Scherer, K. R. & Frijda, N. H. Appraisal theories of emotion: state of the art and future development. *Emot. Rev.* **5**, 119–124 (2013).

30. Belk, R. W. Possessions and the extended self. *J. Consum. Res.* **15**, 139–168 (1988).

31. Pierce, J. L., Kostova, T. & Dirks, K. T. The state of psychological ownership: integrating and extending a century of research. *Rev. Gen. Psychol.* **7**, 84–107 (2003).

32. Peck, J. & Shu, S. B. The effect of mere touch on perceived ownership. *J. Consum. Res.* **36**, 434–447 (2009).

33. Park, J. S. et al. Generative agents: interactive simulacra of human behavior. In *Proc. 36th Annual ACM Symposium on User Interface Software and Technology* 1–22 (ACM, 2023).

34. Smith, C. A. & Ellsworth, P. C. Patterns of cognitive appraisal in emotion. *J. Pers. Soc. Psychol.* **48**, 813–838 (1985).

35. Watson, L. & Spence, M. T. Causes and consequences of emotions on consumer behaviour: a review and integrative cognitive appraisal theory. *Eur. J. Mark.* **41**, 487–511 (2007).

36. Baars, B. J. *A Cognitive Theory of Consciousness* (Cambridge Univ. Press, Cambridge, 1988).

37. Dehaene, S. & Naccache, L. Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition* **79**, 1–37 (2001).

38. Kruglanski, A. W. et al. A theory of goal systems. *Adv. Exp. Soc. Psychol.* **34**, 331–378 (2002).

39. Bless, H. & Schwarz, N. Mental construal and the emergence of assimilation and contrast effects. *Adv. Exp. Soc. Psychol.* **42**, 319–373 (2010).

40. Aaker, D. A. & Keller, K. L. Consumer evaluations of brand extensions. *J. Mark.* **54**, 27–41 (1990).

41. Kruglanski, A. W. & Webster, D. M. Motivated closing of the mind: "seizing" and "freezing." *Psychol. Rev.* **103**, 263–283 (1996).

42. Cacioppo, J. T. & Petty, R. E. The need for cognition. *J. Pers. Soc. Psychol.* **42**, 116–131 (1982).

43. Higgins, E. T. Self-discrepancy: a theory relating self and affect. *Psychol. Rev.* **94**, 319–340 (1987).

44. Thaler, R. H. Mental accounting and consumer choice. *Mark. Sci.* **4**, 199–214 (1985).

45. Sela, A. & Berger, J. Decision quicksand: how trivial choices suck us in. *J. Consum. Res.* **39**, 360–370 (2012).

46. Kelley, H. H. The processes of causal attribution. *Am. Psychol.* **28**, 107–128 (1973).

47. Argyle, L. P. et al. Out of one, many: using language models to simulate human samples. *Polit. Anal.* **31**, 337–351 (2023).

48. Park, J. S. et al. Generative agent simulations of 1,000 people. Preprint at https://arxiv.org/abs/2411.10109 (2024).

49. Klein, R. A. et al. Investigating variation in replicability: a "many labs" replication project. *Soc. Psychol.* **45**, 142–152 (2014).

50. Klein, R. A. et al. Many Labs 2: investigating variation in replicability across samples and settings. *Adv. Methods Pract. Psychol. Sci.* **1**, 443–490 (2018).

51. Ebersole, C. R. et al. Many Labs 3: evaluating participant pool quality across the academic semester via replication. *J. Exp. Soc. Psychol.* **67**, 68–82 (2016).

52. Klein, R. A. et al. Many Labs 4: failure to replicate mortality salience effect with and without original author involvement. *Collabra Psychol.* **8**, 35271 (2022).

53. Eerland, A. et al. Registered replication report: Hart & Albarracín (2011). *Perspect. Psychol. Sci.* **11**, 158–171 (2016).

54. Camerer, C. F. et al. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nat. Hum. Behav.* **2**, 637–644 (2018).

55. Bago, B. et al. Situational factors shape moral judgements in the trolley dilemma in Eastern, Southern and Western countries in a culturally diverse sample. *Nat. Hum. Behav.* **6**, 880–895 (2022).

56. Lynott, D. et al. Replication of "Experiencing physical warmth promotes interpersonal warmth" by Williams and Bargh (2008). *Soc. Psychol.* **45**, 216–222 (2014).

57. Mead, N. L., Baumeister, R. F., Stillman, T. F., Rawn, C. D. & Vohs, K. D. Social exclusion causes people to spend and consume strategically in the service of affiliation. *J. Consum. Res.* **37**, 902–919 (2011).

58. Rucker, D. D. & Galinsky, A. D. Desire to acquire: powerlessness and compensatory consumption. *J. Consum. Res.* **35**, 257–267 (2008).

59. Pham, M. T. & Avnet, T. Ideals and oughts and the reliance on affect versus substance in persuasion. *J. Consum. Res.* **30**, 503–518 (2004).

60. Gourville, J. T. Pennies-a-day: the effect of temporal reframing on transaction evaluation. *J. Consum. Res.* **24**, 395–403 (1998).

61. Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J. & Griskevicius, V. The constructive, destructive, and reconstructive power of social norms. *Psychol. Sci.* **18**, 429–434 (2007).

62. Aggarwal, P. & McGill, A. L. Is that car smiling at me? Schema congruity as a basis for evaluating anthropomorphized products. *J. Consum. Res.* **34**, 468–479 (2007).

63. Bauer, M. A., Wilkie, J. E. B., Kim, J. K. & Bodenhausen, G. V. Cuing consumerism: situational materialism undermines personal and social well-being. *Psychol. Sci.* **23**, 517–523 (2012).

64. Bellezza, S., Gino, F. & Keinan, A. The red sneakers effect: inferring status and competence from signals of nonconformity. *J. Consum. Res.* **41**, 35–54 (2014).

65. Berger, J. & Heath, C. Who drives divergence? Identity signaling, outgroup dissimilarity, and the abandonment of cultural tastes. *J. Pers. Soc. Psychol.* **95**, 593–607 (2008).

66. Berger, J. & Milkman, K. L. What makes online content viral? *J. Mark. Res.* **49**, 192–205 (2012).

67. Burger, J. M. & Caldwell, D. F. The effects of monetary incentives and labeling on the foot-in-the-door effect: evidence for a self-perception process. *Basic Appl. Soc. Psychol.* **25**, 235–241 (2003).

68. Carmon, Z. & Ariely, D. Focusing on the forgone: how value can appear so different to buyers and sellers. *J. Consum. Res.* **27**, 360–370 (2000).

69. Carter, T. J., Ferguson, M. J. & Hassin, R. R. A single exposure to the American flag shifts support toward Republicanism up to 8 months later. *Psychol. Sci.* **22**, 1011–1018 (2011).

70. Caruso, E. M., Vohs, K. D., Baxter, B. & Waytz, A. Mere exposure to money increases endorsement of free-market systems and social inequality. *J. Exp. Psychol. Gen.* **142**, 301–306 (2013).

71. Cialdini, R. B. et al. Reciprocal concessions procedure for inducing compliance: the door-in-the-face technique. *J. Pers. Soc. Psychol.* **31**, 206–215 (1975).

72. Critcher, C. R. & Gilovich, T. Incidental environmental anchors. *J. Behav. Decis. Mak.* **21**, 241–251 (2008).

73. Fishbach, A. & Dhar, R. Goals as excuses or guides: the liberating effect of perceived goal progress on choice. *J. Consum. Res.* **32**, 370–377 (2005).

74. Freedman, J. L. & Fraser, S. C. Compliance without pressure: the foot-in-the-door technique. *J. Pers. Soc. Psychol.* **4**, 195–202 (1966).

75. Gao, L., Wheeler, S. C. & Shiv, B. The "shaken self": product choices as a means of restoring self-view confidence. *J. Consum. Res.* **36**, 29–38 (2009).

76. Gervais, W. M. & Norenzayan, A. Analytic thinking promotes religious disbelief. *Science* **336**, 493–496 (2012).

77. Giessner, S. R. & Schubert, T. W. High in the hierarchy: how vertical location and judgments of leaders' power are interrelated. *Organ. Behav. Hum. Decis. Process.* **104**, 30–44 (2007).

78. Gneezy, U., Keenan, E. A. & Gneezy, A. Avoiding overhead aversion in charity. *Science* **346**, 632–635 (2014).

79. Gray, K. & Wegner, D. M. Moral typecasting: divergent perceptions of moral agents and moral patients. *J. Pers. Soc. Psychol.* **96**, 505–520 (2009).

80. Greenberg, J., Pyszczynski, T., Solomon, S., Simon, L. & Breus, M. Role of consciousness and accessibility of death-related thoughts in mortality salience effects. *J. Pers. Soc. Psychol.* **67**, 627–637 (1994).

81. Greene, J. D. et al. Pushing moral buttons: the interaction between personal force and intention in moral judgment. *Cognition* **111**, 364–371 (2009).

82. Griskevicius, V., Tybur, J. M. & Van den Bergh, B. Going green to be seen: status, reputation, and conspicuous conservation. *J. Pers. Soc. Psychol.* **98**, 392–404 (2010).

83. Hart, W. & Albarracín, D. Learning about what others were doing: verb aspect and attributions of mundane and criminal intent for past actions. *Psychol. Sci.* **22**, 261–266 (2011).

84. Hershfield, H. E. et al. Increasing saving behavior through age-progressed renderings of the future self. *J. Mark. Res.* **48**, S23–S37 (2011).

85. Hsee, C. K. Less is better: when low-value options are valued more highly than high-value options. *J. Behav. Decis. Mak.* **11**, 107–121 (1998).

86. Hyman, H. H. & Sheatsley, P. B. The current status of American public opinion. In *The Teaching of Contemporary Affairs: 21st Yearbook of the National Council for the Social Studies* (ed. Payne, J. C.) 11–34 (National Council for the Social Studies, Washington, 1950).

87. Jacowitz, K. E. & Kahneman, D. Measures of anchoring in estimation tasks. *Pers. Soc. Psychol. Bull.* **21**, 1161–1166 (1995).

88. Jostmann, N. B., Lakens, D. & Schubert, T. W. Weight as an embodiment of importance. *Psychol. Sci.* **20**, 1169–1174 (2009).

89. Kahneman, D., Knetsch, J. L. & Thaler, R. H. Experimental tests of the endowment effect and the Coase theorem. *J. Polit. Econ.* **98**, 1325–1348 (1990).

90. Kay, A. C., Laurin, K., Fitzsimons, G. M. & Landau, M. J. A functional basis for structure-seeking: exposure to structure promotes willingness to engage in motivated action. *J. Exp. Psychol. Gen.* **143**, 486–491 (2014).

91. Kivetz, R. & Simonson, I. Earning the right to indulge: effort as a determinant of customer preferences toward frequency program rewards. *J. Mark. Res.* **39**, 155–170 (2002).

92. Knobe, J. Intentional action and side effects in ordinary language. *Analysis* **63**, 190–194 (2003).

93. Labroo, A. A. & Lee, A. Y. Between two brands: a goal fluency account of brand evaluation. *J. Mark. Res.* **43**, 374–385 (2006).

94. Lorge, I. & Curtiss, C. C. Prestige, suggestion, and attitudes. *J. Soc. Psychol.* **7**, 386–402 (1936).

95. Mogilner, C. & Aaker, J. "The time vs. money effect": shifting product attitudes and decisions through personal connection. *J. Consum. Res.* **36**, 277–291 (2009).

96. Norton, M. I., Mochon, D. & Ariely, D. The IKEA effect: when labor leads to love. *J. Consum. Psychol.* **22**, 453–460 (2012).

97. Oppenheimer, D. M. & Monin, B. The retrospective gambler's fallacy: unlikely events, constructing the past, and multiple universes. *Judgm. Decis. Mak.* **4**, 326–334 (2009).

98. Reb, J. & Connolly, T. Possession, feelings of ownership and the endowment effect. *Judgm. Decis. Mak.* **2**, 107–114 (2007).

99. Reed II, A. Activating the self-importance of consumer selves: exploring identity salience effects on judgments. *J. Consum. Res.* **31**, 286–295 (2004).

100. Risen, J. L. & Gilovich, T. Why people are reluctant to tempt fate. *J. Pers. Soc. Psychol.* **95**, 293–307 (2008).

101. Rottenstreich, Y. & Hsee, C. K. Money, kisses, and electric shocks: on the affective psychology of risk. *Psychol. Sci.* **12**, 185–190 (2001).

102. Rugg, D. Experiments in wording questions: II. *Public Opin. Q.* **5**, 91–92 (1941).

103. Schwarz, N., Hippler, H.-J., Deutsch, B. & Strack, F. Response scales: effects of category range on reported behavior and comparative judgments. *Public Opin. Q.* **49**, 388–395 (1985).

104. Shafir, E. Choosing versus rejecting: why some options are both better and worse than others. *Mem. Cognit.* **21**, 546–556 (1993).

105. Shu, S. B. & Peck, J. Psychological ownership and affective reaction: emotional attachment process variables and the endowment effect. *J. Consum. Psychol.* **21**, 439–452 (2011).

106. Soman, D. Effects of payment mechanism on spending behavior: the role of rehearsal and immediacy of payments. *J. Consum. Res.* **27**, 460–474 (2001).

107. Thompson, D. V., Hamilton, R. W. & Rust, R. T. Feature fatigue: when product capabilities become too much of a good thing. *J. Mark. Res.* **42**, 431–442 (2005).

108. Tversky, A. & Gati, I. Studies of similarity. In *Cognition and Categorization* (eds Rosch, E. & Lloyd, B. B.) 79–98 (Lawrence Erlbaum, Hillsdale, 1978).

109. Tversky, A. & Kahneman, D. The framing of decisions and the psychology of choice. *Science* **211**, 453–458 (1981).

110. Williams, L. E. & Bargh, J. A. Experiencing physical warmth promotes interpersonal warmth. *Science* **322**, 606–607 (2008).

111. Zaval, L., Keenan, E. A., Johnson, E. J. & Weber, E. U. How warm days increase belief in global warming. *Nat. Clim. Chang.* **4**, 143–147 (2014).

112. Richins, M. L. & Dawson, S. A consumer values orientation for materialism and its measurement: scale development and validation. *J. Consum. Res.* **19**, 303–316 (1992).

113. Cialdini, R. B., Trost, M. R. & Newsom, J. T. Preference for consistency: the development of a valid measure and the discovery of surprising behavioral implications. *J. Pers. Soc. Psychol.* **69**, 318–328 (1995).