Research Square

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

Belief uncertainty, unwillingness to trust, and excessive mentalizing underlie misplaced distrust in lonelier individuals

Gabriele Bellucci

gabriele.a.bellucci@gmail.com

Royal Holloway, University of London https://orcid.org/0000-0001-9050-4058

Patrick Suwandjieff

Institute of Neural Engineering, Graz University of Technology

Gernot R. Müller-Putz

Graz University of Technology https://orcid.org/0000-0002-0087-3720

René Riedl

University of Applied Sciences Upper Austria

Frank Krueger

George Mason University https://orcid.org/0000-0003-3420-2463

Article

Keywords:

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9948272/v1

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

Additional Declarations: There is **NO** Competing Interest.

Belief uncertainty, unwillingness to trust, and excessive mentalizing underlie misplaced distrust in lonelier individuals

Abstract

Loneliness is associated with disrupted social behaviors, yet the underlying mechanisms remain poorly understood. Understanding them requires disentangling the contributions of social beliefs, social preferences, and mentalizing abilities to social behaviors. Combining the investment game with computational modeling and neuroimaging, we investigated here how loneliness shapes reciprocity beliefs, willingness to trust and mentalizing across partners with different reciprocity traits. Behaviorally, lonelier individuals were more distrustful of partners, specifically of reciprocating ones, failing to settle on stable cooperative strategies even when partners responded generously. Computational modeling revealed that this pattern reflected reduced willingness to trust, excessive mentalizing, and persistently volatile, negatively-biased beliefs about partner reciprocity. Neuroimaging showed that greater ventromedial prefrontal cortex (VMPFC) activity was associated with greater willingness to trust, and integration of belief representations into trust valuations was linked to stronger coupling of VMPFC with the middle cingulate cortex. However, lonelier individuals with more negative expectations showed reduced VMPFC activity, especially with reciprocating partners, suggesting that loneliness distorts the neurocomputational mechanisms underpinning trusting interactions, leaving lonely individuals unable to capitalize on genuinely trustworthy others. This work provides a computational framework that offers mechanistic insights for understanding and targeting social dysfunction across healthy and clinical populations.

Loneliness is a subjective experience rooted in the perceived quality of one's social network and its ability to buffer life stressors that jeopardize physical and mental health. Loneliness carries health risks comparable to smoking and greater than physical inactivity, diet, and alcohol use[1], and has been linked to both psychiatric (e.g., psychosis) and somatic diseases (e.g., cardiovascular diseases)[2–4]. The scale of this challenge has drawn global attention, prompting the WHO's first report on loneliness countermeasures[5], and structural responses (e.g., social prescribing programs and community shed initiatives) across countries[6,7].

A central question for these policy efforts is what specific social behaviors maintain or deepen loneliness despite individuals' motivation to reconnect. Lonely individuals show a range of cognitive tendencies that could perpetuate social isolation, including greater attention to threatening interactions, heightened awareness of negative evaluations from others, stronger recall of negative social feedback, and negative social expectations[8–11]. These tendencies, together with more negative social attitudes[12,13], have been proposed to foster inflated beliefs about others' malevolent intentions[14–16], with downstream effects on social motivation and social functioning[17–19]. Yet, the link between these cognitive biases and social behaviors has remained surprisingly underspecified. Fully understanding this link requires distinguishing at least three conceptually separable layers: what lonely individuals believe about others (beliefs), how willing they are to act on those beliefs (preferences), and what they actually do (behavior). These layers are often conflated, yet they may be independently affected by loneliness and may interact in ways that explain seemingly contradictory findings in the literature.

Trust, defined as the willingness to accept vulnerability based on one's expectations of others' intentions and behaviors[20–22], provides a particularly informative lens for this decomposition. Trust decisions are grounded in expectations of others' benevolence that might diverge from their actual benevolence[23–25]. Moreover, despite past experiences of partner cooperation, individuals might consider the evidence for a partner's benevolent trait insufficient to trust[26,27]. At extreme levels, the fear of being let down can be extremely deleterious for social relationships[28]. Trust is therefore not merely the output of a rational cost-benefit calculation grounded in beliefs about a partner's cooperativeness, but it also depends on an individual's preference to rely on those beliefs.

Importantly, behavior can further diverge from one's beliefs and preferences. People often trust more than their expectations or preferences would suggest[23,24,29,30] and lonelier individuals have been shown to invest more than their negative expectations would warrant, with higher levels of loneliness predicting a stronger divergence between trusting behavior and

trustworthiness expectations[25,31]. Abnormal engagement in mentalizing, the ability to think about others' mental states[32–34], can produce the opposite pattern. For example, over-attribution of mental states (excessive mentalizing) has been linked to paranoid ideation, fostering suspicious behaviors responsible for the breakdown of trust[35,36]. Moreover, suspiciousness can lead to distrust beyond individual social preferences and beliefs about partner trustworthiness[37]. Accordingly, lonelier individuals with more paranoid-like thinking have been found selectively more distrustful of high reciprocity rates by others[38]. However, whether this reflects excessive mentalizing, negative beliefs, or reduced preference to rely on positive evidence of others' traits has not yet been directly tested[39]. Cases of misplaced distrust, in which a partner's actual trustworthiness warrants trust yet the individual withholds it, are particularly diagnostic, as they let emerge the decoupling between beliefs, preferences and behaviors that cannot be understood by examining any single layer in isolation.

Neural and computational evidence has mostly disregarded these layered dynamics, focusing on disrupted learning mechanisms underpinning lonely individuals' social inferences in brain areas like the medial prefrontal cortex (PFC) associated with social knowledge, social belief updating, and representations of others' personality traits[40–42]. In lonely individuals, atypical engagement of PFC and other dopaminergic regions has been reported in response to social cues[43,44]. In a learning task, lonely individuals integrated more information about negative behaviors of partners with a good reputation and recruited the medial PFC less when learning about them[39]. During repeated trust interactions, lonelier individuals showed reduced functional connectivity of the ventromedial PFC (VMPFC)[45], while stronger functional connectivity between VMPFC and temporoparietal junction (classically associated with social beliefs[46–48]) predicted more positive trustworthiness impressions of others[39]. However, how these belief representations shape decisions to trust or distrust in lonelier individuals remains unknown.

Moreover, previous computational work has mostly applied simple reinforcement learning models to capture basic learning dynamics in lonely individuals. This work has shown that lonelier individuals learn more from negative than positive information about others[39], akin to similar learning patterns in individuals with depression and anxiety[49,50]. Importantly, these basic learning models make assumptions that are often violated in real social interactions, where each agent's behavior depends on individual preferences and strategic inferences about other people's intentions. This is highly problematic, since the ability of computational models to satisfactorily reveal the failures of social neurocognitive processes in loneliness can at best be only as good as their ability to suitably capture the complex dynamics of social reasoning[31].

Here, we used a modified investment game[51] in which participants faced three trustee types differing in their tendency to reciprocate ("reciprocity trait"; **Fig. 1A-B**). To adopt modeling choices that better embed the existing social knowledge into computational models[52] and capture the relationships between beliefs, preferences and behavior central to the present study, we embedded a newly developed model of trust (i.e., vulnerability utility function) within an interactive partially observable Markov decision process (IPOMDP) framework[53,54]. This allowed us to pursue three main aims. First, our experimental design enabled us to test whether lonelier individuals selectively misplace their trust especially with trustworthy others (i.e., trustees with high reciprocity traits). Second, our computational approach allowed us to formally operationalize beliefs, preferences and mentalizing as separate components of the decision-making process to test their role in lonely individuals' trusting behaviors. Third, neuroimaging provided us with a mechanistic understanding of the interplay between these factors during trust decisions in lonelier individuals.

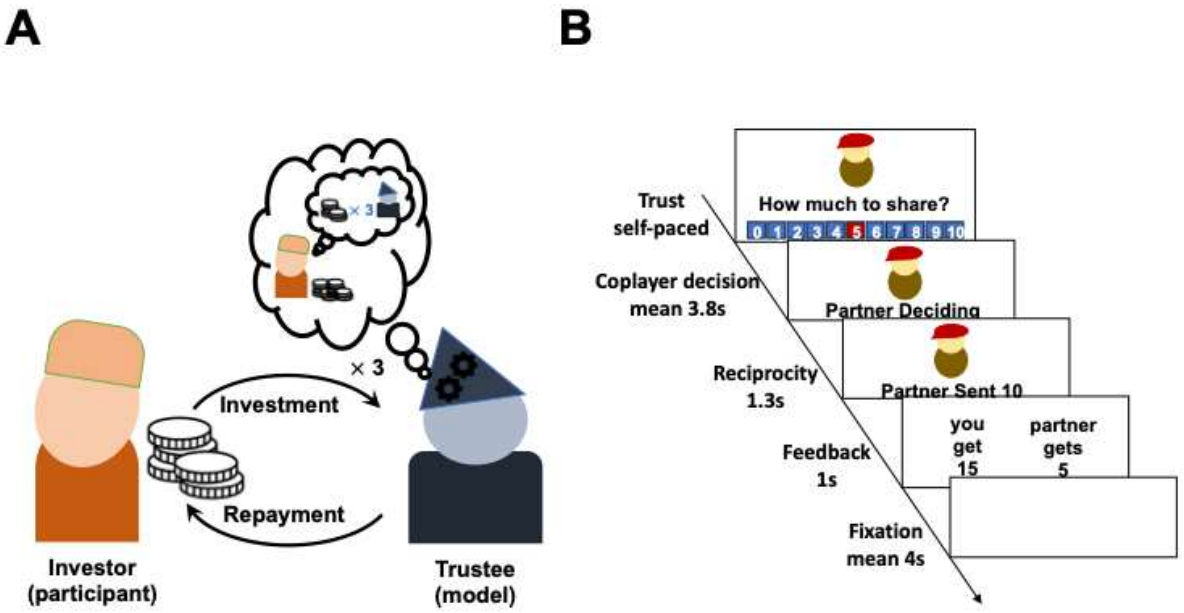


Fig. 1. Task and experimental design. **A.** Participants played the investment game with three different trustees whose "type" (reciprocity trait) was determined by our computational model (i.e., by the χ parameter where higher values mean greater tendency to reciprocate. Trustees behaved as IPOMDP agents that modeled participants' behaviors using an investor model (operationalized by the vulnerability function). Trustees held beliefs about participants' beliefs about themselves and updated these beliefs trial-by-trial via Bayesian updating. Participants' investments are classically considered as a measure of trust, while model repayments as a measure of trustee reciprocity. **B.** Experimental design in the MRI scanner: at the beginning of every trial (Trust), participants are matched with a different trustee represented by a unique avatar. They choose whether to share any of their initial endowment with the trustee. The investment is then tripled and passed on to the trustee. The participant waits until the trustee decides whether to repay any amount (Coplayer decision). Waiting times were based on an empirical distribution of real participants' times from previous experiments. Afterwards, the repayment is shown (Reciprocity) before the amount earned by both players is presented (Feedback). A variable jitter (2-8s with mean=4s) was presented between trials.

Results

The investment game is a two-player, sequential decision-making game where a player, the investor, receives an initial endowment and decides whether to share any of this amount with a second player, the trustee. The amount sent is multiplied (generally by three) and the trustee decides how much to repay (send back to) the investor (**Fig. 1A**). Participants' investments are classically taken to measure trust, while trustees' repayments to measure reciprocity[51]. In our study, each participant played the investment game as investor (**Fig. 1B**). Trustees' behavior was generated based on an IPOMDP framework that tracked participants' behavior and updated beliefs about participants' trusting behavior in a Bayesian fashion. IPOMDP trustees reciprocated based on belief-based simulations of participants' future trust decisions using a stochastic tree search approach (*Supplementary Methods*)[55]. IPOMDP trustees modeled participants' decisions using the vulnerability utility function for action value estimation, which has been proposed to capture fundamental psychological elements of trust[22]. IPOMDP trustees used these simulations of their partner's likely future behavior to estimate action values for reciprocity based on an equality-enforcing utility function and chose actions stochastically using a softmax policy (*Supplementary Methods*). Participants faced three different trustee types, which were implemented using utility functions with a χ parameter that made trustees progressively deviate from equality-enforcing behaviors the lower its value (with $\chi = [.3, .4, .5]$). This χ parameter can be thought of trustees' intrinsic propensity to reciprocate, their general "reciprocity trait", with lower values leading to lower reciprocity traits. Importantly, as trustees' behaviors further depended on their beliefs about participants' trust, trustees' gameplay and level of reciprocity was not deterministically set by their general reciprocity trait (χ parameter) but resulted from an interplay between χ and their beliefs, so that, in principle, an IPOMDP trustee with a low χ parameter (e.g., $\chi = .3$) playing with a trustworthy participant could manifest reciprocity levels as high (if not higher) as an IPOMDP trustee with a high (e.g., $\chi = .5$) χ parameter playing with an untrustworthy participant (see *Supplementary Methods*). We replicated our results in two experiments, an fMRI and an online experiment.

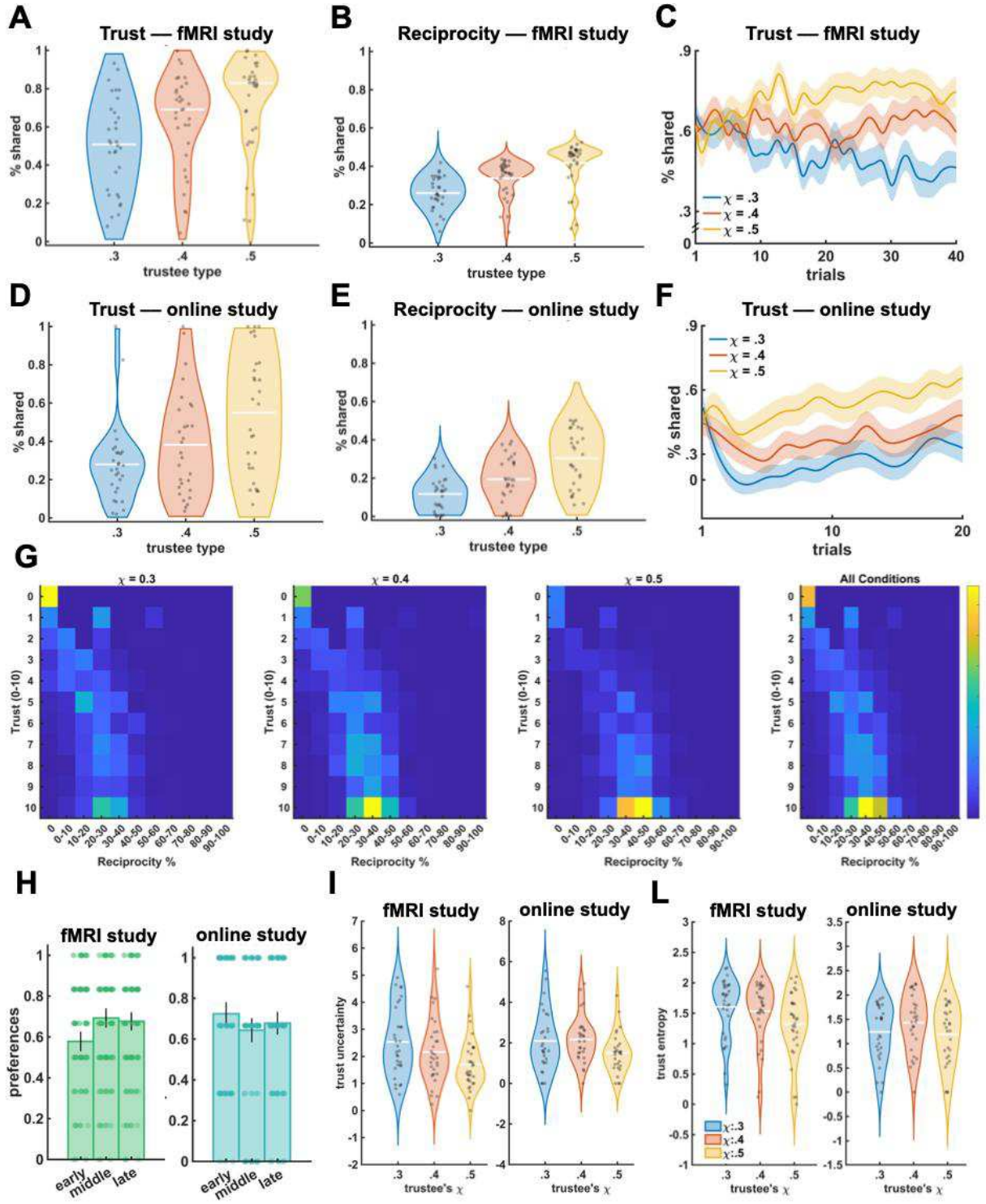


Fig. 2. Effects of trustee trustworthiness on trust. A-D. Mean percentage shared (trust) by participants in the investment game by trustee type in the fMRI (A) and online sample (D). **B-E.** Mean percentage returns (reciprocity) by IPOMDP trustees by trustee type in the fMRI (B) and online (E) sample. **C-F.** Mean trust in the three trustee types across participants over time in the fMRI (C) and online (F) sample. **G.** Participants' policies by partner responses measured as frequencies of unit of shared amount (0-10) across the three trustee types and overall (All Conditions). **H.** Average forced-choice preferences of trustees over three timepoints in the experiment. **I.** Empirical trust

uncertainty measured as the standard deviation of trusting behaviors manifested by participants for each of the three trustee types in both samples. **L.** Average trust entropy measured as the Shannon entropy of choosing a given trust level across trials for each of the three trustee types in both samples. Trustee types were $\chi = [.3, .4, .5]$, where lower values mean lower tendency to reciprocate. Error bars and shaded areas around lines represent standard error of the mean. Datapoints represent participants.

Trust strategies are sensitive to trustee types. We first characterized learning dynamics in the task using model-agnostic analyses (see *Supplementary Results*, *SR*, for all statistics). In both samples, participants were sensitive to trustees' tendency to reciprocate, linearly increasing their trust as a function of trustee reciprocity (*SR 1.1*; **Fig. 2A-D**). Participants' behavioral patterns closely reflected the observed reciprocity of trustees (*SR 1.2*; **Fig. 2B-E**), which in turn tracked participants' trust levels, as evidenced by strong associations between trust variability and reciprocity variability both at the individual and trustee-type level, with this relationship increasing in strength for more reciprocating trustees (*SR 1.2*). This reflected the sophistication of our model that not only generated behavior based on its hidden reciprocity trait (χ) but also on participants' empirical trust. Notably, this bidirectional coupling meant that some trustee types could "masquerade" as another. Reciprocating trustees ($\chi = .5$) in the online sample reciprocated at levels comparable to the unreciprocating trustees ($\chi = .3$) in the fMRI sample, simply because lower average trust in the online sample suppressed reciprocation (*SR 1.3*).

Further, empirical policies show that participants implemented distinct gameplays for each trustee type, with trust increasing over time selectively for more reciprocating trustees (*SR 1.4*; **Fig. 2C-F**). Specifically, with reciprocating trustees ($\chi = .5$), behavioral patterns converged toward a socially beneficial "all-half" strategy, whereby participants shared the full endowment and received half back from their trustees, yielding the highest achievable gains (**Fig. 2G**). Learning was also reflected in forced-choice trials, where partner preference for more reciprocating trustees exceeded chance early on (*SR 1.5*; **Fig. 2H**).

Finally, at the trial level, trusting behaviors manifested short history dependencies and strong autocorrelations, showing strong decreases of trust following low partner reciprocity in the preceding trial (*SR 1.6*). Trust was more consistent across trials with reciprocating trustees ($\chi = .5$), likely reflecting a stable cooperative strategy (see below). Interestingly, participants showed greater reactivity to positive reciprocity by unreciprocating trustees ($\chi = .3$), potentially reflecting coaxing attempts in response to unexpected cooperative signals (*SR 1.6*).

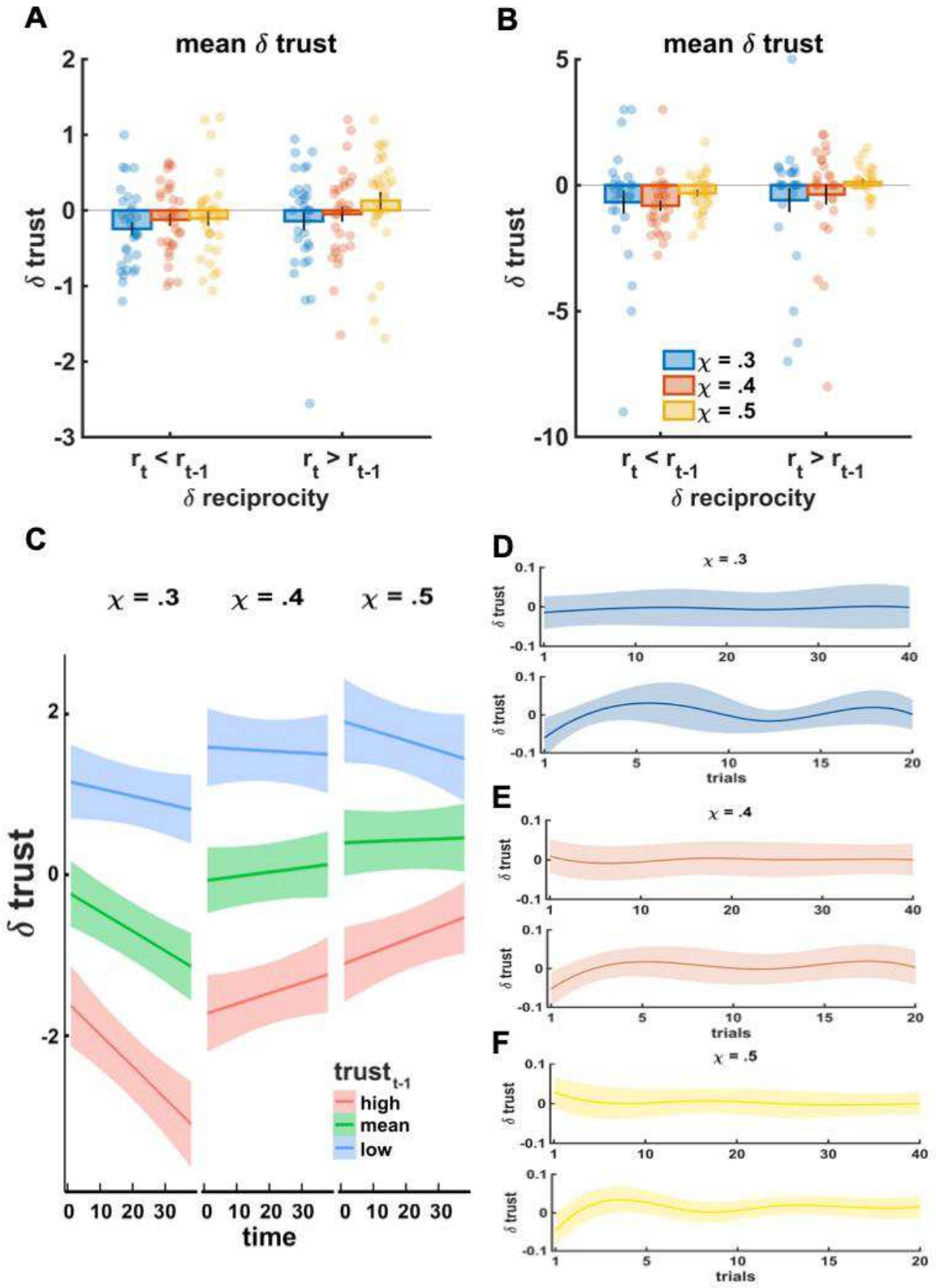


Figure 3. Sequential effects analysis. A. Mean trust changes (current minus previous trial investment) as a function of trial-by-trial changes in trustee reciprocity (previous minus 2-trial back reciprocity), separately for

each trustee type, in the fMRI sample. **B.** Same analysis as **A** for the online sample. **C.** Trust changes over time as a function of previous trust levels and trustee type, showing a steady drop in trust with tendentially unreciprocating trustees and stable or increasing trust with moderately and tendentially reciprocating trustees. **D-F.** Trialwise trust changes over time by trustee type in both samples, showing greater and increasing variability in interactions with tendentially unreciprocating trustees compared to more reciprocating ones. Error bars and shaded areas around lines represent standard error of the mean. Datapoints represent participants.

Greater reciprocity reduces trust variability. To gain further insights into participants' strategies and changes in trust adaptations when playing with trustees with different reciprocal tendencies beyond their average trialwise trust choices, we investigated the second moment of their behavioral patterns (i.e., standard deviation of trust choices), that is, their behavioral variability over time.

We observed greater behavioral variability in interactions with unreciprocating trustees ($\chi = .3$), with progressively lower variability for more reciprocating trustees (*SR 2.1*; **Fig. 2I**). In the fMRI sample, greater trust variability with reciprocating trustees ($\chi = .5$) was additionally driven by the greater variability in those trustees' reciprocal behaviors (*SR 2.2*). These results suggest that good partner traits foster not only others' trust but also more reliable (i.e., less variable) social behaviors.

Notably, trust variability patterns were comparable across samples despite different average trust levels ($BF_{10} = 0.32$, $P(H_1|D) = .24$), indicating they reflect deeper strategic dynamics. Importantly, there may be different reasons driving the observed variability patterns. For instance, greater trust variability could, on one hand, hinge on patterns of increasing trust over time, representing participants' learning about their partners' reciprocity trait. On the other hand, greater trust variability could instead signal reiterated attempts to coax unreciprocating trustees to reciprocate more. Greater variability in trust for unreciprocating trustees ($\chi = .3$) suggests the latter but standard deviation of trust choices alone does not give us insights into whether a participant is moving to a stable strategy with higher shared amounts in the course of the interaction or keeps sharing different amounts in the attempt to convince the partner to settle on a given behavioral strategy.

To disentangle whether greater trust variability reflected learning or coaxing strategies, we further estimated empirical choice entropy as the likelihood of manifesting a given level of trust with a certain trustee type over the course of the interaction. Under a learning account, participants should progressively move away from lower investments, resulting in lower empirical choice entropy. Under a coaxing account, they should instead continue varying their investments without settling on a stable strategy, resulting in higher empirical choice entropy. Results supported the learning account: participants became less likely to share lower investments with

more reciprocating trustees over time, and empirical choice entropy was lower for reciprocating than unreciprocating trustees ($\beta_{\chi=.5} = -0.25, 95\%\ CIs = [-0.42, -0.09]$; **Fig. 2L**). Together, these results suggest that less variable behavioral patterns signal the adaptation of more stable trust strategies over time with more reciprocating partners.

To further characterize how trust strategies adapted over time, we analyzed sequential effects in the game by using trust changes as a function of the previous trial's interaction, in line with the observed short history dependency of trust choices in our experiment and previous work[38,56]. We observed that trust adaptations were modulated by both observed trial-by-trial reciprocity changes and trustee type (*SR 2.3*; left panels of **Fig. 3A-B**). Participants more strongly updated their trust when interacting with more reciprocating trustees, particularly so when interacting with reciprocating ($\chi = .5$) trustees, and higher past trust levels predicted greater future trust changes specifically with these reciprocating trustees (*SR 2.3*).

Adding a time regressor revealed that in the fMRI sample, trust changes became progressively smaller in highly trusting interactions, and trustee type modulated this pattern, with a steady drop of trust toward unreciprocating trustees ($\chi = .3$) over time (*SR 2.4*; **Fig. 3C**). In the online sample, trust changes were weaker but more sustained across the interaction, with again larger changes occurring over time in highly trusting interactions (*SR 2.4*). This was further supported by analyses of variability of trust changes over time. While variability was similar across trustee types early in the experiment, it diverged significantly over time (*SR 2.5*). Specifically, variability decreased in interactions with reciprocating trustees ($\chi = .5$) but increased in interactions with unreciprocating ones ($\chi = .3$; **Fig. 3D-F**). Overall, these results point to unstable social strategies in interactions with unreciprocating others.

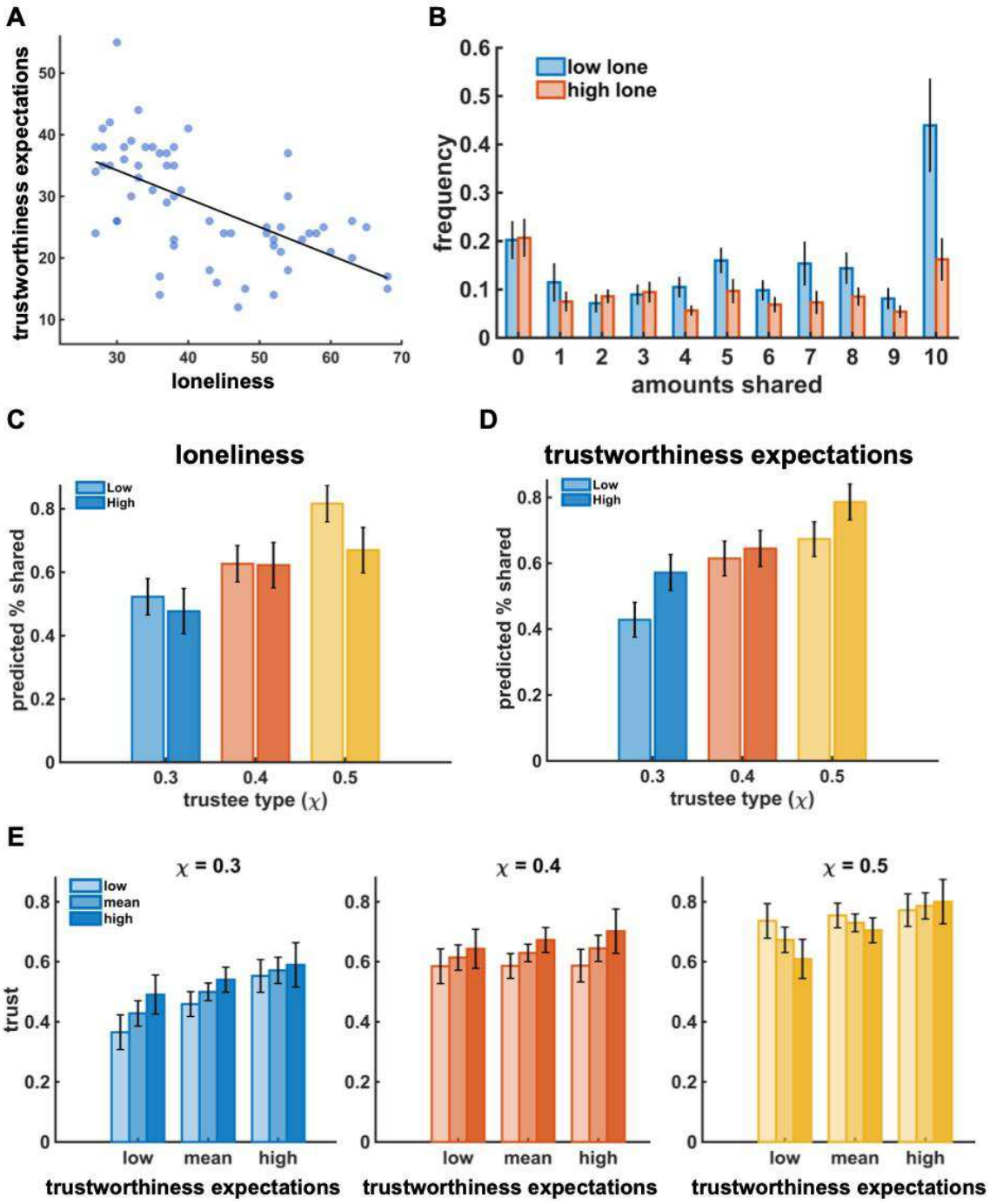


Fig. 4. Relationships of loneliness and general trustworthiness expectations with behavior. A. Correlation between individual loneliness scores and questionnaire-based trustworthiness expectations, showing that lonelier individuals had more negative expectations of others' trustworthiness. **B.** Lonelier individuals shared less overall with their trustees, especially higher amounts. **C.** An interaction between loneliness and trustee type shows that lonelier individuals distrusted tendentially reciprocating trustees significantly more than less reciprocating trustees. **D.** An interaction between trustworthiness expectations and trustee type shows that more negative trustworthiness expectations predicted greater distrust of less reciprocating trustees. **E.** Joint modulation of trust by loneliness and trustworthiness expectations showing that the effect of loneliness on distrust of tendentially reciprocating trustees was specifically driven by lonelier individuals with more negative trustworthiness expectations, whereas more

positive expectations partially offset the negative effects of loneliness. Error bars represent standard error of the mean. Datapoints represent participants.

Lonelier individuals distrust more reciprocating others. Once having understood the overall patterns of trusting behaviors in our participants, we can now turn to investigate how loneliness modulates them. We first tested whether lonelier individuals manifest behavioral strategies divergent from the observed patterns of trust in reciprocating trustees. Specifically, difficulties of lonelier individuals to establish successful trusting relationships may rely on their more negative beliefs about others (coming across as "too suspicious") or on their ambiguous behavior (signaling unreliability).

In line with previous work[25], we observed that lonelier individuals believed others to be less trustworthy ($r_{59} = -.60,\ BF_{10} = 5.45 \times 10^4,\ P(H_1|D) \cong 1$; **Fig. 4A**). Consequently, they trusted less both at the beginning of the interaction ($BF_{10} = 7.43,\ P(H_1|D) = .88$) and on average ($r_{59} = -.38, BF_{10} = 17.71,\ P(H_1|D) = .95$), and were more likely to choose lower investments overall ($\beta_{lone\times amount} = -0.02, 95\%\ CIs = [-0.02, -0.01]$; **Fig. 4B**). This contrasted with the behavioral patterns manifested by participants with more positive trustworthiness beliefs who showed overall greater levels of trust in their trustees ($r_{59} = .44,\ BF_{10} = 101.63,\ P(H_1|D) = .99$).

Importantly, though, lonelier individuals did not distrust all partners equally (**Fig. 4C-E**), but were more likely to specifically distrust reciprocating trustees ($\beta_{lone\times\chi=.5} = -0.11, 95\%\ CIs = [-0.16, -0.06]$), with trust in trustees with lower reciprocity traits ($\chi = .4$) slightly increasing ($\beta_{lone\times\chi=.4} = 0.05, 95\%\ CIs = [0.01, 0.10]$; all compared to $\chi = .3$). Again, these behavioral patterns are just the opposite to what we observed from individuals with more positive expectations of others' trustworthiness who while distrusting unreciprocating trustees ($\beta_{exp\times\chi=.4} = -0.13, 95\%\ CIs = [-0.18, -0.08]$), correctly placed trust in reciprocating ones ($\beta_{exp\times\chi=.5} = 0.03, 95\%\ CIs = [-0.02, 0.08]$; **Fig. 4D**). Importantly, the effect of loneliness on trust in reciprocating trustees was specifically driven by lonelier individuals' generally negative trustworthiness expectations ($\beta_{lone\times exp\times\chi=.4} = 0.23, 95\%\ CIs = [0.12, 0.34]$; $\beta_{lone\times pref\times\chi=.5} = 0.48, 95\%\ CIs = [0.38, 0.59]$), suggesting that more positive expectations might offset the negative effects of loneliness on social behaviors. Interestingly, lonelier individuals with more negative expectations trusted reciprocating trustees ($\chi = .5$) less than lonelier individuals with more positive expectations trusted unreciprocating ones ($\chi = .3$; **Fig. 4E**).

Lonelier individuals were also more distrustful of reciprocating trustees ($\chi = .5$) even when these trustees had reciprocated more in response to their previous distrust

($\beta_{lone\times\chi=.5\times rec_{t-1}\times trust_{t-1}} = 0.27, 95\%\ CIs = [0.04, 0.50]$). Consistently, lonelier individuals manifested higher empirical choice entropy when trusting more reciprocating trustees ($\beta_{lone\times\chi=.4} = 0.55, 95\%\ CIs = [0.33, 0.78]$; $\beta_{lone\times\chi=.5} = 1.23, 95\%\ CIs = [1.01, 1.45]$), as if treating them like unreciprocating partners. This effect was amplified in lonelier individuals with more positive trustworthiness expectations ($\beta_{lone\times exp\times\chi=.4} = 0.88, 95\%\ CIs = [0.57, 1.20]$; $\beta_{lone\times exp\times\chi=.5} = 1.61, 95\%\ CIs = [1.29, 1.92]$), potentially pointing to conflicting behavioral strategies in this group. Together, these results suggest that lonelier individuals had greater uncertainty about the trustworthiness of reciprocating partners and struggled more to settle on a stable strategy when interacting with them. Finally, none of these loneliness effects differed across samples ($\beta_{lone\times exp\times sample} = 0.19, 95\%\ CIs = [-0.20, 0.58]$).

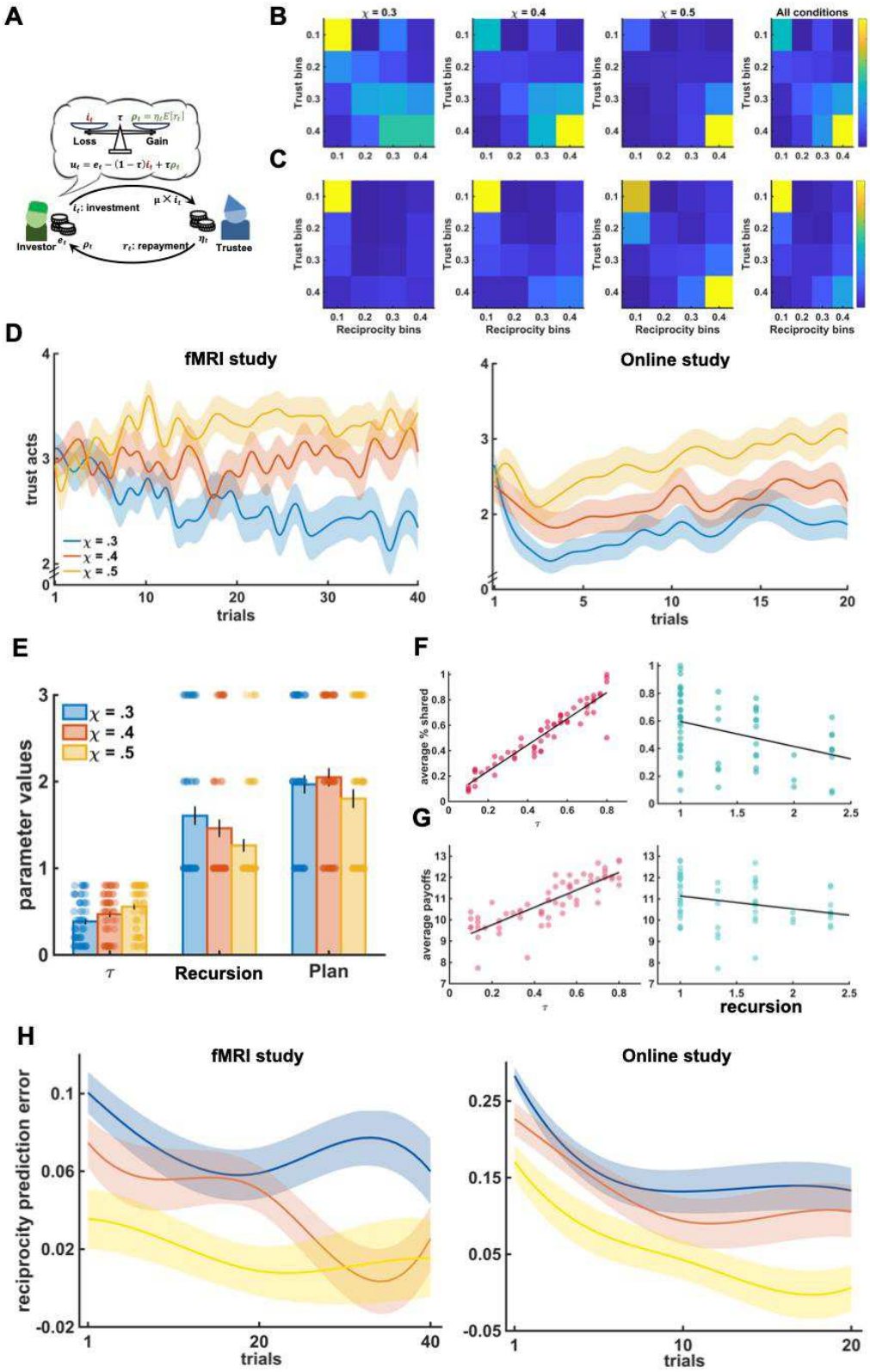


Figure 5. Modeling results. A. Schematic of the vulnerability utility function used to model participants' trust decisions, illustrating the tradeoff between the potential cost of trusting (lost investment s_t) and the expected benefit

(trustee reciprocity r_t), weighted by the individual willingness-to-trust (τ parameter). **B-C.** Action policies in the fMRI (B) and online sample (C) showing the frequency distribution of trust amounts across the three trustee types, demonstrating the shift toward higher investments for more reciprocating trustees. **D.** Trust trajectories over time by trustee type for the reduced action space. **E.** Model parameter estimates by trustee type: willingness to trust (τ) and recursions, and planning horizon (plan), showing significantly greater willingness to trust for more reciprocating trustees and more recursions for less reciprocating ones. **F.** Correlations of willingness to trust (left panel) and recursion depth (right panel) with average trust across participants. **G.** Correlations of willingness to trust (left panel) and recursion depth (right panel) with average individual payoffs. **H.** Reciprocity prediction errors (absolute difference between model-based expected and observed reciprocity) over time by trustee type, showing decreasing violations as participants learned trustees' reciprocity tendencies. Error bars and shaded areas represent standard error of the mean. Datapoints represent participants.

Computational modeling. The model-agnostic analyses above gave us insights into the overall effects of loneliness on trust in reciprocating partners. However, they still leave unanswered the question as to what mechanisms are driven lonelier individuals' distrust in reciprocating others and their uncertainty about these partners' trustworthiness. As mentioned above, these behavioral patterns can hinge on different mechanisms observed in lonely individuals in the current literature, such as their biased beliefs about others' reciprocity, their greater reluctance to rely on their partner, or their socio-cognitive abilities. Computational modeling allows us to characterize these processes separately and with a finer resolution by further probing their potential interactions. Moreover, contrary to regression analyses that limit us to only test sequential effects of closely consecutive trials, computational modeling enables us to investigate influences of those processes from the entire choice history.

To this end, we modeled participants as IPOMDP agents making Bayesian inferences about their partners' reciprocity and choosing whether to trust based on action utilities computed by the vulnerability function (i.e., the investor's internal model; **Fig. 5A**; see *Supplementary Methods*)[22]. Recursive thinking—a core property of mentalizing[53,57]—was operationalized as nested Bayesian inference that hierarchically embedded an agent's model of the partner and the partner's model of themselves. Following classic definitions of trust[20,21], the vulnerability function defines trust as a tradeoff between the potential negative consequences of trusting (i.e., losing the investment i_t) and the partner's positive behaviors (i.e., the trustee's repayment r_t). Individual willingness to trust (i.e., tendency to rely on one's expectations of another's behavior) is captured by a participant-specific parameter ($\tau \in [0,1]$) that operationalizes a belief-reliance mechanism and defines one's readiness to be vulnerable to a partner based on learned beliefs about the partner's behavior (here, reciprocity). Finally, partner reciprocity beliefs were represented by a probability distribution over the possible repayments based on the history of the interaction ($p(R_t = r_t|\mathcal{D}_{t-1})$, where $\mathcal{D}_{t-1}$ includes the history of repayments until time t). Trust action utilities are then so computed:

$$u_t(i_t, \mathcal{D}_{t-1}) = e_t - (1-\tau)s_t + \tau\left(\eta_t \int_0^1 r_t \, p_t(r_t|\mathcal{D}_{t-1})dr\right).$$

For tractability, we reduced the action and parameter spaces our IPOMDP agents made inferences about. Importantly, though, the reduced action space (binned trust "acts") maintained all the features of trust dynamics observed in the raw data both in terms of action policies (**Fig. 5B-C**) and trust changes over time across trustee types (**Fig. 5D**). We also fitted the IPOMDP using the inequality-aversion function from the literature as internal model of the investor[58,59]. Model comparisons show that the IPOMDP approach outperformed non-IPOMDP ones, and that the vulnerability function as investor model outperformed the inequity-aversion function irrespective of the approach used (**Fig. S1**). Moreover, models with biased belief updating and discounting mechanisms turned out to fit the data worse than models without (**Tab. S1**; *Supplementary Methods*). IPOMDP models were fitted using a stochastic POMCP procedures that revealed good parameter estimate consistency for the winning model (**Fig. S2**; *Supplementary Methods*). Parameter recovery and behavioral correspondence of model simulations using the winning model were also good (**Fig. S3-4**).

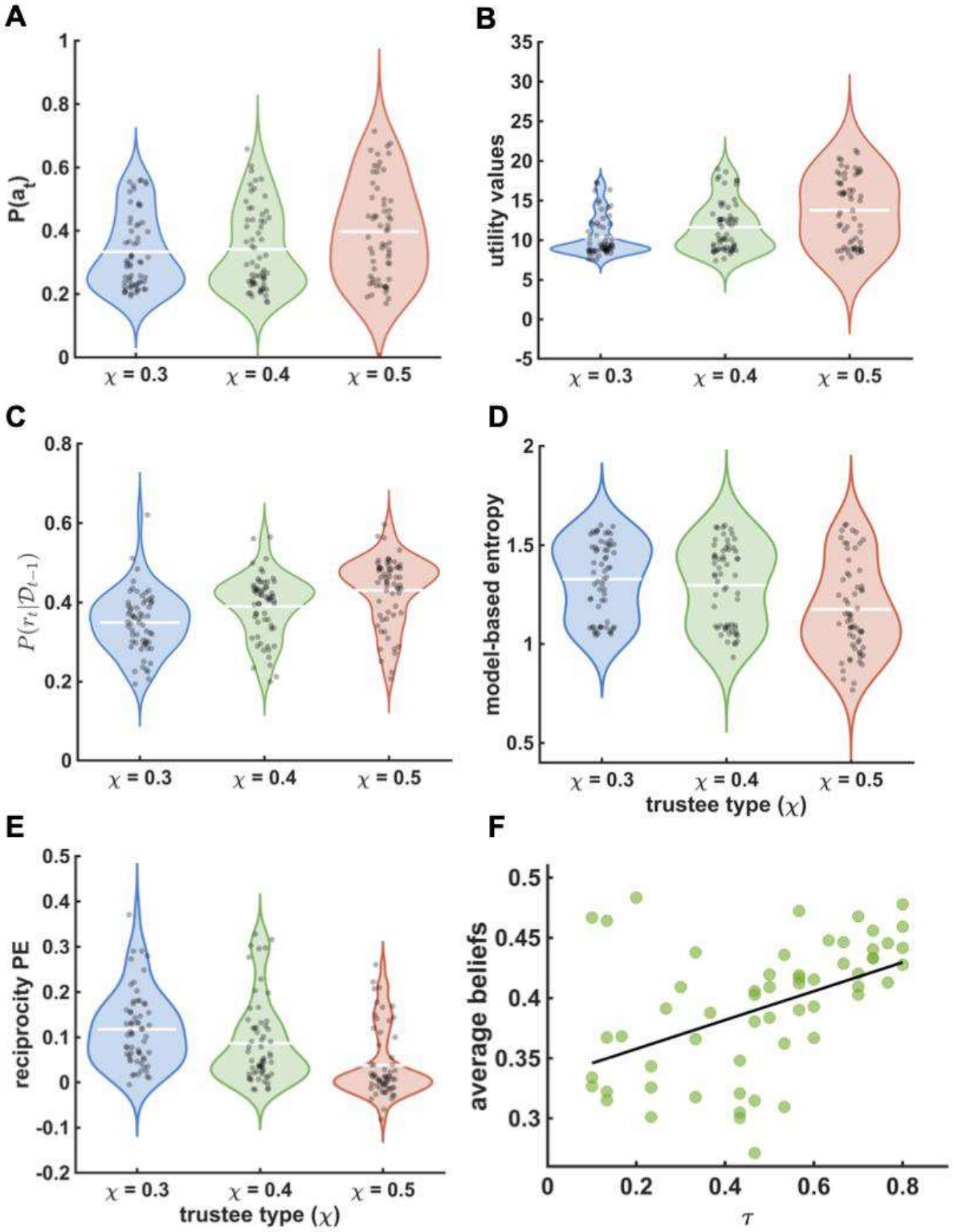


Figure 6. Parameter variables. A. Model-based policies (action probabilities) by trustee type, showing more deterministic policies for more reciprocating trustees. **B.** Model-based action utilities by trustee type. **C.** Participants' model-based beliefs about trustee reciprocity by trustee type, capturing trustees' intrinsic tendency to reciprocate. **D.** Model-based action uncertainty (Shannon entropy of the trialwise action probability distributions averaged by

trustee type) shows lower uncertainty for more reciprocating trustees. **E.** Prediction errors (PE) about partner behavior (reciprocity) computed as the difference between partner reciprocity beliefs and observed reciprocity were lower for more reciprocating trustees. **F.** Correlation between average model-based beliefs and willingness to trust (τ) across participants, showing that greater willingness to trust was associated with more positive beliefs about trustee reciprocity. Error bars represent standard error of the mean. Datapoints represent participants.

Reciprocity increases willingness to trust and reduces recursive thinking. We first investigated the general patterns of the above-mentioned three mechanisms (i.e., willingness to trust, mentalizing and partner beliefs) and their contributions to overall trusting behaviors. Participants' willingness to trust (i.e., τ parameter) changed significantly as a function of trustee's reciprocity traits ($BF_{10} = 6.00 \times 10^4$, $P(H_1|D) \cong 1$), with greater willingness to trust for more reciprocating trustees ($\chi = .3$: $M = 0.39, SD = 0.25$; $\chi = .4$: $M = 0.47, SD = 0.26$; $\chi = .5$: $M = 0.56, SD = 0.26$; **Fig. 5E**). Post-hoc t-tests revealed strong evidence for significant differences of all pairwise comparisons (BF_{10} ranging from 5.73 to 2.17×10^4). Moreover, participants engaged in shallower recursive thinking (as indexed by number of recursions) in interactions with more reciprocating trustees ($\beta_{trustee} = -0.17, 95\%\ CIs = [-0.29, -0.05]$; $\chi = .3$: $M = 1.61, SD = 0.84$; $\chi = .4$: $M = 1.46, SD = 0.79$; $\chi = .5$: $M = 1.26, SD = 0.58$; **Fig. 5E**), while no evidence was found for significantly different planning horizons across trustee types ($BF_{10} = 0.24$, $P(H_1|D) = .19$; **Fig. 5E**).

To characterize the specific contributions of these computational measures to participants' trusting behaviors, we investigated their associations with behavioral outcomes. We observed that on average, greater willingness to trust led to higher levels of trust ($r_{59} = .95$, $BF_{10} = 7.43 \times 10^{25}$, $P(H_1|D) \cong 1$; **Fig. 5F**, left panel) and more successful interactions as measured by individual average payoffs ($r_{59} = .82$, $BF_{10} = 1.99 \times 10^{12}$, $P(H_1|D) \cong 1$; **Fig. 5G**, left panel), while deeper recursive thinking was associated with lower levels of trust ($\varrho_{59} = -.40$, $BF_{10} = 23.99$, $P(H_1|D) = .96$; **Fig. 5F**, right panel) and less successful interactions ($\varrho_{59} = -.34$, $BF_{10} = 2.70$, $P(H_1|D) = .73$; **Fig. 5G**, right panel).

Moreover, model-based action probabilities ($BF_{10} = 12.59$, $P(H_1|D) = .93$; **Fig. 6A**) and action utilities ($BF_{10} = 1.34 \times 10^{10}$, $P(H_1|D) \cong 1$; **Fig. 6B**) were significantly higher for more reciprocating trustees, suggesting that participants' policies were more deterministic in interactions with reciprocating others, corroborating the model-agnostic evidence on more stable strategies with these trustees. On average, model-based beliefs about partner reciprocity were more positive for more reciprocating trustees, consistent with trustees' latent reciprocity traits (**Fig. 6C**; $BF_{10} = 2.89 \times 10^7$, $P(H_1|D) \cong 1$). Importantly, lower model-based action uncertainty (i.e., entropy: **Fig. 6D**; $BF_{10} = 1.11 \times 10^3$, $P(H_1|D) \cong 1$) and less reciprocity prediction errors (**Fig. 6E**; $BF_{10} = 6.73 \times 10^7$, $P(H_1|D) \cong 1$) for more reciprocating trustees indicate greater decision

certainty and more accurate reciprocity beliefs for reciprocating partners.

Importantly, knowing that someone is trustworthy not only triggers greater trust but also fosters an overall willingness to view others as benevolent and reliable partners. Consequently, in our task, more positive beliefs about partner reciprocity strongly correlated with greater willingness to trust (**Fig. 6F**; $r_{59} = .42, BF_{10} = 47.71, P(H_1|D) = .98$), and participants were more likely to update their beliefs upward about the reciprocity of trustees with more positive reciprocity traits (as measured by Kullback-Leibler divergence of model-based beliefs across trials: $\beta_{\chi=.5} = 0.09, 95\%\ CIs = [0.04, 0.14]$). This meant that while over time reciprocity prediction errors decreased as participants learned partner reciprocity traits ($\beta_{time} = -0.23, 95\%\ CIs = [-0.33, -0.13]$), reciprocity predictions of more reciprocating trustees were more likely to be more accurate from the outset and thus changed less strongly as compared to those of unreciprocating ones ($\beta_{\chi=.5\times time} = 0.20, 95\%\ CIs = [0.08, 0.32]$; **Fig. 5H**). Consequently, action uncertainty also decreased over time for reciprocating trustees ($\beta_{\chi=.5\times time} = -0.18, 95\%\ CIs = [-0.27, -0.09]$). In line with the above trust variability results, these findings point to more stable and deterministic trust strategies with more reciprocating partners and more uncertain behavioral patterns with less reciprocating ones. Such uncertainty might have triggered greater belief inference demands, as suggested by deeper recursive thinking in interactions with unreciprocating partners (**Fig. 5E**).

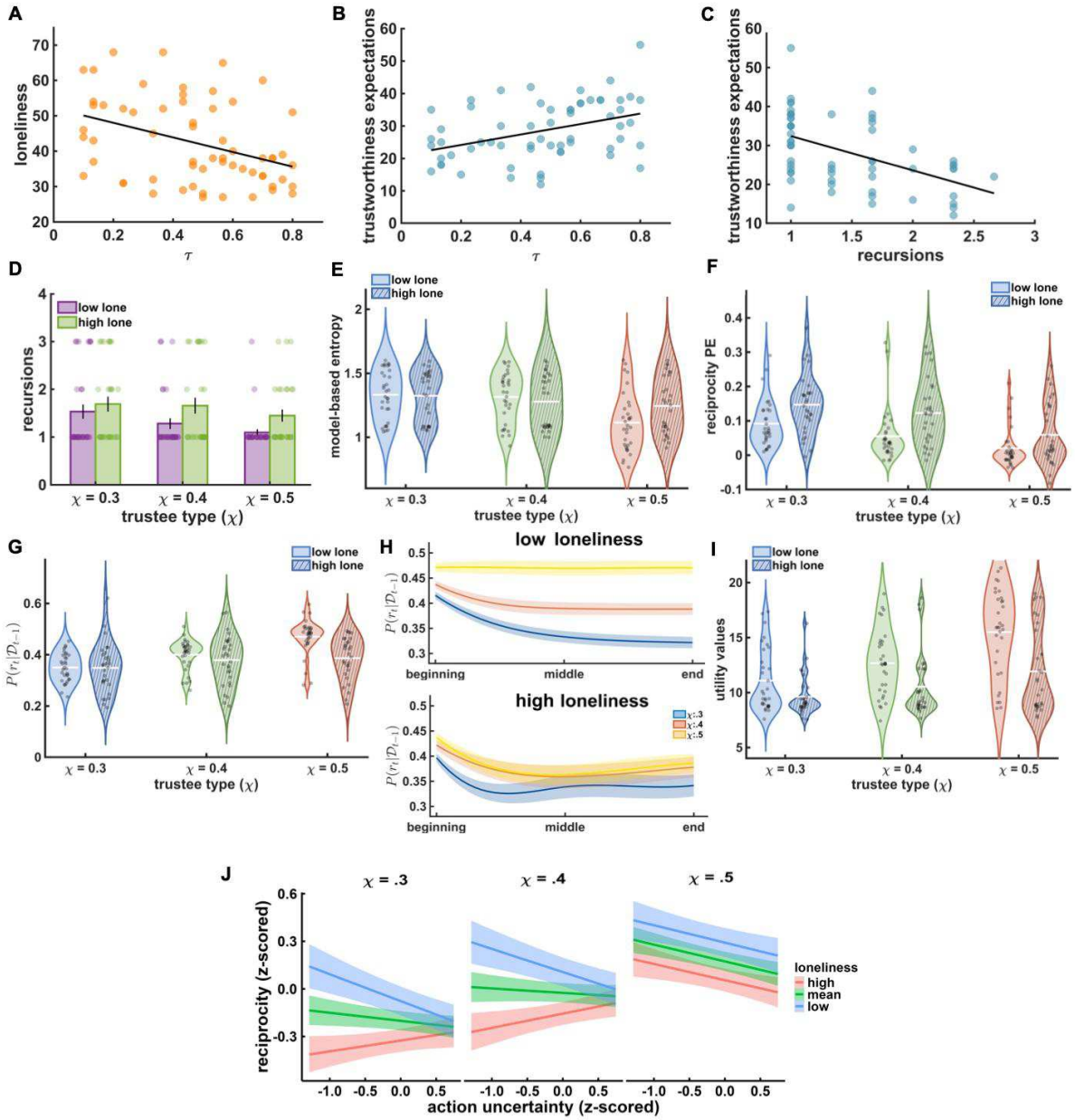


Figure 7. Loneliness and model dynamics. A. Correlation between individual loneliness scores and willingness to trust (τ parameter), showing that lonelier individuals were less willing to trust. **B.** Correlation between questionnaire-based trustworthiness expectations and willingness to trust (τ), showing that more positive trustworthiness expectations were associated with greater willingness to trust. **C.** Correlation between trustworthiness expectations and recursions, showing that more positive trustworthiness expectations predicted less recursions. **D.** Recursions as a function of loneliness and trustee type, showing that lonelier individuals engaged in more recursions especially when interacting with reciprocating trustees. **E.** Lonelier individuals had greater model uncertainty (i.e., entropy) for more reciprocating trustees. **F.** Prediction errors (PE) computed as the difference between partner reciprocity beliefs and observed reciprocity were greater in lonelier individuals, showing that their beliefs were more inaccurate about their partners. **G.** Lonelier individuals had more negative beliefs about their partners, specifically reciprocating trustees. **H.** This was due to lonelier individuals not sufficiently distinguishing between trustees with different reciprocity traits. **I.** Lonelier individuals had increasingly lower trust utilities for more reciprocating trustees. **J.** Lonely individuals' action uncertainty elicited lower reciprocity, especially from more reciprocating trustees. Error bars and shaded areas represent standard error of the mean. Datapoints represent participants.

Loneliness is associated with more uncertainty in beliefs and behaviors. Model-agnostic analyses suggested that lower trust in lonelier individuals may stem from heightened cautiousness when interacting with trustworthy others. We hypothesized that this reflects greater suspiciousness about partners' trustworthiness, which may be associated with more uncertain beliefs about others' cooperative intentions and a tendency to discount particularly positive beliefs. Our computational model allows us to test these three mechanisms directly through the recursion parameter, whose higher values point to excessive mentalizing (i.e., an operationalization of suspiciousness), model-based beliefs about partner reciprocity (estimated trial-by-trial via Bayesian inference), and the τ parameter, which captures a belief-reliance mechanism (i.e., the degree to which a decision-maker relies on their beliefs about the partner's trustworthiness).

We observed that lonelier individuals had lower τ values ($r_{59} = -.38$, $BF_{10} = 20.99$, $P(H_1|D) = .96$; **Fig. 7A**) and showed deeper recursive thinking especially for more reciprocating trustees ($\beta_{lone\times\chi=.5} = 0.65, 95\%\ CIs = [0.01, 1.27]$; **Fig. 7D**). The opposite pattern emerged in individuals with generally more positive trustworthiness expectations about others, who showed higher τ values ($r_{59} = .39$, $BF_{10} = 26.18$, $P(H_1|D) = .96$; **Fig. 7B**) and shallower recursive thinking ($r_{59} = -.50$, $BF_{10} = 7.14 \times 10^2$, $P(H_1|D) \cong 1$; **Fig. 7C**). Model-based beliefs about the reciprocity of reciprocating trustees were more volatile in lonelier individuals ($\beta_{lone\times\chi=.5} = 0.41, 95\%\ CIs = [0.31, 0.51]$), particularly in those with generally more negative trustworthiness expectations ($\beta_{lone\times exp\times time\times\chi=.5} = -0.75, 95\%\ CIs = [-1.32, -0.18]$). This was accompanied by higher belief uncertainty about reciprocating trustees in lonelier individuals ($\beta_{lone\times\chi=.5} = 0.29, 95\%\ CIs = [0.21, 0.38]$; **Fig. 7E**), again amplified in those with more negative trustworthiness expectations ($\beta_{lone\times exp\times\chi=.4} = -0.21, 95\%\ CIs = [-0.40, -0.02]$; $\beta_{lone\times exp\times\chi=.5} = -0.52, 95\%\ CIs = [-0.73, -0.32]$). Importantly, while reciprocity prediction errors decreased over time as participants learned trustees' reciprocity traits ($\beta_{time} = -0.21, 95\%\ CIs = [-0.32, -0.10]$), they remained persistently high in lonelier individuals throughout the interaction ($\beta_{lone\times time\times\chi=.4} = 0.29, 95\%\ CIs = [0.03, 0.55]$; $\beta_{lone\times time\times\chi=.5} = 0.25, 95\%\ CIs = [0.02, 0.48]$; **Fig. 7F**), especially in those with generally more negative trustworthiness expectations ($\beta_{lone\times exp\times time\times\chi=.5} = 0.65, 95\%\ CIs = [0.07, 1.23]$).

These results suggest that lonelier individuals might have been consistently poor at predicting their partner reciprocity over time, possibly because their beliefs were less accurate than those of less lonely participants. Indeed, model-based beliefs about more reciprocating

trustees were more negative in lonelier individuals ($\beta_{lone\times\chi=.4} = -0.36, 95\%\ CIs = [-0.43, -0.29]$;$\beta_{lone\times\chi=.5} = -0.42, 95\%\ CIs = [-0.50, -0.35]$; **Fig. 7G**), particularly in those with more negative trustworthiness expectations ($\beta_{lone\times exp\times time} = 0.51, 95\%\ CIs = [0.17, 0.85]$). Crucially, beliefs about reciprocating trustees became increasingly more negative over time in lonelier individuals ($\beta_{lone\times time\times\chi=.5} = -0.73, 95\%\ CIs = [-0.89, -0.57]$), such that they ultimately held similarly negative beliefs about all trustees regardless of trustees' latent reciprocity trait (**Fig. 7H**).

Importantly, action utilities were lower in lonelier individuals ($\beta_{lone} = -0.17, 95\%\ CIs = [-0.34, -0.001]$) and while in general, action utilities increased over time for more reciprocating trustees (indicative of learning; $\beta_{time\times\chi=.4} = 0.17, 95\%\ CIs = [0.08, 0.26]$; ($\beta_{time\times\chi=.5} = 0.12, 95\%\ CIs = [0.04, 0.21]$), lonelier individuals' action utilities became increasingly smaller over time especially for reciprocating trustees ($\beta_{lone\times time\times\chi=.5} = -0.32, 95\%\ CIs = [-0.46, -0.18]$; **Fig. 7I**). Consequently, action policies also became more random over time in lonelier individuals ($\beta_{lone\times time} = -0.33, 95\%\ CIs = [-0.50, -0.17]$). Since trustees' actual reciprocity also depended on participants' own behavior, lonelier individuals' more ambiguous trust signals may have elicited less reciprocation from their partners. Indeed, higher action uncertainty elicited overall lower reciprocity ($\beta_{eH} = -0.21, 95\%\ CIs = [-0.25, -0.17]$). In particular, lonelier individuals not only received overall lower reciprocity ($\beta_{lone} = -0.25, 95\%\ CIs = [-0.36, -0.15]$), but reciprocity was particularly lower when they manifested higher action uncertainty with reciprocating trustees ($\beta_{lone\times eH\times\chi=.5} = -0.23, 95\%\ CIs = [-0.34, -0.12]$; **Fig. 7J**). On the contrary, lonelier individuals' higher action uncertainty led to more reciprocity in interactions with unreciprocating trustees, suggesting that their more unstable behavioral strategy reflected continuous probing of these partners, which enabled them to receive higher returns from those partners than the ones less lonely individuals were able to get by withdrawing from trust altogether (**Fig. 7J**). Hence, quickly settling on a behavioral strategy might prevent less cooperative partners from the chance to cooperate (as likely happened to less lonely individuals), while not doing so jeopardizes the development of a fruitful relationship with cooperative ones (as likely happened to lonelier individuals). Together, these results suggest that lonely individuals' more ambiguous behaviors and more negative social beliefs combine to set off a self-reinforcing cycle of distrust with trustworthy partners.

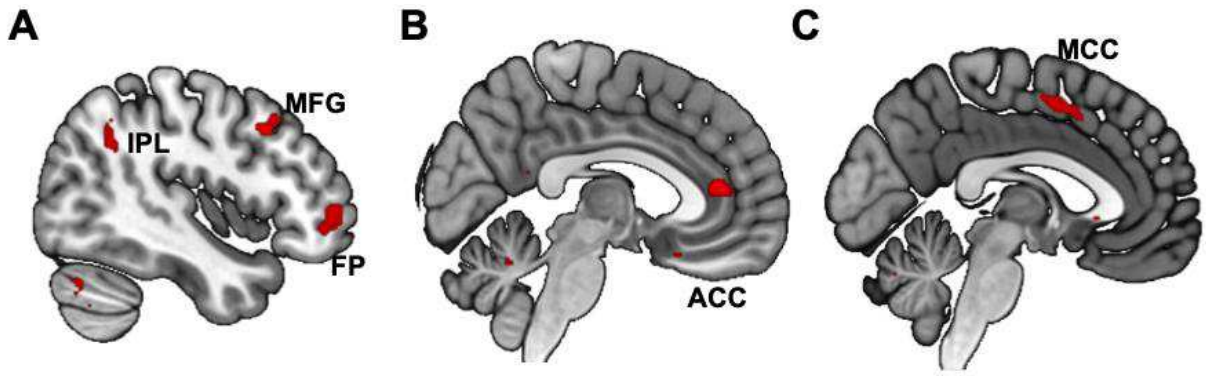


Figure 8. Trustee types modulate trust-related activity. A. During trust decisions, interactions with reciprocating trustees ($\chi = .5$) elicited stronger neural activations in the dorsolateral prefrontal cortex (middle frontal gyrus, MFG), parietal cortex (inferior parietal lobe, IPL), and ventrolateral prefrontal cortex (frontal pole, FP) as opposed to unreciprocating trustees ($\chi = .3$). **B.** The anterior cingulate cortex (ACC), on the contrary, was more strongly engaged by interactions with reciprocating trustees as compared with moderately reciprocating trustees ($\chi = .4$). **C.** Waiting reciprocal decisions of less reciprocating trustees ($\chi = .3$ and $\chi = .4$) elicited stronger activations in the middle cingulate cortex (MCC) as compared to reciprocating trustees ($\chi = .5$).

Lonelier individuals show reduced engagement of the neural representations of trust. We then sought to determine the neural correlates associated with trustee types, decisions to trust, and individual willingness to trust (τ parameter). We found that interactions with different trustee types elicited activations in various brain regions (**Tab. S2**). Most notably, we observed that interactions with reciprocating trustees ($\chi = .5$) elicited stronger activations associated with trust decisions (e.g., in the ventrolateral and dorsolateral prefrontal cortex, parietal cortex, AI and ACC; **Fig. 8A-B & Tab. S2**) and partner returns (e.g., precuneus and angular gyrus; see **Tab. S2**). On the contrary, waiting for a reciprocity decision by less reciprocating trustees ($\chi = .3$ and $\chi = .4$) more strongly activated the MCC as compared to reciprocating trustees ($\chi = .5$; **Fig. 8C & Tab. S2**).

Moreover, PCC and VMPFC were parametrically modulated by the amount invested during trust decisions (**Fig. 9A & Tab. S3**), while bilateral caudate activations were modulated by higher payoffs (**Fig. 9B & Tab. S4**). Finally, at the subject-level, greater willingness to trust (higher τ values) were associated with MCC activity during trustee reciprocity (**Fig. 9C & Tab. S5**), and with neural responses to personal gains in MCC, VMPFC, IPL, and temporal cortex (**Fig. 9D & Tab. S5**).

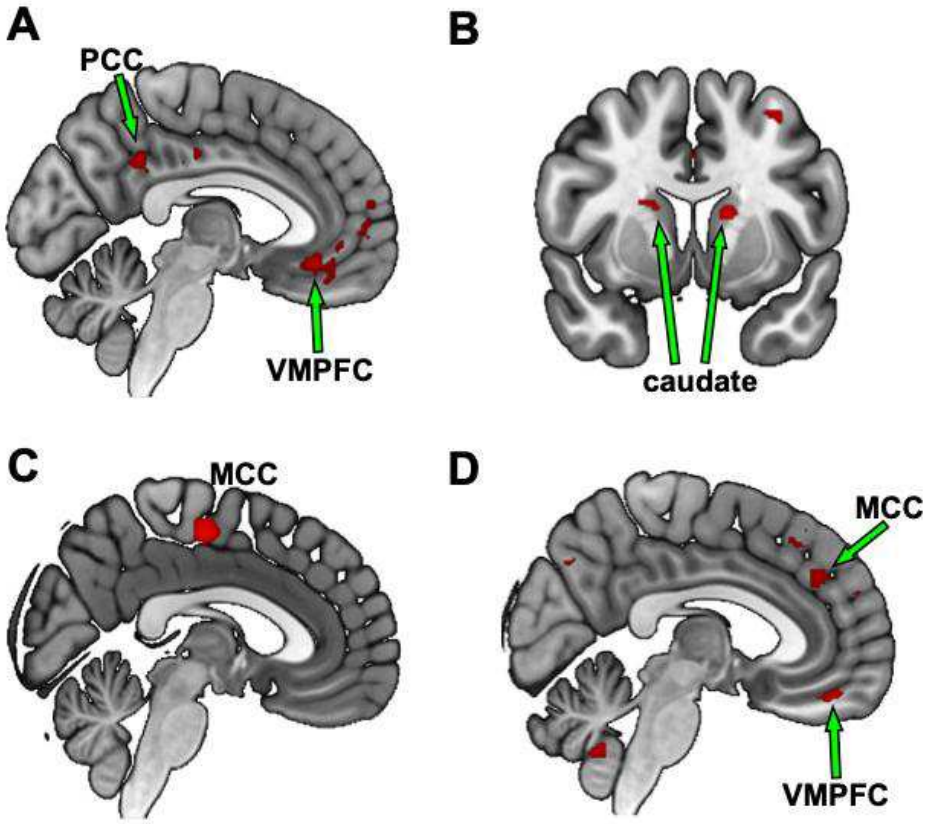


Figure 9. Neural correlates of trust and personal gains. A. The ventromedial prefrontal cortex (VMPFC) and posterior cingulate cortex (PCC) were parametrically associated with greater trialwise trust across trustees. **B.** The bilateral caudate was parametrically associated with greater trialwise personal gains during feedback. **C.** Model-based willingness to trust (τ parameter) was associated with the middle cingulate cortex (MCC) during the trustees' decisions. **D.** Willingness to trust was further associated with neural responses to personal gains in the VMPFC and MCC during feedback.

Given that the activity in the VMPFC (previously involved in both inferences about self/others' traits and loneliness[39,45,60]) was associated with both greater trust (as measured by trialwise investments) and willingness to trust (as measured by the τ parameter), we more closely investigated the role of this brain region in loneliness during trust decisions in our study (**Fig. 10A**). We observed that lonelier individuals with generally more negative trustworthiness expectations engaged the VMPFC significantly less with reciprocating trustees but more with unreciprocating ones ($\beta_{lone\times exp\times \chi=.5} = 0.36, 95\%\ CIs = [0.04, 0.67]$; **Fig. 10B**). Given that lonelier individuals also formed more negative beliefs about partner reciprocity during the game and showed more variable trusting behaviors, we analyzed VMPFC functional pathways to investigate whether trust value signals in this region might be differentially modulated by behaviorally-relevant belief representations. A whole-brain functional connectivity analysis revealed that during trust decisions, VMPFC was more strongly coupled with an MCC region (**Fig. 10C**) that was also more

activated when participants interacted with less reciprocating trustees (**Fig. 8C**).

Specifically, stronger coupling between these two brain regions was associated with more random behavior when beliefs about partner reciprocity were negative, but with more deterministic behavior when beliefs were positive ($\beta_{bel\times pol} = 0.36, 95\%\ CIs = [0.08, 0.65]$; **Fig. 10D**). These results suggest that when negative beliefs about partner reciprocity represented in the MCC are integrated into trust values in the VMPFC, social strategies become less deterministic (possibly signaling probing or coaxing attempts), whereas the integration of more positive beliefs helps individuals settle on a more predictable course of actions. Given its central role in this orchestrating mechanism, lonelier individuals' reduced VMPFC engagement when interacting with reciprocating partners may compromise their ability to both effectively probe the partner's intentions and reliably maintain an unambiguous trust policy, with potentially detrimental consequences for the formation of successful social relationships.

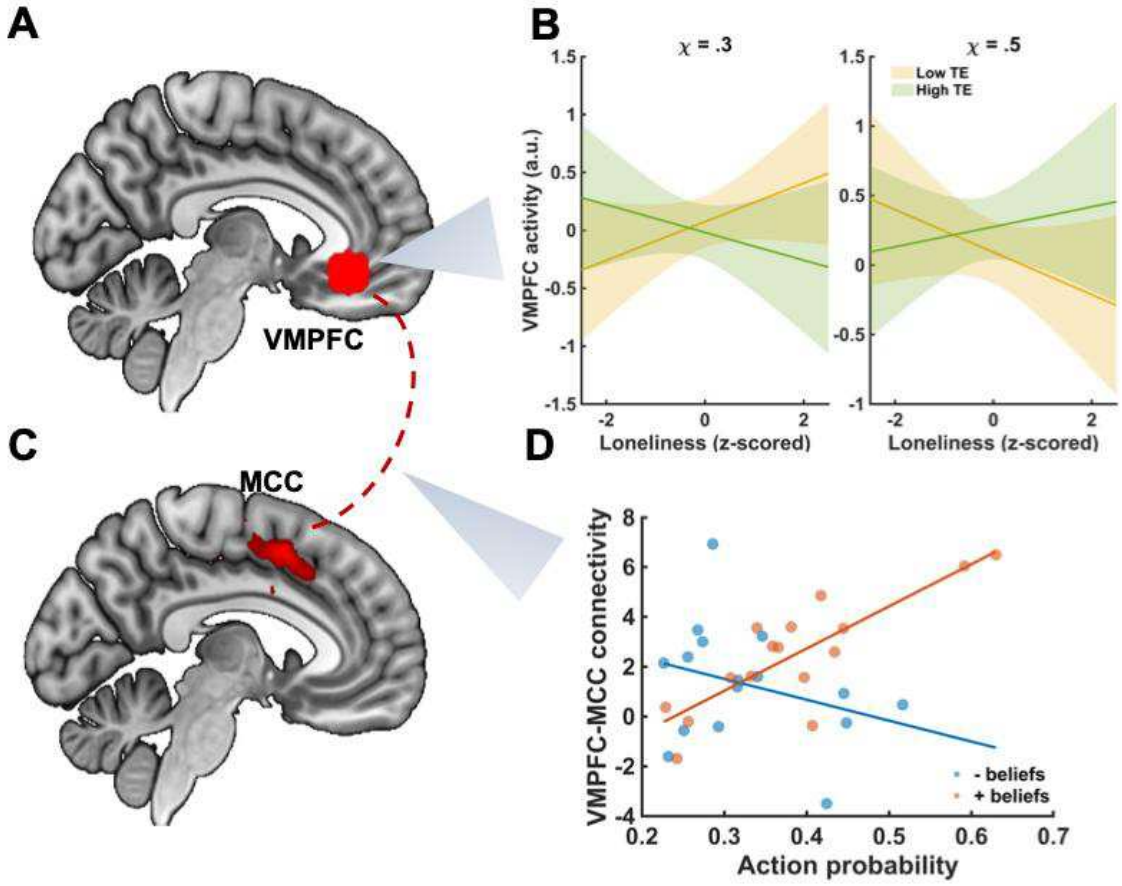


Figure 10. Integration of trust values and belief representations in lonelier individuals. A-B. The ventromedial prefrontal cortex (VMPFC; A) was less strongly activated in lonelier individuals with more negative trustworthiness expectations when interacting with reciprocating trustees ($\chi = .5$; B). **C-D.** Further, the VMPFC was functionally

coupled with the middle cingulate cortex (MCC: x = -6, y = 8, z = 50) during decisions to trust (C) and stronger coupling between these two brain regions was associated with more random policies when beliefs were negative but more deterministic ones when beliefs were positive (D).

Discussion

Using a modified investment game with trustees varying in reciprocity, we showed that lonelier individuals held more negative beliefs about others and trusted less, particularly reciprocating partners, failing to settle on cooperative strategies even when these partners responded generously. Computational modeling revealed that this reflected deeper recursive thinking and a reduced reliance on beliefs about partner reciprocity. Neuroimaging further revealed that lonelier individuals engaged the VMPFC less with reciprocating partners, likely impairing the integration of beliefs into trust valuations needed to capitalize on genuinely trustworthy relationships.

Behavioral analyses demonstrated that lonelier individuals showed increased action uncertainty across different types of trustees, suggesting that loneliness is associated with greater difficulty in determining appropriate trust responses. This finding extends previous research on loneliness and social cognition[61,62] by providing evidence that lonely individuals experience heightened uncertainty in their social decision-making processes, leading to ambiguous behavioral strategies. Moreover, as in previous studies[25,39], expectations of others' trustworthiness were negatively associated with individual loneliness, and interacted with partner reciprocity traits in predicting individual trust choices. Specifically, more positive expectations likely compensated against the negative effects of loneliness, promoting more trusting behaviors. This could explain the different effects of temporary and chronic loneliness on social behaviors[63–65]. In particular, temporary feelings of loneliness might act as an evolutionary-built warning signal about inadequate social network[65], as they might leverage one's generally positive expectations of others to foster affiliative behaviors that could lead to the establishment of better social relationships. However, when feelings of loneliness persist long enough due to repeatedly unsatisfactory social encounters, they might begin deteriorating one's social expectations, thereby having the opposite effect of promoting social withdrawal and rejection. Future work could test this hypothesis in longitudinal studies of loneliness on social behaviors, which are to date still missing, especially in early midlife populations (25-40 years)[66].

Moreover, previous evidence suggested that lonelier individuals are less sensitive to positive signals of partner reciprocity but a direct test of the underlying mechanisms underpinning loneliness' modulation of social behaviors with differently cooperative others had been missing[67]. Here, we demonstrated that despite appropriate behavioral responses to unreciprocating

partners, lonelier individuals showed difficulties in recognizing and responding to reciprocating others, especially if they held more negative expectations of others' trustworthiness. This pattern of heightened uncertainty may reflect the vigilant cognitive style that characterizes loneliness[65], where individuals become more cautious and deliberative in their social judgments due to past experiences of social rejection or exclusion. Our results suggest that with their more volatile behavior, lonelier individuals may misplace their trust in the wrong partners and inadequately trust reliable others, putting at risk their ability to establish meaningful, satisfying bonds with others[62].

Our model further revealed that this increased uncertainty did not merely reflect impaired belief updating, but rather systematic differences in lonely individuals' preferences for relying on social information about their partner and their engagement of socio-cognitive processes. Specifically, lonely individuals' distrust was driven by a reduced willingness to trust and deeper recursive thinking. Both point to a heightened suspiciousness that contributes to underweighting of partner reciprocity beliefs, a computational pattern similar to other mental health disorders such as schizophrenia[68]. Relying less on beliefs about others' reciprocity promotes more uncertain behaviors that make partners less cooperative, which in turn reinforces lonely individuals' negative social expectations, inducing a spiral of distrust. Indeed, towards the end of the interactions, we observed that in lonelier individuals, beliefs about partner reciprocity tended to be all similarly negative irrespective of the intrinsic reciprocity trait of the partner.

Moreover, deep recursive thinking (a hallmark feature of mentalizing[57,69]) in trusting interactions was not beneficial for the establishment of successful interactions. This is in line with previous evidence for more positive judgments of "moral intuitionists" using shallower and faster decision-making processes[70,71] and with accounts of "hypermentalizing" as core feature of social cognitive impairments in clinical populations[72,73]. Specifically, if deep recursive thinking can in general leave impressions of being strategic and hence unreliable, as one would change their behavior depending on the circumstances, lonelier individuals' engagement in deeper recursive thinking can partly explain previous evidence that people interacting with lonely individuals report having left with generally negative impressions about them[70,74,75].

Importantly, to our knowledge, our results provide the first behavioral evidence to previous theoretical predictions that lower sophistication of mentalizing abilities (i.e., shallower recursive thinking) should favor prosocial behaviors such as trust and cooperation, particularly puzzling from an evolutionary perspective[76,77]. Deep recursive thinking may hinder successful relationships by promoting paranoid-like behaviors that affect interactions in various ways[67,78]. For example, inappropriate increases of trust may signal suspicious cautiousness that discourages cooperative

partners. Withdrawing from an uncooperative partner may deprive them of the chance to cooperate. Finally, probing behaviors may prompt distrust and uncertainty that can irritate benevolent partners, scare cautious ones and alert the strategic. Hence, lonely individuals' excessive mentalizing may induce uncertain and cautious behaviors that undermine the core ingredients of well-functioning relationships: confidence and a sense of safety[79,80].

Evidence in our study for lonelier individuals' more random behavior leading to lower levels of partner reciprocity, particularly during interactions with more reciprocating trustees, indicates that they inadvertently hamper a potentially successful trusting exchange with noisier behaviors that prompt others' uncooperativeness. These dynamics are akin to what in a clinical context has been termed social thinning[81], which refers to social and behavioral patterns that constrain and diminish a person's network of affiliative or supportive relationships over time. In this framework, failing to trust reliable partners is not merely a behavioral error but a self-reinforcing mechanism whereby distrustful behavior may discourage reciprocation from genuinely cooperative partners, providing apparent confirmation of one's negative expectations, and ultimately limiting opportunities for social learning and support[82].

Neuroimaging results revealed that participants exhibited stronger MCC activity during trust decisions with less reciprocating trustees and stronger VMPFC activity for greater trust. The VMPFC has previously been associated with valuation processes in both social and non-social contexts[83,84], while the MCC is pivotal to action uncertainty in social cognition and decision-making under uncertainty[85,86]. In the context of trust behavior, enhanced MCC activity likely reflects the increased computational burden of navigating social interactions under conditions of uncertainty[87]. On the contrary, the VMPFC likely encodes representations of others' trustworthiness that underpin trust decisions[88,89]. Importantly, in our study, the VMPFC was less engaged by lonelier individuals during trust choices with reciprocating partners, indicating a failure of adequate trust valuations of trustworthy partners, which may lead to their distrustful behavior.

In addition, we found that the interplay between VMPFC and MCC was central for integration of belief representations into trust values underlying action choices. Specifically, functional coupling between these two brain regions governed the translation of belief representations into action, promoting stable strategies under positive beliefs but erratic behavior under negative ones. Reduced activations in the VMPFC in lonelier individuals with more negative trustworthiness expectations might play a role in lonelier individuals' ability to appropriately guide behavioral adaptations with updated beliefs about others' cooperativeness. This may lead lonely individuals to misplace their trust and leave them unable to act on evidence of trustworthiness

precisely when it matters most.

In conclusion, the present study provides multidimensional insights into how loneliness is associated with systematic alterations in the behavioral, computational, and neural mechanisms underlying trust decisions. These findings advance our understanding of the cognitive and neural mechanisms by which loneliness affects social behavior and may inform the development of more effective interventions for addressing the social difficulties experienced by lonely individuals[90]. For example, greater behavioral uncertainty and deeper recursive thinking in lonely individuals indicate that social interactions may be particularly demanding for these individuals, potentially contributing to the withdrawal that often characterizes chronic loneliness[91]. The demonstration that loneliness modulates specific aspects of social cognition, rather than producing global social impairments, provides a more nuanced view of how social experiences shape social behavior and points toward targeted approaches for understanding and treating social dysfunction.

Materials and Methods

Participants. In the fMRI study, 32 participants (age: *M* = 28.1, *SD* = 10.9; 20 males, 12 females) were recruited; in the online study, 29 participants (age: *M* = 39.3, *SD* = 12.1; 13 males, 14 females, 1 preferred not to say and 1 selected "other" without further specifications). Inclusion criteria: right-handed, age between 18 and 65 years, normal or corrected vision, no history of psychiatric disorders/mental illnesses. All participants had normal hearing and normal or corrected-to-normal vision. The fMRI study was approved by the Ethics Committee at [*masked for double-blinded review*] and the online study by the Ethics Committee at [*masked for double-blinded review*]. All participants provided written informed consent before the experiment.

Task. We implemented a modified version of a multi-round investment game (**Fig. 1A**)[51]. All participants played the role of the investor, even though participants were led to believe that the role was randomly assigned at the beginning of the experiment. At the start of each trial, the avatar of the other player was displayed. Participants (investors) received an initial endowment of €10 and indicated, on a horizontal scale from zero to ten euros, how much they wanted to share with the other player (trust). The chosen amount was tripled during the transaction and after this, the trustee made a repayment (reciprocity). At the end of the trial, they saw the kept and given amounts and the total earnings for both players. After this a new trial started. Participants played with three different IPOMDP-trustees with three different "reciprocity traits" (*Supplementary Methods*).

Model-agnostic analyses. Mixed-effect Bayesian regressions were implemented in model-agnostic analyses. For all models, we used uninformative, Gaussian beta priors, i.e., $\beta \sim \mathcal{N}(0,5)$. Stan was used to fit models with the no-U-turn sampler for efficient exploration of posterior estimates (https://mc-stan.org) with four chains, 1,000 warm-up steps and 6,000 samples. The 95% highest posterior density intervals were calculated from these samples. Visual inspection of the traces and the Gelman-Rubin statistics ($\hat{R}$) were used to assess convergence[92] (*Supplementary Methods*).

Computational modeling analyses. We modeled participants as Interactive Partially Observable Markov Decision Process (IPOMDP) agents making Bayesian inferences about their partners' reciprocity and computing trust action utilities via a vulnerability function. The vulnerability function formalizes trust as a tradeoff between the costs of potential betrayal and the

expected benefits of partner reciprocity, scaled by a participant-specific belief-reliance parameter (τ) that captures the degree to which a decision-maker relies on learned beliefs about their partner when making trust choices. The IPOMDP framework additionally captures two socio-cognitive processes: recursive thinking, which operationalizes mentalizing by embedding each agent's model of the other in a hierarchical Bayesian inference, and planning horizon, which reflects how far ahead participants anticipate future interactions. To fit the model, we performed a grid search over τ parameters (8 values: 0.1-0.8), recursion levels (1-3), and planning horizons (1-3) for each participant and trustee condition, selecting the parameter combination minimizing the Bayesian Information Criterion (see *Supplementary Methods*). Model comparison confirmed that the IPOMDP approach outperformed a set of ten non-IPOMDP alternatives, and that the vulnerability function provided a better account of participants' choices than the inequity-aversion function, irrespective of the modeling approach used (see **Fig. S1**). Models with biased belief-updating and discounting mechanisms did not improve fit over the unbiased baseline. Parameter recovery analyses confirmed that all three parameters were reliably identifiable from the data (see *Supplementary Results*).

Neuroimaging analyses.

Scanning parameters. Functional images were collected using a T2*-weighted echo planar imaging (EPI) sequence optimized for blood oxygenation level-dependent (BOLD) contrast (430 volumes; repetition time [TR] = 2000 ms; echo time [TE] = 35 ms; field of view = 240 mm; slice thickness = 3 mm; matrix size = 64 × 64; flip angle = 75°) on a 3T MAGNETOM Vida MRI system (Siemens, Germany). High-resolution structural images were also acquired for each participant using a T1-weighted magnetization-prepared rapid gradient echo (MP-RAGE) sequence (TR = 7321 ms; TE = 3 ms; inversion time [TI] = 400 ms; field of view = 240 mm; slice thickness = 1.2 mm; 196 slices).

Image preprocessing. Neuroimaging data were preprocessed using SPM12 (v. 6905; http://www.fil.ion.ucl.ac.uk/spm/software/spm12/) in MATLAB R2022b (The Mathworks, Natick, Massachusetts; http://www.mathworks.com/). Functional images were corrected for slice acquisition time, realigned to the mean image, co-registered to their structural images using the unified segmentation procedure[93], normalized into MNI space using deformation fields from the segmentation procedure (resampling voxel size: 2×2×2 mm^3), and spatially smoothed using a Gaussian filter (8×8×8 mm^3 full width at half maximum, FWHM) to decrease spatial noise.

First-level neuroimaging analyses. At the first level, a general linear model (GLM1) was estimated with 7 regressors to test for main effects of each of the task phases (i.e., "trust", "coplayer decision", "reciprocity", "feedback"; see **Fig. 1B**) with associated parametric modulators across trustee types. Specifically, for the trust and reciprocity phases, GLM1 had amounts invested (trust) and returned (reciprocity) as parametric modulators, respectively. For the feedback phase, GLM1 had amounts earned in that trial by the participant as parametric modulators. Further, GLM2 contained the same regressors for each trustee type separately. Motion parameters were included as regressors of no-interest. A temporal high-pass filter with a cutoff of 128 s was applied. Contrast analyses between beta regressors were performed on the first level and later used for second-level whole-brain analyses.

Second-level neuroimaging analyses. To examine neural signatures of trusting behaviors, one-sample *t*-tests on main effects and parametric modulators were performed at the second (group) level. To study group differences, contrast images from the first level were entered into a two-sample *t*-test. All results were whole-brain corrected for multiple comparisons using a voxel-defining threshold of $p < .001$ and a family-wise error of $p < .05$ (cluster-level corrected)[94].

Whole-brain regression analyses. To investigate the relationships of individual willingness to trust (τ parameter estimated for each participant by the winning computational model) with neural responses in the game, contrast images from the first level were regressed against τ parameter values in a group-level whole-brain regression analysis (using the same multiple comparison parameters as above). A similar whole-brain regression model was run with average individual trust.

Task-dependent functional connectivity analysis. To investigate the potential functional pathways of the VMPFC, a task-dependent functional connectivity analysis was implemented using a whole-brain psychophysiological interaction analysis (PPI)[95,96] with the significant VMPFC cluster derived from the main effect of the trust choice phase as seed region (10 mm radius). The PPI-GLM consisted of a task regressor, a physiological regressor entailing deconvolved BOLD signal from the seed region and a regressor for the interaction term. Movement parameters were entered as regressors of no interest. Significant connectivity was assessed with the same multiple comparison parameters as above.

Labeling and data visualization. The SPM Anatomy toolbox (v. 2.2)[97] and MRIcron (http://people.cas.sc.edu/rorden/mricron/install.html/) were used for anatomical labeling. MRIcroGL (https://www.mccauslandcenter.sc.edu/mricrogl/home/) was used for visualizations.

Data availability.

Behavioral data are openly available on OSF (https://osf.io/hwtxk/overview?view_only=5bc05bc96c464998a38596cdc5c50800). Neuroimaging data are openly available on NeuroVault (https://identifiers.org/neurovault.collection:23800).

Code availability.

Code for analyses is openly available on OSF (https://osf.io/hwtxk/overview?view_only=5bc05bc96c464998a38596cdc5c50800).

References

1. Holt-Lunstad, J. Social connection as a critical factor for mental and physical health: evidence, trends, challenges, and future implications. *World Psychiatry* **23**, 312–332 (2024).
2. Hawkley, L. C. & Capitanio, J. P. Perceived social isolation, evolutionary fitness and health outcomes: a lifespan approach. *Philos. Trans. R. Soc. B Biol. Sci.* **370**, 20140114 (2015).
3. Hawkley, L. C., Burleson, M. H., Berntson, G. G. & Cacioppo, J. T. Loneliness in everyday life: Cardiovascular activity, psychosocial context, and health behaviors. *J. Pers. Soc. Psychol.* **85**, 105–120 (2003).
4. Michalska da Rocha, B., Rhodes, S., Vasilopoulou, E. & Hutton, P. Loneliness in Psychosis: A Meta-analytical Review. *Schizophr. Bull.* **44**, 114–125 (2018).
5. WHO launches commission to foster social connection. https://www.who.int/news/item/15-11-2023-who-launches-commission-to-foster-social-connection.
6. Danielle, K. & Steiner, A. Men's Sheds - A toolkit for the development and sustainability of sheds in the UK and Ireland. *Glasg. Caled. Univ.* https://doi.org/https://menssheds.ie/wp-content/uploads/2021/08/MENS-SHED-TOOLKIT-FOR-UK-AND-IRELAND.pdf (2021) doi:https://menssheds.ie/wp-content/uploads/2021/08/MENS-SHED-TOOLKIT-FOR-UK-AND-IRELAND.pdf.
7. Golding, B. *Shoulder to Shoulder: Broadening the Men's Shed Movement*. (Common Ground Research Networks, 2021).
8. Duck, S., Pond, K. & Leatham, G. Loneliness and the Evaluation of Relational Events. *J. Soc. Pers. Relatsh.* **11**, 253–276 (1994).
9. Gardner, W. L., Pickett, C. L., Jefferis, V. & Knowles, M. On the Outside Looking In: Loneliness and Social Monitoring. *Pers. Soc. Psychol. Bull.* **31**, 1549–1560 (2005).

10. Pickett, C. L. & Gardner, W. L. The Social Monitoring System: Enhanced Sensitivity to Social Cues as an Adaptive Response to Social Exclusion. in *The social outcast: Ostracism, social exclusion, rejection, and bullying* 213–226 (Psychology Press, New York, NY, US, 2005).
11. Bellucci, G. Positive attitudes and negative expectations in lonely individuals. *Sci. Rep.* **10**, 18595 (2020).
12. Lay-Yee, R., Campbell, D. & Milne, B. Social attitudes and activities associated with loneliness: Findings from a New Zealand national survey of the adult population. *Health Soc. Care Community* **30**, 1120–1132 (2022).
13. Gong, X. & Nikitin, J. "When i feel lonely, i'm not nice (and neither are you)": the short- and long-term relation between loneliness and reports of social behaviour. *Cogn. Emot.* **35**, 1029–1038 (2021).
14. Freeman, D., Garety, P. A., Kuipers, E., Fowler, D. & Bebbington, P. E. A cognitive model of persecutory delusions. *Br. J. Clin. Psychol.* **41**, 331–347 (2002).
15. Kesting, M.-L., Bredenpohl, M., Klenke, J., Westermann, S. & Lincoln, T. M. The impact of social stress on self-esteem and paranoid ideation. *J. Behav. Ther. Exp. Psychiatry* **44**, 122–128 (2013).
16. Steenkamp, L., Weijers, J., Gerrmann, J., Eurelings-Bontekoe, E. & Selten, J.-P. The relationship between childhood abuse and severity of psychosis is mediated by loneliness: an experience sampling study. *Schizophr. Res.* **241**, 306–311 (2022).
17. Fett, A.-K. J. *et al.* The relationship between neurocognition and social cognition with functional outcomes in schizophrenia: A meta-analysis. *Neurosci. Biobehav. Rev.* **35**, 573–588 (2011).
18. Halverson, T. F. *et al.* Pathways to functional outcomes in schizophrenia spectrum disorders: Meta-analysis of social cognitive and neurocognitive predictors. *Neurosci. Biobehav. Rev.* **105**, 212–219 (2019).

19. Turner, D. T. *et al.* A Meta-Analysis of Social Skills Training and Related Interventions for Psychosis. *Schizophr. Bull.* **44**, 475–491 (2018).

20. Mayer, R. C., Davis, J. H. & Schoorman, F. D. An Integrative Model Of Organizational Trust. *Acad. Manage. Rev.* **20**, 709–734 (1995).

21. Rousseau, D. M., Sitkin, S. B., Burt, R. S. & Camerer, C. Not So Different After All: A Cross-Discipline View Of Trust. *Acad. Manage. Rev.* **23**, 393–404 (1998).

22. Bellucci, G. A Model of Trust. *Games* **13**, 39 (2022).

23. Dunning, D., Anderson, J. E., Schlösser, T., Ehlebracht, D. & Fetchenhauer, D. Trust at zero acquaintance: More a matter of respect than expectation of reward. *J. Pers. Soc. Psychol.* **107**, 122–141 (2014).

24. Fetchenhauer, D. & Dunning, D. Do people trust too much or too little? *J. Econ. Psychol.* **30**, 263–276 (2009).

25. Bellucci, G. & Park, S. Q. Loneliness is associated with more trust but worse trustworthiness expectations. *Br. J. Psychol.* **115**, 641–664 (2024).

26. Stenzel, K., Weiß, M. & Hein, G. A Negative Reputation Reduces Trust Despite Trustworthy Behavior. *Psychophysiology* **62**, e70102 (2025).

27. Bellucci, G. & Park, S. Q. Honesty biases trustworthiness impressions. *J. Exp. Psychol. Gen.* **149**, 1567–1586 (2020).

28. Unoka, Z., Seres, I., Áspán, N., Bódi, N. & Kéri, S. Trust Game Reveals Restricted Interpersonal Transactions in Patients With Borderline Personality Disorder. *J. Personal. Disord.* **23**, 399–409 (2009).

29. Dunning, D., Fetchenhauer, D. & Schlösser, T. Why People Trust: Solved Puzzles and Open Mysteries. *Curr. Dir. Psychol. Sci.* **28**, 366–371 (2019).

30. Fetchenhauer, D. & Dunning, D. Why So Cynical?: Asymmetric Feedback Underlies Misguided Skepticism Regarding the Trustworthiness of Others. *Psychol. Sci.* **21**, 189–193 (2010).

31. Bellucci, G., Imada, H., Fett, A.-K. & Ito, A. Neurocomputational mechanisms of maladaptive behaviors in loneliness. *Trends Cogn. Sci.* https://doi.org/10.1016/j.tics.2025.07.007 (2025) doi:10.1016/j.tics.2025.07.007.

32. Frith, C. D. & Frith, U. The Neural Basis of Mentalizing. *Neuron* **50**, 531–534 (2006).

33. Dennett, D. C. Beliefs about beliefs [P&W, SR&B]. *Behav. Brain Sci.* **1**, 568–570 (1978).

34. Premack, D. & Woodruff, G. Does the chimpanzee have a theory of mind? *Behav. Brain Sci.* **1**, 515–526 (1978).

35. Bo, S., Sharp, C., Fonagy, P. & Kongerslev, M. Hypermentalizing, attachment, and epistemic trust in adolescent BPD: Clinical illustrations. *Personal. Disord. Theory Res. Treat.* **8**, 172–182 (2017).

36. Fonagy, P. & Luyten, P. A developmental, mentalization-based approach to the understanding and treatment of borderline personality disorder. *Dev. Psychopathol.* **21**, 1355–1381 (2009).

37. Kazinka, R., Vilares, I. & MacDonald, A. W. A Computational Model of Non-optimal Suspiciousness in the Minnesota Trust Game. *Comput. Psychiatry* **6**, (2022).

38. Bellucci, G., Keramati, M., Hanssen, E. & Fett, A.-K. Willingness to trust is reduced by loneliness and paranoia. *PsyArXiv* https://doi.org/10.31234/osf.io/vnqys (2024) doi:10.31234/osf.io/vnqys.

39. Bellucci, G. & Park, S. Q. Neurocomputational mechanisms of biased impression formation in lonely individuals. *Commun. Biol.* **6**, 1–13 (2023).

40. Ma, N. *et al.* Traits are represented in the medial prefrontal cortex: an fMRI adaptation study. *Soc. Cogn. Affect. Neurosci.* **9**, 1185–1192 (2014).

41. Piva, M. *et al.* The dorsomedial prefrontal cortex computes task-invariant relative subjective value for self and other. *eLife* **8**, e44939 (2019).

42. Krueger, F., Barbey, A. K. & Grafman, J. The medial prefrontal cortex mediates social event knowledge. *Trends Cogn. Sci.* **13**, 103–109 (2009).

43. Inagaki, T. K. *et al.* Yearning for connection? Loneliness is associated with increased ventral striatum activity to close others. *Soc. Cogn. Affect. Neurosci.* **11**, 1096–1101 (2016).

44. Tomova, L. *et al.* Acute social isolation evokes midbrain craving responses similar to hunger. *Nat. Neurosci.* **23**, 1597–1605 (2020).

45. Lieberz, J. *et al.* Loneliness and the Social Brain: How Perceived Social Isolation Impairs Human Interactions. *Adv. Sci.* **8**, 2102076 (2021).

46. Igelström, K. M., Webb, T. W. & Graziano, M. S. A. Neural Processes in the Human Temporoparietal Cortex Separated by Localized Independent Component Analysis. *J. Neurosci.* **35**, 9432–9445 (2015).

47. Engelmann, J. B., Meyer, F., Ruff, C. C. & Fehr, E. The neural circuitry of affect-induced distortions of trust. *Sci. Adv.* **5**, eaau3413 (2019).

48. Saxe, R., Xiao, D.-K., Kovacs, G., Perrett, D. I. & Kanwisher, N. A region of right posterior superior temporal sulcus responds to observed intentional actions. *Neuropsychologia* **42**, 1435–1446 (2004).

49. Lamba, A., Frank, M. J. & FeldmanHall, O. Anxiety Impedes Adaptive Social Learning Under Uncertainty. *Psychol. Sci.* **31**, 592–603 (2020).

50. Vandendriessche, H. *et al.* Contextual influence of reinforcement learning performance of depression: evidence for a negativity bias? *Psychol. Med.* **53**, 4696–4706 (2023).

51. Berg, J., Dickhaut, J. & McCabe, K. Trust, Reciprocity, and Social History. *Games Econ. Behav.* **10**, 122–142 (1995).

52. FeldmanHall, O. & Nassar, M. R. The computational challenge of social learning. *Trends Cogn. Sci.* **25**, 1045–1057 (2021).

53. Baker, C. L., Jara-Ettinger, J., Saxe, R. & Tenenbaum, J. B. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nat. Hum. Behav.* **1**, 1–10 (2017).

54. Khalvati, K. *et al.* Modeling other minds: Bayesian inference explains human choices in group decision-making. *Sci. Adv.* **5**, eaax8783 (2019).

55. Silver, D. & Veness, J. Monte-Carlo Planning in Large POMDPs. *Adv. Neural Inf. Process. Syst.* **23**, (2010).

56. Koshelev, M., Lohrenz, T., Vannucci, M. & Montague, P. R. Biosensor Approach to Psychopathology Classification. *PLOS Comput. Biol.* **6**, e1000966 (2010).

57. Camerer, C. F., Ho, T.-H. & Chong, J.-K. A Cognitive Hierarchy Model of Games*. *Q. J. Econ.* **119**, 861–898 (2004).

58. Fehr, E. & Schmidt, K. M. A Theory of Fairness, Competition, and Cooperation*. *Q. J. Econ.* **114**, 817–868 (1999).

59. Reiter, A. M. F. *et al.* Self-reported childhood family adversity is linked to an attenuated gain of trust during adolescence. *Nat. Commun.* **14**, 6920 (2023).

60. Hiser, J. & Koenigs, M. The Multifaceted Role of the Ventromedial Prefrontal Cortex in Emotion, Decision Making, Social Cognition, and Psychopathology. *Biol. Psychiatry* **83**, 638–647 (2018).

61. Cacioppo, J. T. & Hawkley, L. C. Perceived social isolation and cognition. *Trends Cogn. Sci.* **13**, 447–454 (2009).

62. Qualter, P. *et al.* Loneliness Across the Life Span. *Perspect. Psychol. Sci.* **10**, 250–264 (2015).

63. Shamay-Tsoory, S. G. & Kanterman, A. Away from the herd: loneliness as a dysfunction of social alignment. *Soc. Cogn. Affect. Neurosci.* **19**, nsae005 (2024).

64. Loneliness in the Modern Age: An Evolutionary Theory of Loneliness (ETL). in *Advances in Experimental Social Psychology* vol. 58 127–197 (Academic Press, 2018).

65. Hawkley, L. C. & Cacioppo, J. T. Loneliness Matters: A Theoretical and Empirical Review of Consequences and Mechanisms. *Ann. Behav. Med.* **40**, 218–227 (2010).

66. Mund, M., Freuding, M. M., Möbius, K., Horn, N. & Neyer, F. J. The Stability and Change of Loneliness Across the Life Span: A Meta-Analysis of Longitudinal Studies. *Personal. Soc. Psychol. Rev. Off. J. Soc. Personal. Soc. Psychol. Inc* **24**, 24–52 (2020).

67. Bellucci, G., Keramati, M., Hanssen, E. & Fett, A.-K. Willingness to trust is reduced by loneliness and paranoia. *Commun. Psychol.* https://doi.org/10.1038/s44271-025-00384-6 (2025) doi:10.1038/s44271-025-00384-6.

68. Petrovic, P. & Sterzer, P. Resolving the Delusion Paradox. *Schizophr. Bull.* **49**, 1425–1436 (2023).

69. Houlihan, S. D., Kleiman-Weiner, M., Hewitt, L. B., Tenenbaum, J. B. & Saxe, R. Emotion prediction as computation over a generative theory of mind. *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.* **381**, 20220047 (2023).

70. Everett, J. A. C., Pizarro, D. A. & Crockett, M. J. Inference of trustworthiness from intuitive moral judgments. *J. Exp. Psychol. Gen.* **145**, 772–787 (2016).

71. Critcher, C. R., Inbar, Y. & Pizarro, D. A. How quick decisions illuminate moral character. *Soc. Psychol. Personal. Sci.* **4**, 308–315 (2013).

72. Sharp, C. & and Vanwoerden, S. Hypermentalizing in Borderline Personality Disorder: A Model and Data. *J. Infant Child Adolesc. Psychother.* **14**, 33–45 (2015).

73. Blakemore, S. J., Sarfati, Y., Bazin, N. & Decety, J. The detection of intentional contingencies in simple animations in patients with delusions of persecution. *Psychol. Med.* **33**, 1433–1441 (2003).

74. Levine, E. E., Barasch, A., Rand, D., Berman, J. Z. & Small, D. A. Signaling emotion and reason in cooperation. *J. Exp. Psychol. Gen.* **147**, 702–719 (2018).

75. Rand, D. G. Cooperation, Fast and Slow: Meta-Analytic Evidence for a Theory of Social Heuristics and Self-Interested Deliberation. *Psychol. Sci.* **27**, 1192–1206 (2016).

76. Devaine, M., Hollard, G. & Daunizeau, J. Theory of Mind: Did Evolution Fool Us? *PLOS ONE* **9**, e87619 (2014).

77. Yamagishi, T., Kikuchi, M. & Kosugi, M. Trust, Gullibility, and Social Intelligence. *Asian J. Soc. Psychol.* **2**, 145–161 (1999).

78. Bell, V. *et al.* Do loneliness and social exclusion breed paranoia? An experience sampling investigation across the psychosis continuum. *Schizophr. Res. Cogn.* **33**, 100282 (2023).

79. Murray, S. L. *et al.* The social-safety system: Fortifying relationships in the face of the unforeseeable. *J. Pers. Soc. Psychol.* **120**, 99–130 (2021).

80. Simpson, J. A. Psychological Foundations of Trust. *Curr. Dir. Psychol. Sci.* **16**, 264–268 (2007).

81. McCrory, E., Foulkes, L. & Viding, E. Social thinning and stress generation after childhood maltreatment: a neurocognitive social transactional model of psychiatric vulnerability. *Lancet Psychiatry* **9**, 828–837 (2022).

82. Savulescu, J., Roache, R., Davies, W. & Loebel, J. P. *Psychiatry Reborn: Biopsychosocial Psychiatry in Modern Medicine*. (Oxford University Press, 2020).

83. Bellucci, G., Camilleri, J. A., Eickhoff, S. B. & Krueger, F. Neural signatures of prosocial behaviors. *Neurosci. Biobehav. Rev.* **118**, 186–195 (2020).

84. Moneta, N., Garvert, M. M., Heekeren, H. R. & Schuck, N. W. Task state representations in vmPFC mediate relevant and irrelevant value signals and their behavioral influence. *Nat. Commun.* **14**, 3156 (2023).

85. Lockwood, P. L., Apps, M. A. J., Valton, V., Viding, E. & Roiser, J. P. Neurocomputational mechanisms of prosocial learning and links to empathy. *Proc. Natl. Acad. Sci.* **113**, 9763–9768 (2016).

86. Apps, M. A. J., Rushworth, M. F. S. & Chang, S. W. C. The Anterior Cingulate Gyrus and Social Cognition: Tracking the Motivation of Others. *Neuron* **90**, 692–707 (2016).

87. Eisenberger, N. I. The pain of social disconnection: examining the shared neural underpinnings of physical and social pain. *Nat. Rev. Neurosci.* **13**, 421–434 (2012).

88. Delgado, M. R., Frank, R. H. & Phelps, E. A. Perceptions of moral character modulate the neural systems of reward during the trust game. *Nat. Neurosci.* **8**, 1611–1618 (2005).

89. Bellucci, G., Molter, F. & Park, S. Q. Neural representations of honesty predict future trust behavior. *Nat. Commun.* **10**, 5184 (2019).

90. Masi, C. M., Chen, H.-Y., Hawkley, L. C. & Cacioppo, J. T. A Meta-Analysis of Interventions to Reduce Loneliness. *Personal. Soc. Psychol. Rev.* **15**, 219–266 (2011).

91. Heinrich, L. M. & Gullone, E. The clinical significance of loneliness: A literature review. *Clin. Psychol. Rev.* **26**, 695–718 (2006).

92. Gelman, A. & Rubin, D. B. Inference from Iterative Simulation Using Multiple Sequences. *Stat. Sci.* **7**, 457–472 (1992).

93. Ashburner, J. & Friston, K. J. Unified segmentation. *NeuroImage* **26**, 839–851 (2005).

94. Eklund, A., Nichols, T. E. & Knutsson, H. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proc. Natl. Acad. Sci.* **113**, 7900–7905 (2016).

95. Friston, K. J. *et al.* Psychophysiological and Modulatory Interactions in Neuroimaging. *NeuroImage* **6**, 218–229 (1997).

96. Gitelman, D. R., Penny, W. D., Ashburner, J. & Friston, K. J. Modeling regional and psychophysiologic interactions in fMRI: the importance of hemodynamic deconvolution. *NeuroImage* **19**, 200–207 (2003).

97. Eickhoff, S. B. *et al.* Assignment of functional activations to probabilistic cytoarchitectonic areas revisited. *NeuroImage* **36**, 511–521 (2007).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SM.docx
- STables.docx
- SFigures.docx