Expecting the unexpected: **the paranoid style of belief updating across species**

Erin J. Reed[1,2], Stefan Uddenberg[3], Christoph D. Mathys[4,5], Jane R. Taylor[6], Stephanie M. Groman[6], and Philip R. Corlett[6*]

[1]Interdepartmental Neuroscience Program, Yale School of Medicine, New Haven, CT, USA. [2]Yale MD-PhD Program, Yale School of Medicine, New Haven, CT, USA. [3]Department of Psychology, Princeton University, Princeton, NJ, USA. [4]Scuola Internazionale Superiore di Studi Avanzati (SISSA), Trieste, Italy. [5]Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich and ETH Zurich, Zurich, Switzerland. [6]Department of Psychiatry, Connecticut Mental Health Center, Yale University, New Haven, CT, USA. *email: philip.corlett@yale.edu

Keywords: paranoia, reversal-learning, unexpected uncertainty, hierarchical Gaussian filter, methamphetamine, delusions, computational psychiatry

Abstract

Paranoia is the belief that harm is intended by others. It may arise from selective pressures to infer and avoid social threats, particularly in ambiguous or changing circumstances. We propose that uncertainty may be sufficient to elicit learning differences in paranoid individuals, without social threat. We used reversal learning behaviour and computational modelling to estimate belief updating across individuals with and without mental illness, online participants, and rats exposed to chronic methamphetamine, an elicitor of paranoia in humans. Paranoia is associated with a strong but immutable prior on volatility, accompanied by elevated sensitivity to perceived changes in the task environment. Methamphetamine exposure in rats recapitulates this impaired uncertainty-driven belief updating and rigid anticipation of a volatile environment. Our work provides evidence of fundamental, domain-general learning differences in paranoid individuals. This paradigm enables further assessment of the interplay between uncertainty and belief-updating across individuals and species.

Paranoia is excessive concern that harm will occur due to deliberate actions of others[1]. It manifests along a continuum of increasing severity[2-5]. Fleeting paranoid thoughts prevail in the general population[6]. A survey of over 7,000 individuals found that nearly 20% believed people were against them at times in the past year; approximately 8% felt people had intentionally acted to harm them[4]. At a national level, paranoia may fuel divisive ideological intolerance. Historian Richard Hofstadter famously described catastrophizing, context insensitive political discourse as the 'paranoid style':

> ***"The paranoid spokesman sees the fate of conspiracy in apocalyptic terms—he traffics in the birth and death of whole worlds, whole political orders, whole systems of human values. He is always manning the barricades of civilization. He constantly lives at a turning point."***[7]

At its most severe, paranoia manifests as rigid beliefs known as delusions of persecution. These delusions occur frequently in psychotic illness, including nearly 90% of first episode patients[8]. However, paranoid beliefs are common across psychiatric and neurologic disorders, such as anxiety[9], depression[10], epilepsy[11], and Alzheimer's disease[12]. Psychostimulants elicit severe paranoid states. Methamphetamine evoked new paranoid ideation in nearly half of 274 respondents, particularly after repeated exposure (86%) or escalating dose (68%)[13]. Of those who became paranoid, the majority engaged in evasive defence strategies (hiding or fleeing), but 37% obtained weapons, and 15% attacked others. There is a clear need to better manage paranoia, and to understand and address its broader societal impact.

Paranoia has thus far defied explanation in mechanistic terms, either at the levels of behaviour or brain function. Obvious links with fear processing and social cognition, including sophisticated Game Theory driven approaches (such as the Dictator Game[14,15]) have largely re-described the phenomenon — people who are paranoid self-report difficulties with trust. Those difficulties are recapitulated in laboratory tasks that require trust[16]. However, large-scale online work with inter-personal, Game Theory motivated tasks has shown that paranoia is not driven by personal threat per se, but by negative social representations of others[14,15]. We and others have argued that such reputations are learned[17,18], via the same fundamental learning mechanisms[19] that stimulate non-social learning in non-human species[20]. We hypothesize that domain-general learning differences, particularly in the processing of uncertainty, underlie paranoia.

In prior work, we have shown that prediction errors, mismatches between expectation and experience that drive learning in non-human species[21], contribute to the formation of causal beliefs and delusions in humans[22,23]. However, delusion maintenance, which we conceive of as impaired belief updating, has yet to be related definitively to specific learning mechanisms. Higher order beliefs or expectations about the noisiness of the environment may constrain whether we update beliefs or dismiss surprises as probabilistic anomalies. Expected uncertainty, also described as risk, provides one such constraint: the perceived probabilistic variability in an environment[24]. The higher the expected uncertainty, the less surprising an atypical outcome may be, and the less compelling it is for driving belief updates. Unexpected uncertainty, in contrast, describes perceived change in the underlying statistics of the environment[25-27]. This perception promotes new learning and revision of past beliefs[24,28]. Hofstadter's description of 'paranoid style' evokes the concept of unexpected uncertainty — i.e., living 'constantly…at a turning point.'[7] Excessive unexpected uncertainty is consistent with evolutionary theories attributing paranoia to the need to flexibly categorize or re-categorize social threats[16]. On the other hand, persecutory delusions are resistant to belief updating by definition, and even subclinical paranoia has been associated with reduced sensitivity to meaningful information in a task environment[29].

To address this seeming paradox – excessive and deficient belief updating in paranoia – we behaviourally and computationally dissected learning mechanisms in settings of expected and unexpected uncertainty. Given our premise that paranoid learning arises from domain-general mechanisms, we invited participants to complete a non-social, three-option probabilistic learning task. Participants learn and update reward associations in response to perceived probabilistic variability of outcomes, anticipated but temporally uncertain exchange of reward probabilities between options (reversal events), and unanticipated changes in the underlying probabilities themselves (context change). This task challenges participants to update beliefs about the value of each option and the volatility of the task environment. The Hierarchical Gaussian Filter (HGF)[30,31], a generative model of Bayesian belief, allows us to infer parameters governing learning rates from expected and unexpected variation in the task environment, initial beliefs (i.e., priors) for task volatility, and readiness to learn about changes in the task volatility itself. Beliefs concerning the values of each option update according to prediction errors weighted by belief precision; volatility prediction errors drive updates at higher levels of belief (i.e., beliefs about context). We examined the behavioural and computational correlates of paranoia both in-person and in a large online sample, spanning patients and healthy controls with varying degrees of paranoia. We also undertook a pre-clinical replication in rodents exposed chronically to saline or methamphetamine[32]. We predicted that paranoia-related learning differences would be particularly prominent in settings of contextual change. We observed elevated sensitivity to unexpected uncertainty resulting in excessive revision of option-outcome associations, accompanied by elevated volatility priors and deficient learning about contextual change (metavolatility).

Results

We analysed belief updating across three reversal-learning experiments (Fig. 1): an in laboratory pilot of patients and healthy controls, stratified by stable, paranoid personality trait (Experiment 1); four online task variants administered to participants via the Amazon Mechanical Turk (MTurk) marketplace (Experiment 2); and a re-analysis of data from rats on chronic, escalating doses of methamphetamine, a translational model of paranoia (Experiment 3)[32].

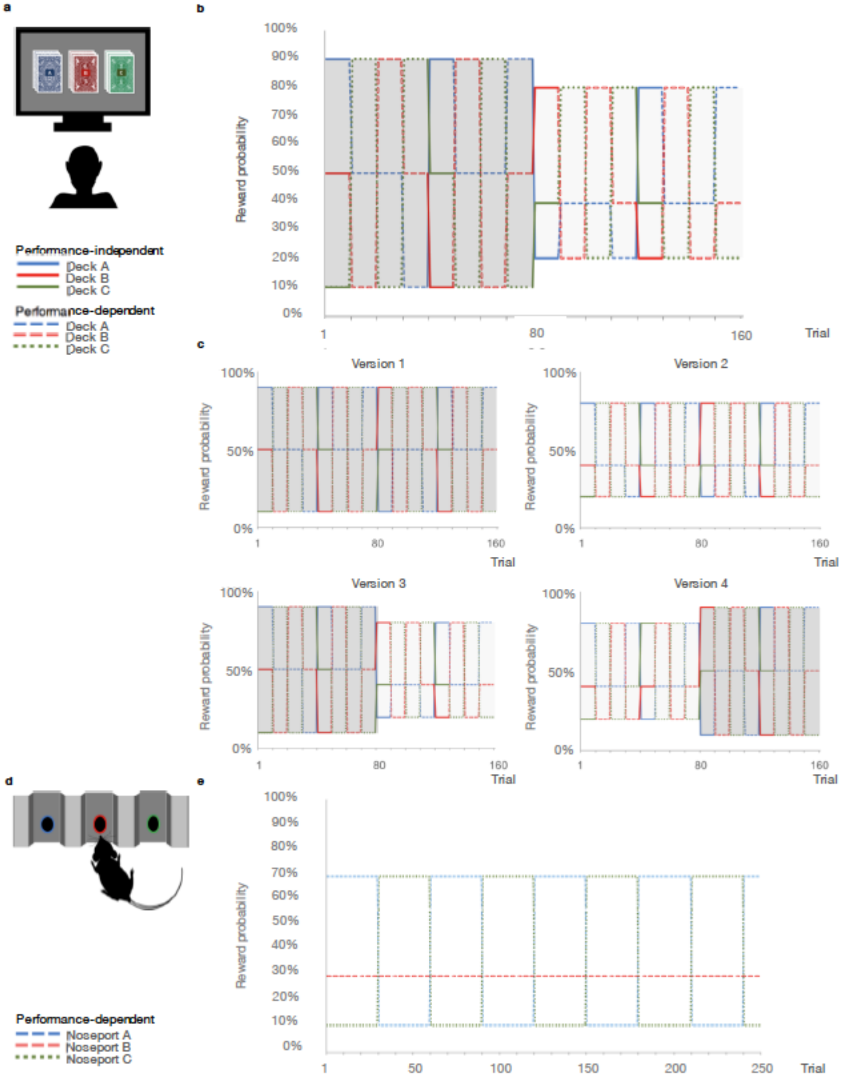


Fig. 1. Probabilistic reversal learning task. a, Human paradigm: participants choose between three ecks of cards with different, unknown probabilities of reward and loss. **b,** Reward contingency schedule for in laboratory experiment. On trial 81, the probability context shifts from 90%, 50%, and 10% (dark grey) to 80%, 40%, and 20% without warning (light grey). **c**, Reward contingency schedules for online experiment. **d,** Rat paradigm: subjects choose between three noseports with different probabilities of sucrose pellet reward. **e,** Reward contingency schedule for rat experiment[39].

Experiment 1. First, we explored trans-diagnostic associations between paranoia and performance on a reversal-learning paradigm. Participants (*n*=32) with and without psychiatric diagnoses (anxiety, depression, bipolar disorder, schizophrenia, and schizoaffective disorder) completed questionnaire versions of the *Structured Clinical Interview for DSM-IV Axis II Personality Disorders* (SCID-II) screening assessment[33], Beck's Anxiety Inventory (BAI)[34], Beck's Depression Inventory (BDI)[35], and demographic assessments (Table 1). Approximately two-thirds of participants endorsed three or fewer items on the SCID-II paranoid personality subscale (median=1 item). Participants who endorsed four or more items were classified as high paranoia (*n*=11), consistent with the diagnostic threshold for paranoid personality disorder. Low paranoia (*n*=21) and high paranoia groups did not differ significantly by age, nor were there significant group associations with gender, educational attainment, ethnicity, or race, although a larger percentage of paranoid participants identified as racial minorities or “not specified” (Table 1). Diagnostic category (i.e., healthy control, mood disorder, or schizophrenia spectrum) was significantly associated with paranoia group membership, χ^2 (2, *n*=32)=12.329, *P*=0.002, Cramer’s V=0.621, as was psychiatric medication usage, χ^2 (1, *n*=32)=9.871, *P*=0.003, Cramer’s V=0.555. These differences were due largely to the higher proportion of healthy controls in the low paranoia group. SCID-II paranoia scores correlated with symptoms of anxiety and depression (BAI: Pearson’s *r*=0.611, *P*=0.0002, 95% Confidence Interval (CI)=[0.315,0.906]; BDI: Pearson’s *r*=0.564, *P*=0.001, CI=[0.257, 0.872]). As expected, paranoia, BAI, and BDI scores were significantly elevated in the high paranoia group relative to low paranoia controls (Table 1; paranoia: mean difference (MD)=0.536, CI=[0.455,0.618], *t*(30)=13.476, *P*=2.92E-14, Hedges’ *g*=5.016; BAI: MD=0.585, CI=[0.239, 0.931], *t*(30)=3.453, *P*=0.002, Hedges’ *g*=1.285, MD=-0.585; BDI: MD=0.427, CI=[0.078, 0.775], *t*(11.854)=2.67, *P*=0.021, Hedges’ *g*=1.255).

Participants completed a three-option reversal-learning task in which they chose between three decks of cards with hidden reward probabilities (Fig. 1 a and b). They selected a deck on each turn and received positive or negative feedback (+100 or -50 points, respectively). They were instructed to find the best deck with the caveat that this deck may change. Undisclosed to participants, reward probabilities switched among decks after selection of the highest probability option in nine out of ten consecutive trials (“reversal events”). Reward probability context changed from 90%, 50%, and 10% chance of reward to 80%, 40%, and 20% between the first and second halves of the task (“contingency context change”; block 1=80 trials, 90-50-10%; block 2=80 trials, 80-40-20%). High paranoia subjects achieved fewer reversals (MD=-2.31, CI=[-4.504, -0.111,], *t*(30)=-2.145, *P*=0.04, Hedges’ *g*=0.798), but total points earned did not significantly differ (Table 1).

Experiment 2. We replicated the effects of paranoia on reversal-learning in a larger online sample. We also tested alternative task versions to control for the contingency context change (Fig. 1c). Version 1 (*n*=45 low paranoia, 20 high paranoia) provided a constant contingency context of 90-50-10% reward probabilities; version 2 (*n*=69 low paranoia, 18 high paranoia) provided a constant context of 80-40-20%; version 3 (*n*=56 low paranoia, 16 high paranoia) replicated Experiment 1 with a context change from 90-50-10% to 80-40-20%; version 4 (*n*=64 low paranoia, 19 high paranoia) provided the reverse context change, 80-40-20% to 90-50-10%. Demographic and mental health questionnaire responses did not differ significantly across task versions (Table 2). Total points and reversals achieved suggest variations in task difficulty (Table 2, version effects: points earned, *F*(3)=232.88, *P*=4.16E-18, η_p^2=0.245; reversals achieved, *F*(3)=4.329, *P*=0.005, η_p^2=0.042), but there was no significant association between task version and attrition rate (52.7%, 52.9%, 54.6%, and 53.1% attrition, respectively; χ^2(3)=0.167, *P*=0.983, Cramer’s V=0.015).

Across task versions, high paranoia participants endorsed higher BAI and BDI scores (*n*=73 high paranoia, 234 low paranoia; BAI: *F*(1)=38.752, *P*=1.63E-09, η_p^2=0.115; BDI: *F*(1)=74.528, *P*=3.62E-16, η_p^2=0.20; Table

2). Both correlated with paranoia (BAI: Pearson's r=0.450, P=1.09E-16, CI=[0.348, 0.55]; BDI: Pearson's r=0.543, P=6.26E-25, CI=[0.448, 0.638]). Trial-by-trial reaction time did not differ significantly between low and high paranoia (Table 2), but high paranoia participants earned fewer total points (F(1)=6.175, P=0.014, η_p^2=0.020) and achieved fewer reversals (F(1)=5.762, p=0.017, η_p^2=0.019; Table 2). Deck choice perseveration after negative feedback (lose-stay behaviour) did not significantly differ by paranoia group, but choice switching after positive feedback (win-switch behaviour) was elevated in high paranoia (block 1: F(1)=7.117, P=0.008, η_p^2=0.023; block 2: F(1)=9.918, P=0.002, η_p^2=0.032; Table 2).

Experiment 3. To translate across species, we performed a new analysis of published data from rats exposed to chronic methamphetamine[32]. Rats chose between three operant chamber noseports with differing probabilities of sucrose reward (70%, 30%, and 10%; Fig.1 d and e). Contingencies switched between the 70% and 10% noseports after selection of the highest reinforced option in 21 out of 30 consecutive trials (Fig. 1e). Rats were tested for 26 within-session reversal blocks (Pre-Rx, n=10 per group), administered saline or methamphetamine according to a 23-day schedule mimicking the escalating doses and frequencies of chronic human methamphetamine users[32], and tested once per week for four weeks following completion of the drug regimen (Post-Rx; n=10 saline, 7 methamphetamine)[32]. Relative to rats exposed to saline, those rats exposed to methamphetamine exhibited increased win-switch behaviour and perseveration after negative feedback[32].

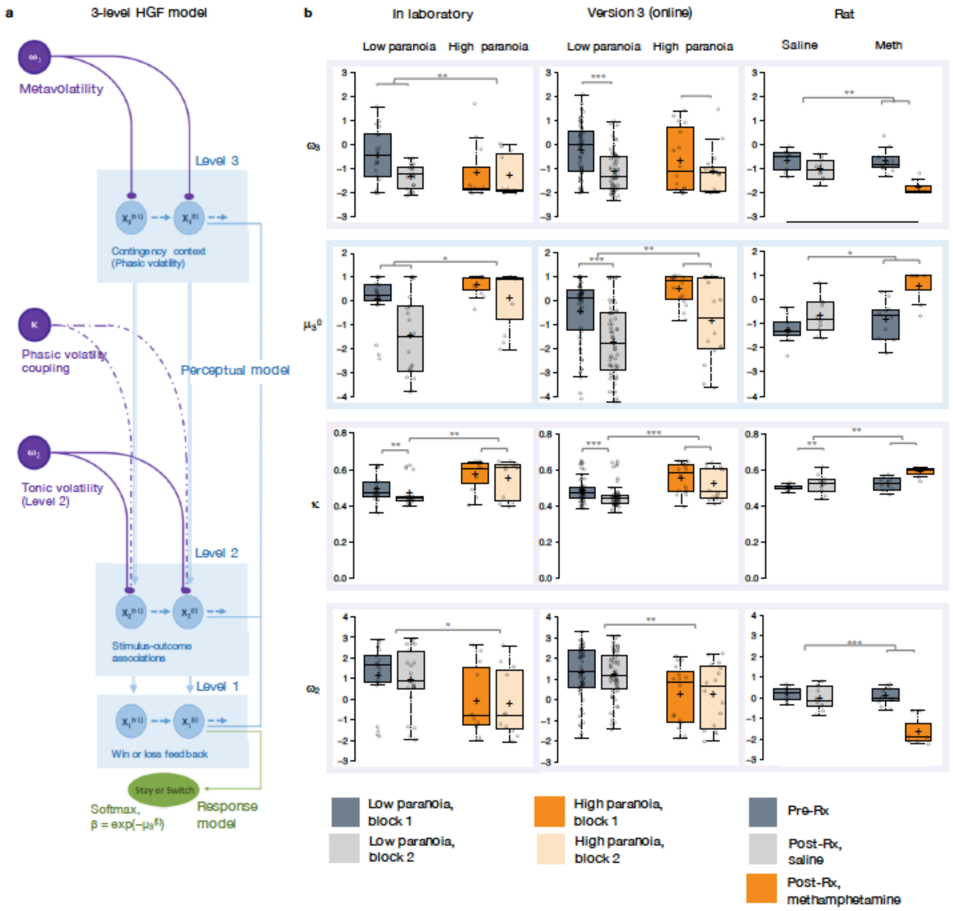


Fig. 2. Hierarchical Gaussian Filter (HGF) model parameters. a, 3-level HGF perceptual model (blue) with a softmax decision model (green). Level 1 (x_1) corresponds to trial-by-trial perception of win or loss feedback. Level 2 (x_2) represents stimulus-outcome associations (e.g., deck values). Level 3 (x_3) models perception of the overall reward contingency context. The impact of phasic volatility upon x_2 is captured by κ, the coupling parameter. Tonic volatility modulates x_3 and x_2 via ω_3 and ω_2, respectively. μ_3^0 is the initial value of the third level volatility belief. **b,** Estimated parameters replicate across high paranoia groups (orange) in the in laboratory experiment (n=21 low paranoia, 11 high paranoia); analogous online task (version 3, n=56 low paranoia, 16 high paranoia); and rats exposed to chronic, escalating methamphetamine (n=10 per group Pre-Rx; n=10 saline, 7 methamphetamine Post-Rx). Error bars denote standard error (SEM); *p ≤ 0.05, **p ≤ 0.01, ***p ≤ 0.001.

Computational modelling. We employed hierarchical Gaussian filter (HGF) modelling to compare belief updating across individuals with low and high paranoia, as well as across human participants and rats exposed to methamphetamine (Table 3). We paired a three-level perceptual model with a softmax decision model dependent upon third level volatility (Fig. 2a). We inverted the model from subject data (trial-by-trial choices and feedback) to estimate parameters for each individual (Fig. 2b). Level 1 (x_1) characterizes trial-by-trial perception of task feedback (win or loss in humans, reward or no reward in rats), Level 2 (x_2) distinguishes stimulus-outcome associations (deck or noseport values), and Level 3 (x_3) renders perception of the overall reward contingency context (i.e., volatility or variance of the deck or noseport values). Belief trajectories were unique to each subject due to the probabilistic, performance-dependent nature of the task, but we estimated initial beliefs (priors) for x_2 and x_3 (μ_2^0 and μ_3^0, respectively). We also estimated $\omega_{2,}$ the contribution of tonic (expected) volatility on learning stimulus-outcome associations, and κ, the coupling or impact of phasic (unexpected) volatility (x_3) on the x_2 belief trajectory. In the setting of our tasks, these two parameters best capture the effects of expected and unexpected uncertainty in updating stimulus-outcome associations. Higher coupling (κ) implies faster belief updating in response to perceived change at the level above, whereas lower or more negative values suggest slower updating. Diminished ω_2 indicates more rigid beliefs about the underlying risk or probability of each option. The third parameter, $\omega_{3,}$ characterizes perception of 'metavolatility,' the tonic volatility of the volatility itself (i.e., how stable the changes in underlying contingencies of the decks might be)[36]. The lower ω_3 is, the slower a subject is to update beliefs about contextual volatility.

Priors did not differ between groups at x_2 (Supplementary Table 1) but paranoid individuals and rats exposed to methamphetamine exhibited elevated μ_3^0, an initial perception of higher contingency context volatility (Fig. 2b, blue). In Experiment 1, we observed an interaction between task block and paranoia group ($F(1)=5.344$, $P=0.028$, $\eta_p^2=0.151$; Table 1). μ_3^0 differed between high and low paranoia in both blocks (block 1, $F(1)=4.232$, $P=0.048$, $\eta_p^2=0.124$, MD=0.658, CI=[0.005,1.312]; block 2, F(1)=7.497, $P=0.010$, $\eta_p^2=0.20$, MD=1.598, CI=[0.406, 2.789]), but only low paranoia subjects significantly updated their priors between block 1 and block 2 ($F(30)=39.841$, $P=5.85E-07$, $\eta_p^2=0.570$, MD=1.504, CI=[1.017, 1.99]). In Experiment 2, the analogous task design (version 3) demonstrated significant effects of block ($F(1)=64.652$, $P=1.54E-11$, $\eta_p^2=0.480$, MD=1.303, CI=[0.980,1.627]) and paranoia ($F(1)=6.366$, $P=0.014$, $\eta_p^2=0.083$, MD=0.909, CI=[0.191, 1.628]; Table 1). Rats showed a similar effect following methamphetamine exposure with a significant time (Pre-Rx, Post-Rx) by treatment (methamphetamine, saline) interaction ($F(1)=5.159$, $P=0.038$, $\eta_p^2=0.256$; pre versus post methamphetamine effect: $F(15)=12.186$, $P=0.003$, MD=1.265, CI=[-0.493, 2.037]; Pre-Rx mean [standard error]= -1.25 [0.56] saline, -0.77 [0.80] methamphetamine; Post-Rx: m=-0.69 [0.74] saline, 0.58 [0.73] methamphetamine). Random effects meta-analyses confirmed significant cross-experiment replication of elevated μ_3^0 in human participants with paranoia (in laboratory and online version 3; MD_{META}= 1.110, CI=[0.927, 1.292], z_{META}=11.929 , p=8.356E-33) and across humans with paranoia and rats exposed to methamphetamine (MD_{META}=2.090, CI=[0.123, 4.056], z_{META}=2.083, p=0.037).

Paranoid participants and methamphetamine exposed rats updated stimulus-outcome associations more strongly in response to perceived phasic volatility (e.g., correctly or incorrectly inferred reversals; Fig. 2b). κ showed significant paranoia group and block effects across the in laboratory experiment and online version 3 (Table 1; paranoia effects, in laboratory: $F(1)=7.599$, $P=0.010$, $\eta_p^2=0.202$, MD=0.081, CI=[0.021, 0.140]; online version 3: $F(1)=13.521$, $P=0.0005$, $\eta_p^2=0.162$, MD=0.068, CI=[0.031-0.104]; MD_{META} = 0.079, CI=[0.063, 0.095], z_{META}=9.502 p=2.067E-21); see Supplementary Table 1 for block effects). κ increased from baseline in rats on methamphetamine, yielding significant effects of treatment ($F(1)=13.356$, $P=0.002$, $\eta_p^2=0.471$, MD=0.045, CI=[0.019, 0.072]) and time ($F(1)=9.132$, $P=0.009$, $\eta_p^2=0.378$, MD=0.041, CI=[0.012, 0.069]);

however, the interaction between time and treatment did not reach statistical significance (Supplementary Table 1; Pre-Rx *m*=0.499 [0.015] saline, 0.523 [0.040] methamphetamine; Post-Rx: *m*=0.518 [0.053] saline, 0.585 [0.029] methamphetamine). Replication of group effects was significant across all three experiments (MD_{META}=2.063, CI=[0.341, 3.785], z_{META}=2.348, p=0.019).

Tonic volatility and metavolatility ($\omega_{2,}\ \omega_3$) were decreased in paranoid participants and rats exposed to methamphetamine (Fig. 2b). In laboratory and online (version 3), paranoid individuals were slower to update stimulus-outcome associations in response to expected stochastic variance within the contingency context (Table 1; ω_2 paranoia effect, in laboratory: *F*(1)=4.186, *P*=0.050, η_p^2=0.122, MD=-1.188, CI=[-2.375, -0.002]; online version 3: *F*(1)=8.7, *P*=0.004, η_p^2=0.111, MD=-0.993, CI=[-1.665, -0.322]; MD_{META}=-1.154 , CI=[-1.455, -0.853], z_{META}=-7.521, p=5.450E-14). The effects of methamphetamine exposure in rats were consistent (MD_{META}=-1.992 , CI=[-3.318, -0.665], z_{META}=-2.943, p=0.003) yet more striking, with a strongly negative ω_2 accounting for the more pronounced lose-stay behaviour in rats (time by treatment interaction, *F*(1)=18.454, *P*=0.001, η_p^2=0.552; pre versus post methamphetamine: *F*(1)=42.242, *P*=1.0E-5[32], η_p^2=0.738, MD=-1.604, CI=[-2.130, -1.078]; Pre-Rx *m*=0.198 [0.33] saline, -0.036 [0.42] methamphetamine; Post-Rx: *m*=-0.023 [0.56] saline, -1.640 [0.71] methamphetamine). Metavolatility (ω_3) was similarly lower across paranoia and methamphetamine exposed groups (in laboratory, online version 3, and rats: MD_{META}=-1.155, CI=[-2.139, -0.171], z_{META}=-2.3, p=0.021), suggesting resistance to updating beliefs about the overall contingency context. In laboratory, we observed a block by paranoia group interaction (Table 1, *F*(1)=6.948, *P*=0.010, η_p^2=0.188). Post-hoc tests differentiated first and second blocks for the low paranoia group only (*F*(1)=26.640, *P*=1.5E-5, η_p^2=0.470, MD=-0.876, CI=[-1.222, -0.529]). The paranoia effect did not reach statistical significance for online version 3 (block effect only, *F*(1)=14.932, *P*=0.0002, η_p^2=0.176, MD=-0.692, CI=[-1.050, -0.335]; Supplementary Table 1), but meta-analytic random effects analysis confirms a significant paranoia group difference (in laboratory and online version 3: MD_{META}=-0.341, CI=[-0.522, -0.159], z_{META}=-3.68, p=0.0002). Methamphetamine exposure decreased ω_3 in rats (time by treatment interaction, (*F*(1)=9.058, *P*=0.009, η_p^2=0.376; pre versus post methamphetamine: *F*(1)=30.668, P=5.7E-5, η_p^2=0.672, MD=-1.210, CI=[-1.676, -0.745]; Pre-Rx m=-0.692 [0.44] saline, -0.607 [0.51] methamphetamine; Post-Rx: *m*=-1.044 [0.44] saline, -1.817 [0.32] methamphetamine).

We applied False Discovery Rate (FDR) corrections for modelling parameters. κ group effects survived corrections within each experiment (Supplementary Table 2). In addition to κ, μ_3^0 survived for experiment 1; μ_3^0 and ω_2 survived in online version 3; and μ_3^0, ω_2, and ω_3 survived in experiment 3 as group effects. Such correction is not yet standard practice with this modelling approach[36-38] but we believe it should be, and when effects survive correction we should increase our confidence in them.

Paranoia effects across task versions. To examine the relationship between contingency context change and paranoia within our HGF parameters, we performed split-plot, repeated measures ANOVAs across all four task versions (Experiment 2; paranoia group and task version between subject factors). Paranoia group effects were specific to versions of the task in which we explicitly manipulated uncertainty via context change (Fig. 3, Supplementary Table 2). Specifically, we observed paranoia by version interactions for κ ($F(3)=4.178$, $P=0.006$, $\eta_p^2=0.040$) and ω_2 ($F(3)=2.809$, $P=0.040$, $\eta_p^2=0.027$; Table 2). Post-hoc tests confirmed that significant paranoia group effects were restricted to version 3 (κ: $F(1)=12.230$, $P=0.001$, $\eta_p^2=0.039$, MD=0.068, CI=[0.03,0.106]; ω_2: $F(1)=8.734$, $P=0.003$, $\eta_p^2=0.028$, MD=-0.993, CI=[-1.655, -0.332]) and a trend for version 4 (ω_2: $F(1)=2.909$, $P=0.089$, $\eta_p^2=0.010$, MD=-0.528, CI=[-1.138, 0.081], Fig. 3a). μ_3^0 also exhibited a paranoia by version trend (Table 2, $F(3)=2.329$, $P=0.075$, $\eta_p^2=0.023$), largely driven by version 3 ($F(1)=6.206$, $P=0.013$, $\eta_p^2=0.020$, MD=0.909, CI=[0.191, 1.628]; Fig. 3a). There were no significant paranoia effects or interactions for ω_3 (Supplementary Table 2).

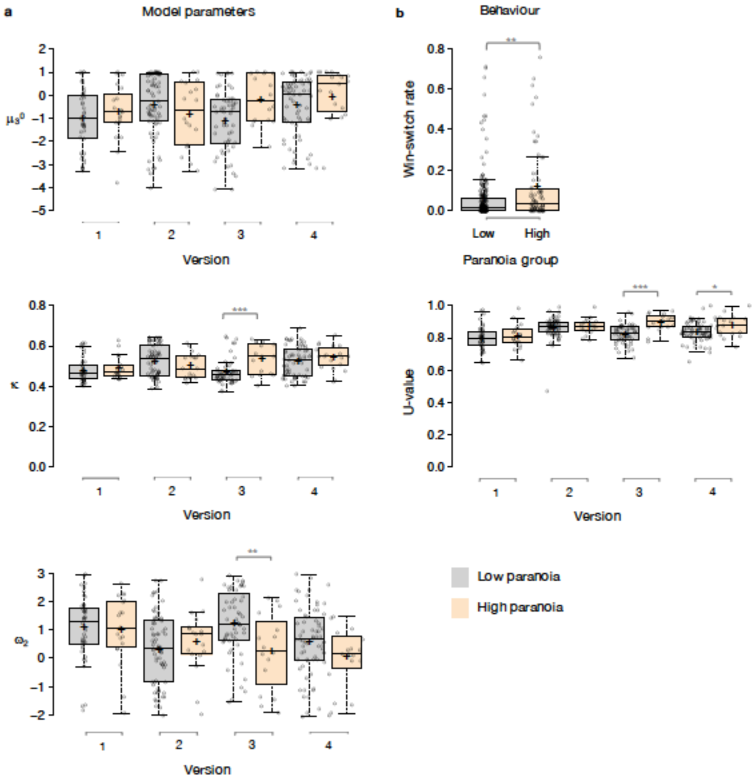


Fig. 3. Paranoia effects across task versions. a, HGF parameters μ_3^0, κ, and ω_2 show version 3 specific trends and effects of paranoia group membership (Experiment 2, n=234 low paranoia, 73 high paranoia, collapsed across task versions). **b,** Behaviourally, paranoid participants switched between decks more frequently after positive feedback, across all task versions and blocks (paranoia group effect; version trend, p=0.081). In versions 3 and 4 only, paranoid participants showed higher U-values, suggesting increasingly stochastic switching rather than perseverative returns to a previously rewarding option. Error bars denote standard error (SEM); *p ≤ 0.05, **p ≤ 0.01, ***p ≤ 0.001.

Covariate analyses. We completed three ANCOVAs for each HGF parameter derived from Experiment 2: demographics (age, gender, ethnicity, and race); mental health factors (medication usage, diagnostic category, BAI score, and BDI score); and metrics and correlates of global cognitive ability (educational attainment, income, and cognitive reflection; Supplementary Tables 3 and 4). For κ, our metric of unexpected uncertainty, the paranoia by version interaction remained robust across all three ANCOVAs (demographics: $F(3)=3.753$, $P=0.011$, $\eta_p^2=0.037$; mental health: $F(3)=4.417$, $P=0.005$, $\eta_p^2=0.049$; cognitive: $F(3)=4.304$, $P=0.005$ $\eta_p^2=0.043$). The paranoia by version trend of μ_3^0 diminished with inclusion of demographic, mental health, and cognitive covariates (demographic: $F(3)=1.997$, $P=0.119$, $\eta_p^2=0.020$; mental health: $F(3)=1.942$, $P=0.123$, $\eta_p^2=0.022$; cognitive: $F(3)=2.193$, $P=0.089$, $\eta_p^2=0.022$). The paranoia by version interaction for ω_2 was robust to mental health and cognitive factors ($F(3)=3.617$, $P=0.014$, $\eta_p^2=0.041$; $F(3)=3.017$, $P=0.030$, $\eta_p^2=0.030$). A paranoia group effect and paranoia by version trend remained with inclusion of demographics (ω_2 , paranoia effect: $F(1)=4.275$, $P=0.040$, $\eta_p^2=0.014$; interaction: $F(3)=2.507$, $P=0.059$, $\eta_p^2=0.025$).

Multiple regression. We examined the effects of paranoia, anxiety, and depression on κ within the online version 3 dataset by multiple regression analysis. A significant regression equation was found (F(3,68)=3.681, p=0.016), with an R^2 of 0.140 (Supplementary Fig.1). Participants' predicted κ equalled 0.486 + 0.062 (PARANOIA)+0.012 (BDI) -0.006 (BAI). Paranoia was a significant predictor of κ (β=0.343, t=2.470, p=0.016, CI=[0.012, 0.113]) but depression and anxiety were not (BDI: β=0.086, t=0.423, p=0.674, CI=[-0.043, 0.066]; BAI: β=-0.043, t=-0.218, p=0.828, CI=[-0.063, 0.050]). Examination of correlation plots for κ (Supplementary Fig. 2) revealed a much stronger relationship when analyses were restricted to individuals with paranoia scores greater than 0 (i.e., endorsement of at least one item); among participants who denied all questionnaire items, a minority (seven out of 33) exhibited elevated κ. To account for the possibility that some individuals with severe paranoia may avoid disclosing sensitive information, we performed additional analyses of participants who endorsed one or more paranoia item. The correlation between paranoia and κ in the first block of the task increases from r=0.3, p=0.011, CI=[0.074, 0.497] (all participants, n=72) to r=0.588, p=8.0E-5, CI=[0.335, 0.762] (participants with paranoia > 0, n=39). In this subset, a significant regression equation was also found (F(3,35)=6.322, p=0.002), with an R^2 of 0.351 (Supplementary Fig.1). Participants' predicted κ was equal to 0.432 + 0.150 (PARANOIA)+0.013 (BDI) -0.004 (BAI). Paranoia was a significant predictor of κ (β=0.538, t=2.983, p=0.005, CI=[0.048, 0.252]) but depression and anxiety were not (BDI: β=0.111, t=0.494, p=0.624, CI=[-0.041, 0.067]; BAI: β=-0.035, t=-0.163, p=0.872, CI=[-0.057, 0.049]).

Behaviour and simulations. Win-switching was the prominent behavioural feature of both paranoid participants and rats exposed to methamphetamine (Table 1, Table 2,[32]). Collapsed across blocks and task versions, our Experiment 2 data demonstrated a main effect of paranoia group (Fig. 3b; *F*(1)=9.207, *P*=0.003, η_p^2=0.030, MD=0.059, CI=[0.021, 0.097]; version trend: *F*(3)=2.263 *P*=0.081, η_p^2=0.022; low paranoia: *m*=0.06 [0.01], high paranoia: *m*=0.12 [0.02]). To elucidate whether this behaviour was stochastic or predictable (e.g., switching back to a previously rewarding option), we calculated U-values[39], a metric of behavioural variability employed by behavioural ecologists (increasingly an inspiration for human behavioural analysis[40]), particularly with regards to predator-prey relationships[41]. When a predator is approaching a prey animal, the prey's best course of action is to behave randomly, or in a *protean* fashion, in order to evade capture[41]. The more protean or stochastic the behaviour, the closer to the U-value is to 1. Across task blocks, paranoid participants exhibited elevated choice stochasticity (paranoia by version interaction, *F*(3)=3.438, *P*=0.017, η_p^2=0.033; Table 2). Post-hoc tests indicate that this stochasticity was specific to versions with contingency context change, suggesting a relationship to unexpected uncertainty (Fig. 3b; version 3, *F*(1)=17.585, *P*=3.6E-5, η_p^2=0.056, MD=0.071, CI=[0.038, 0.104]; version 4, *F*(1)=6.397, *P*=0.012, η_p^2=0.021, MD=0.039, CI=[0.009, 0.07]).

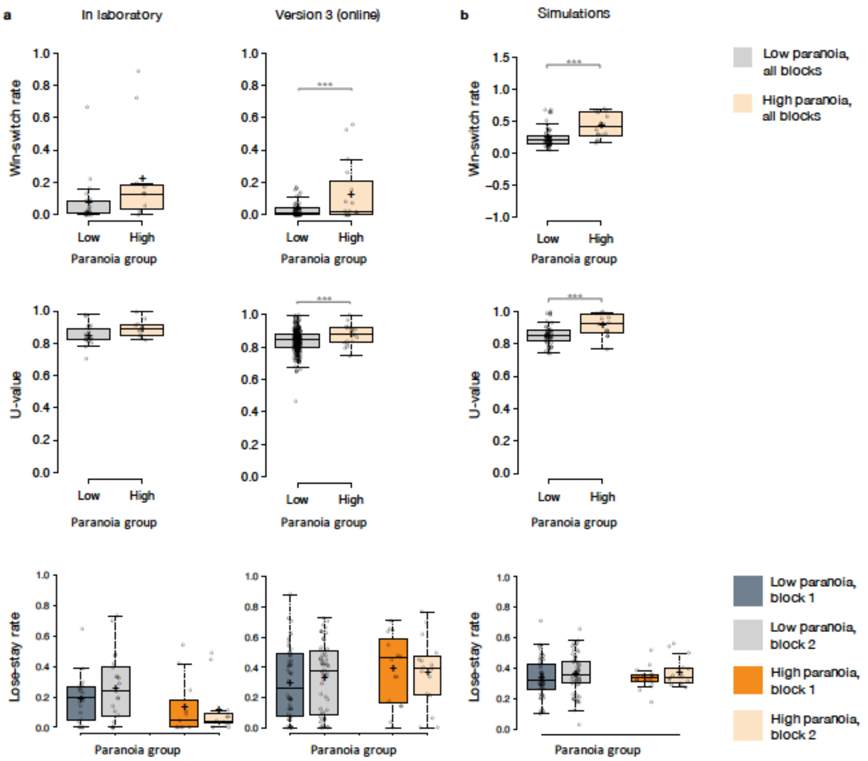


Fig. 4. Behavioural data and simulations. a, Behavioural switching patterns replicate across in laboratory and online version 3 experiments (win-switch: in laboratory paranoia group trend, p=0.068; version 3 paranoia effect; U-value: in laboratory paranoia group trend, p=0.079; version 3 paranoia effect). Perseveration after negative feedback (lose-stay behaviour) did not significantly differ between paranoia groups or task block. **b,** Simulated data generated from HGF perceptual parameters (version 3) replicates win-switch and U-value behaviours (win-switch paranoia effect; U-value paranoia effect). Ten simulations were performed per subject. Rates and U-values were averaged across simulations. Error bars denote standard error (SEM); n=21 low paranoia, 11 high paranoia (in laboratory); n=56 low paranoia, 16 high paranoia (online, version 3); $^*p \leq 0.05$, $^{**}p \leq 0.01$, $^{***}p \leq 0.001$.

To test the propriety of our model, we simulated data for each subject in online version 3 and determined whether or not key behavioural effects (Fig. 4a, Table 1, Supplementary Table 5) were present. Using individually estimated HGF parameters to generate ten simulations per participant, we recapitulated both elevated win-switch behaviour (paranoia effect, $F(1)$=15.394, P=2.01E-4, η_p^2=0.180, MD=0.186, CI=[0.091, 0.28]) and choice stochasticity (U-value; paranoia effect, F(1)=13.362, P=0.0005, η_p^2=0.160, MD=0.065, CI=[0.030, 0.101]) in simulated paranoid participants (Fig. 4b; simulated win-switch rate, low paranoia: m=0.24 [0.02], high paranoia: m=0.43 [0.04]; simulated U-value, low paranoia: m=0.851 [0.008], high paranoia: m=0.916 [0.016]). Neither real nor simulated data showed any significant relationship between lose-stay behaviour and paranoia (Table 1, Table 2, Supplementary Table 5). To demonstrate the effects of parameters on task performance, we performed additional simulations in which we doubled or halved a single parameter at a time from the baseline average of low paranoia participants. These results confirmed the impact of κ, ω_2, and ω_3 on win-shift behaviour (Supplementary Fig. 3). Parameter recovery revealed significant correlations for κ and ω_2 between original subject parameters and those estimated from simulations (Supplementary Fig. 4; ω:

r=0.702, p=2.52E-11, CI=[0.557, 0.805]; κ: r=0.305, p=0.011, CI=[0.072, 0.506]). Higher level parameters (ω_3, μ_3^0) were less consistently recovered, as noted in previous publications[42].

Clustering analysis. Given the apparent similarity in effects of paranoia and methamphetamine in humans and rats, respectively (Fig. 2b), we formally tested for latent structure in our data using two-step cluster analysis[43]. This approach automatically determines the optimal number of clusters. We analysed μ_3^0, κ, ω_2, and ω_3 estimates derived from the first block of experiment 1 and online version 3 (pre-context change data, because rats do not experience a context shift) with Post-Rx rat data. We identified two clusters with good cohesion and separation (average silhouette coefficient=0.7; cluster size ratio=2.46; Fig. 5a). All parameters contributed to clustering; κ contributed most strongly (Fig. 5b). Relative to the overall distribution, Cluster 1 was characterized by high κ and μ_3^0, and decreased ω_2 and ω_3. Cluster 2 parameters fell close to the overall distribution median, with κ and μ_3^0 scores lower than Cluster 1, ω_2 and ω_3 higher (Fig. 5a). Cluster 1 membership was significantly associated with high paranoia and methamphetamine exposure, χ^2(1, *n*=121)=29.447, *P*=5.75E-8, Cramer's V=0.493 (Fig. 5c). Notably, no participants in the low paranoia group with paranoia scores above zero were ascribed Cluster 1 membership. The cluster solution was robust to validation by split-half analysis, removal of the rat subjects, and removal of human participants (Supplementary Fig. 5).

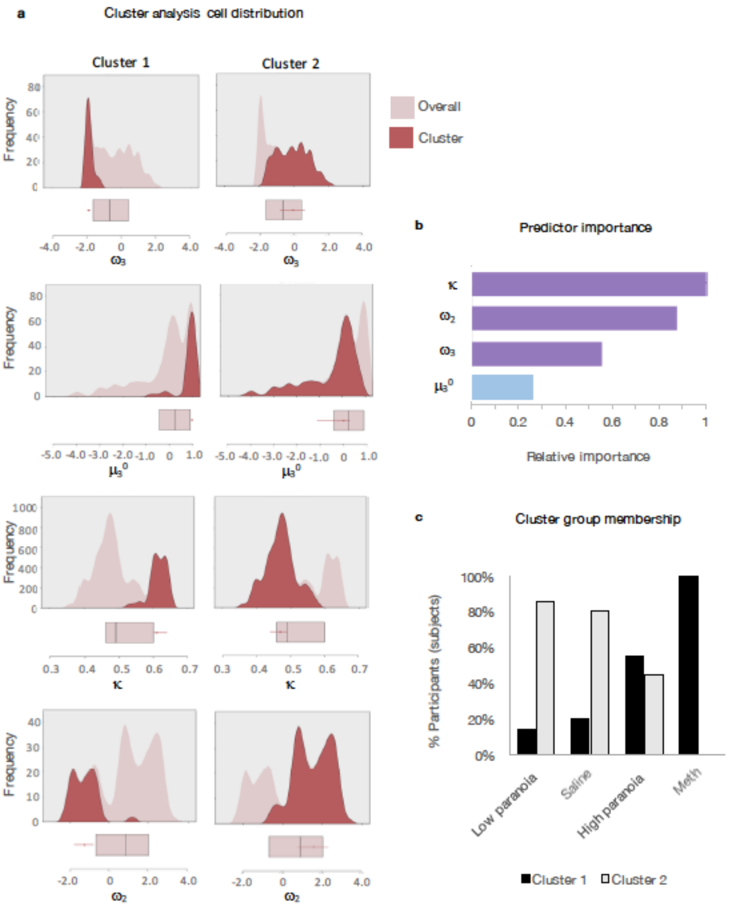


Fig. 5. Cluster analysis of HGF parameters. Two-step cluster analysis of model parameters across rat and human data sets (rat, post-Rx; in laboratory and online version 3, block 1). Automated clustering yielded an optimal two clusters with good cohesion and separation (average silhouette coefficient=0.7; cluster size ratio=2.46). **a,** Parameter density plots for overall distributions (light pink) and cluster-specific distributions (red). Box-plots of overall median, 25th quartile, and 75th quartile are aligned below each plot (pink), with cluster medians and quartiles superimposed (red). Relative to the overall distribution, Cluster 1 (n=35) medians are elevated for μ_3^0 and κ, decreased for ω_2 and ω_3. Cluster 2 (n=86) falls within each overall distribution. **b,** Predictor importance of included parameters. **c,** Distribution of cluster identities within groups (low paranoia: n=77; high paranoia: n=27; rat-saline: n=10; rat-methamphetamine: n=7). Cluster 1 membership is significantly associated with paranoia and methamphetamine groups (χ^2(1, n=121)=29.447, p=5.75E-8).

Discussion

We have shown that inferential learning differs in paranoid individuals according to a specific pattern of beliefs about contextual volatility and response to uncertainty. During probabilistic reversal-learning with three options, paranoid individuals and rats chronically exposed to methamphetamine have higher initial expectations of task volatility (μ_3^0). In other words, they start the task anticipating more changes in stimulus-outcome associations. These same subjects respond more strongly to perceived volatility in updating stimulus-outcome associations (κ). This coincides with decreased ω_2, reflecting more stable beliefs about the underlying probability values themselves. Consequently, inferred contextual volatility manifests in excessive perception of reversal events, increasing behavioural switching.

We replicated an elevated prior on environmental volatility (μ_3^0) and higher sensitivity to this volatility (κ) previously observed in HGF analyses of 2-choice probabilistic reversal-learning in medicated and unmedicated patients with schizophrenia[44]. Unlike prior work, we assessed the trans-diagnostic symptom of paranoia across the continuum of health and illness, provided three choice options to differentiate stochastic switching from perseverative returns, and explicitly manipulated unexpected uncertainty across task versions. The version that shifts from an easier to discern contingency context to a more difficult context was associated with paranoia group differences in μ_3^0, κ, and ω_2. Furthermore, this context change elicited decreases in metavolatility learning (ω_3) among low paranoia controls relative to their first block baseline, rendering them more similar to high paranoia participants. Paranoid individuals behave as if that the world is always more volatile, demanding continual updating of associations. Low paranoia individuals behave similarly under more difficult, uncertain conditions. Although our domain-general paradigm lacks any sizable, tangible threat, uncertainty itself may be aversive[45], threatening the brain's ability to make clear predictions about future states and actions.

Unexpected uncertainty, the perception of change in the probabilities of the environment — particularly "unsignaled context switches"[25] — is thought to promote abandonment of old associations and new learning. Our analysis of covariates warrants specific focus on κ, the sensitivity to unexpected uncertainty. Other parameter-paranoia associations did not endure after controlling for demographic factors (age, gender, ethnicity, and race). These factors are strong predictors of paranoia[46-48]. It is notable too that κ was the most powerful discriminator of the two clusters of human and animal participants. However, the rodent data are less impacted by concerns about covariates, and there, the other parameters (prior on volatility, expected uncertainty, meta-volatility) were all changed by methamphetamine. We conclude that κ is the parameter most robustly associated with paranoia.

Multiple neurobiological manipulations may induce win-switching behaviour. Lesions of the mediodorsal thalamus in non-human primates[49] or neurons projecting from the amygdala to orbitofrontal cortex in rats[50] engender win-switching. The human hippocampus appears to be sensitive to volatility during belief learning, as are the anterior cingulate cortex and insula[27]. However, unexpected uncertainty, and the κ parameter of the HGF in particular[51], are thought to be signalled via the locus coeruleus and noradrenaline (i.e., neural gain)[25-28]. This mechanism is thought to coordinate rapid shifts in cortical networks through patterns of widespread norepinephrine release, modulating exploratory versus exploitative behaviours (i.e., switching and staying)[52-55] and responding to stress[56-58], unexpected uncertainty[25],[27] and subliminal fear cues[59] to coordinate fight-or-flight responses[58]. In fact, visual fear stimuli presented below the threshold of conscious perception activate the locus coeruleus, amygdala, and fronto-temporal orienting regions, suggesting a neural ‘alarm’ system for rapid threat detection[59]. The dual role of the locus coeruleus in recognizing and responding to threats as well as unexpected uncertainty suggests that dysfunction could produce both paranoia and the inferential abnormalities we observed. Methamphetamine may induce similar dysfunction. Acute moderate doses increase pre-synaptic catecholamine release, particularly noradrenaline[60], and induce exploratory locomotive effects modulated through adrenoceptors on dopamine neurons[61]. Unlike acute binge paradigms, the schedule of methamphetamine administration completed by the rats in these analyses preserves methamphetamine-induced locomotor hyperactivity[32,62].

Perturbations of noradrenergic gain impede new learning while appearing falsely to enhance behavioural flexibility. In rats, excessive release of noradrenaline from the locus coeruleus into the anterior cingulate cortex drives disengagement from model-consistent performance in a three-option counter prediction task[28]. This manipulation includes stochastic switching and insensitivity to feedback or context change— a type of behavioural “flexibility” that is ultimately inflexible. Our data suggest that in paranoia, increased gain under uncertainty may similarly shunt away incoming information, leaving only reflexive, habitual responses. Although participants engage in choice switching— in an increasingly stochastic fashion— our cluster analyses show that excessive κ is associated with diminished metavolatility learning, rendering these subjects less flexible in updating context beliefs. In this fashion, excessive switching behaviour may be indicative of fixed higher-level beliefs.

Disengagement from model-congruent behaviour has been observed in paranoia and psychosis[63,64]. Evolutionarily, departure from predictable, rational modes of behaviour might offer an adaptive mechanism for escape from intractable threat. As a protean defence mechanism, behavioural stochasticity impedes predators' abilities to create accurate, actionable countermeasures[41,65,66]. If driven by excessive noradrenergic gain, protean defence may represent an extreme state along a heavily conserved, continuous common mechanism underlying vigilance and false alarms[67-69], arousal-linked attentional biases and selective processing (i.e., focusing on narrow, most salient features versus broader context[70]; attending to and learning from predisposition-conforming features[55]), and behavioural and cognitive flexibility in response to unexpected uncertainty and Bayesian surprise (i.e., prediction error)[53,54,71]. We hypothesize that individuals with stable, trait-level paranoia, rather than having specific deficits in inferring others' reputations[16], exhibit disturbances across the domains of behavioural flexibility and stochasticity, false alarms and attentional bias, and inferential response to unexpected uncertainty. Our data suggest that that these perturbations exist outside of social settings and may be elicited in nonhuman models. We propose that protean defence and its attendant behavioural stochasticity might be one useful translational marker of paranoia.

We conclude that this model provides a robust tool for computational dissection of learning mechanisms across species. Social interactions play a rich and undeniable role, but translational, domain-general approaches may ultimately facilitate biological insights into paranoia, psychosis and delusions[72,73]. Whilst we contend that our task is relatively free of social features (certainly compared to others[15]), the possibility remains that the elevated U-values in our participants are reflective of attempts (and perhaps failures) to predict our intentions as experimenters. Indeed, this is a possibility raised previously with regards to simple conditioned behaviours in experimental animals. Even during Pavlovian conditioning, animals may attempt to infer a generative model of the task environment, which might, ultimately, include the experimenter arranging the contingencies[74,75]. It is possible that all instances of human cognitive testing involve an element of inference by the participant with regards to the intentions of the experimenter, whether or not the task at hand is explicitly social, and indeed, all cognitive functions may be aimed at or modulated by such inferences[76].

In summary, a strong belief in the volatility of the world necessitates hypervigilance and a facility with change. However, in paranoia, that belief (in the volatility of the world) is itself resistant to change, making it difficult to reassure, teach, or change the minds of people who are paranoid. They remain "on guard" even under stable conditions. Whether promoting recovery from paranoia-associated illness[77] or interpersonal collaboration, our domain-general approach reaffirms the merit of trying to establish stable, predictable environments. We note with interest the apparent relationship between conspiratorial ideation and societal crisis situations (terrorist attacks, plane crashes, natural disasters or war) throughout history, with peaks around the great fire of Rome (AD 64), the industrial revolution, the beginning of the cold war, 9/11, and contemporary financial crises[78]. Perhaps these broader trends are a macrocosmic version of the unexpected uncertainty manipulation that drove promiscuous switching behaviour in our task, particularly in high paranoia participants. Rather than proving adaptive, their behaviour ultimately increases the noise of their task experience with sampling of sub-optimal options and exposure to misleading positive feedback. In today's world of escalating uncertainty and volatilty – particularly environmental climate change – our findings suggest that the paranoid style of inference may prove particularly maladaptive for coordinating collaboratve solutions.

Methods

Experiments were conducted at Yale University and the Connecticut Mental Health Center (New Haven, CT) in strict accordance with Yale University's Human Investigation Committee and Institutional Animal Care and Use Committee. Informed consent was provided by all research participants.

Experiment 1. English-speaking participants aged 18 to 65 (*n*=34) were recruited from the greater New Haven area through public fliers and mental health provider referrals. Exclusion criteria included history of cognitive or neurologic disorder (e.g., dementia), intellectual impairment, or epilepsy; current substance dependence or intoxication; cognition-impairing medications or doses (e.g. opiates, high dose benzodiazepines); history of special education; and colour blindness. Participants were classified as healthy controls (*n*=18), schizophrenia spectrum patients (schizophrenia or schizoaffective disorder; *n*=8), and mood disorder patients (depression, bipolar disorder, generalized anxiety disorder, post-traumatic stress disorder; *n*=8) on the basis of clinician referrals and/or self-report. Participants were compensated $10 for enrolment with an additional $10 upon completion. Two healthy controls were excluded from analyses due to failure to complete the questionnaires and suspected substance use, respectively.

Experiment 2. 332 participants were recruited online via Amazon Mechanical Turk (MTurk). The study advertisement was accessible to MTurk workers with a 90% or higher HIT approval rate located within the United States. To discourage bot submissions and verify human participation, we required participants to answer open-ended free response questions; submit unique, separate completion codes for the behavioural task and questionnaires; and enter MTurk IDs into specific boxes within the questionnaires. All submissions were reviewed for completion code accuracy, completeness of responses (i.e., declining no more than 30% of questionnaire items), quality of free response items (e.g., length, appropriate grammar and content), and use of virtual private servers (VPS) to submit multiple responses and/or conceal non-US locations (Dennis VPS paper, 2018). Upon approval, workers were compensated $6. Those who scored in the top 25% on the card game (reversal-learning task) earned a $2 bonus. We rejected or excluded 19 submissions that geolocation services (https://www.iplocation.net/) identified as originating outside of the United States or from suspected server farms, 4 submissions for failure to manually enter MTurk ID codes, and 2 submissions for insufficient questionnaire completion. Submissions with grossly incorrect completion codes were rejected without further review.

Experiment 3. Subject information, behavioural data acquisition, and behavioural analyses were described previously [32]. Long Evans rats (Charles River; *n*=20) ranged from 7 to 9 weeks of age. Rats were exposed to escalating doses and frequency of saline (*n*=10) or methamphetamine (*n*=10, 3 withdrawn during dosing), imitating patterns of human methamphetamine users[62,79]. Prior to dosing (Pre-Rx), rats completed 26 within-session reversal sessions, including up to 8 reversals per session. Post-dosing (Post-Rx), rats completed one test session per week for four weeks. Computational model parameters were estimated from each session and averaged across treatment conditions to yield one Pre-Rx and Post-Rx set of parameters per rat.

Behavioural task. Participants completed a 3-option probabilistic reversal-learning paradigm. Three decks of cards were displayed on a computer monitor for 160 trials. Participants selected a deck on each trial by pressing the predesignated key. We advised participants that each deck contained winning and losing cards (+100 and -50 points), but in different amounts. We also stated that the best deck may change. Participants were instructed to find the best deck and earn as many points as possible. Probabilities switched between decks when the highest probability deck was selected in 9 out of 10 consecutive trials (performance-dependent reversal). Every 40 trials the participant was provided a break, following which probabilities automatically reassigned (performance-independent reversal).

In Experiment 1, the task was presented via Eprime® 2.0 software (Psychology Software Tools, Sharpsburg, PA). Participants were limited to a 3-second response window, after which the trial would time out and record a null response. A fixation cross appeared during variable inter-trial intervals (jittering). Task pacing remained independent of response time. In block 1 (trials 1-80) the reward probabilities (contingency context) of the three decks were 90%, 50%, and 10% (90-50-10%). Without cue or warning, the context changed to 80%, 40%, and 20% (80-40-20%) upon initiation of block 2 (trials 81-160).

In Experiment 2, the task was administered via web browser link from the MTurk marketplace. We changed the task timing to self-paced and eliminated null trials and inter-trial jittering. A progress tracker was provided every 40 trials. Workers were randomly assigned to one of four task versions. Version 1 had a constant contingency context of 90-50-10%. Version 4 maintained a constant context of 80-40-2. Version 3 replicated the 90-50-10% (block 1) to 80-40-20% (block 2) context change of Experiment 1. Version 4 presented the reversed context change, 80-40-20% (block 1) to 90-50-10% (block 2). We analysed attrition rates across the four versions.

Questionnaires. Following task completion, questionnaires were administered via the Qualtrics® survey platform (Qualtrics Labs, Inc., Provo, UT). Items included demographic information (age, gender, educational attainment, ethnicity, and race) and mental health questions (past or present diagnosis, medication use, *Structured Clinical Interview for DSM-IV Axis II Personality Disorders* (SCID-II)[33], Beck's Anxiety Inventory (BAI)[34], Beck's Depression Inventory (BDI)[35]. We removed the single suicidality question from the BDI for Experiment 2. Experiment 2 included additional items: income, three cognitive reflection questions (Supplementary Table 4), and three free response items ('What do you think the card game was testing?', 'Did you use any particular strategy or strategies? If yes, please describe', and 'Did you find yourself switching strategies over the course of the game?'). We quantified trait-level paranoia using the paranoid personality subscale of the SCID-II, and we included an ideas of reference item from the schizotypy subscale ('When you are out in public and see people talking, do you often feel that they are talking about you?') This item, along with other SCID-II items, has previously been included as a metric of paranoia in the general population[5,80]. Participants who endorsed 4 or more paranoid personality items (i.e., the cut-off for the top third identified in Experiment 1) were classified as 'high paranoia.' Each participant's SCID-II, BAI, and BDI scores were normalized by total scale items answered. Distributions of SCID-II scores are shown in Supplementary Fig. 6. Response rates were higher than 90% for all questionnaire items and scales (Supplementary Table 6).

Behavioural analysis. We analysed tendencies to choose alternative decks after positive feedback (win-switch) and select the same deck after negative feedback (lose-stay). Win-switch rates were calculated as the number of trials in which the participant switched after positive feedback divided by the number of trials in which they received positive feedback. Lose-stay rates were calculated as number of trials in which a participant persisted after negative feedback divided by total negative feedback trials. In Experiment 1, we excluded post-null trials from these analyses. To further characterize switching behaviour, we calculated U-values, a measure of choice stochasticity:

$$U{-}value = -\Sigma_{i=1}^{\beta} \frac{\log(\alpha_i)\ x\ \alpha_i}{\log(\beta)} \qquad (1)$$

where β is the number of possible choice options (i.e., card decks or noseports) and α equals the relative frequency of choice option i [39]. To avoid any choice counterbalancing effects across reversals, choice frequencies were determined by the underlying probabilities of the decks rather than their physical attributes (e.g., deck position or colour). Additional behavioural analyses included trials to first reversal, trials to post-reversal recovery, and trials to post-reversal switch. The latter two were restricted to the first reversal in the first block. Trials post-reversal were counted from the first-negative feedback trial following the true reversal event. Recovery was defined as switching to the best deck and staying for at least one additional trial.

Computational modelling and simulations. We utilized the freely available HGF toolbox v5.3.1 (https://translationalneuromodeling.github.io/tapas/) in MATLAB and Statistics Toolbox Release 2016a (MathWorks ®, Natick, MA)[30,31]. The HGF employs an "observing the observer" framework consisting of a perceptual model (a generative model of the agent's inferences about the task environment) and a decision model that reconciles the agent's actions. HGF parameter values are inferred from observed agent decisions and trial-by-trial feedback (i.e., win or loss outcomes) through variational model inversion[30,31]. Our model schema consisted of a 3-level HGF multi-arm bandit configuration for binary outcomes, paired with the softmax-mu03 decision model. The softmax mu03 model tests the hypothesis that beliefs about environmental volatility dynamically influence behaviour. The inverse decision temperature (β) is set to the inverse volatility estimate $\exp(-\mu_3^{(k)})$ where *k* denotes the current trial, permitting simultaneous estimation of μ^0, κ and ω. We inspected each subject's x_1, x_2, and x_3 trajectories and optimized the default perceptual model configuration file by changing the second level κ prior mean from log(1) to log(0.6). The first level κ remained fixed at log(1). Third level trajectory were regularized by use of the autoregressive HGF configuration option.

Perceptual parameters were estimated separately for blocks 1 and 2, with block 1 μ_2^0 and μ_3^0 comprising the μ_2 and μ_3 prior means in block 2. To evaluate the validity of our model, we subsequently simulated participant choices using trial-by-trial outcome data and estimated perceptual parameters from online version 3 participants. We performed ten simulations per subject and calculated win-shift rates, U-values, and lose-stay rates to compare with our actual data. Code for parameter estimation and simulations are detailed in the Supplementary Methods.

Statistics. Unless otherwise specified, statistical analyses and effect size calculations were performed in IBM SPSS Statistics, Version 25 (IBM Corp., Armonk, NY), with an alpha of 0.05. Box-plots were created with the web tool BoxPlotR[81]. Model parameters were corrected for multiple comparisons using the Benjamini Hochberg (False Discovery Rate) method. Bonferroni corrections were largely consistent (Supplementary Table 2)

To compare questionnaire item means between two groups (Table 1, low versus high paranoia), we conducted independent samples t-tests. To compare questionnaire item means across paranoia groups and task versions (Table 2, fixed factors), we employed univariate analyses. Associations between characteristic frequencies and subject group or task version were evaluated by Chi-Square Exact tests (two groups) or Monte Carlo tests (more than 2 groups). Pearson correlations established the associations between paranoia and BDI scores, BAI scores, win-switch rates, and κ. We selected two-tailed p-values where applicable and assumed normality. Multiple regressions were conducted with κ estimates from the first task block (dependent variable) and paranoia, BAI, and BDI scores from online version 3.

To compare HGF parameter estimates and behavioural patterns (win-switch, U-value, lose-stay) across block, paranoia group (Experiment 1, Experiment 2 version 3), and/or task version (Experiment 2), we employed repeated measures and split-plot ANOVAs (i.e., block designated within-subject factor, paranoia group and task version as between subject). We similarly evaluated Experiment 3 parameter estimates for treatment by time interactions. For Experiment 2, we performed ANCOVAs for μ_3^0, κ, ω_2, and ω_3 to evaluate three sets of covariates: (1) demographics (age, gender, ethnicity, and race); (2) mental health factors (medication usage, diagnostic category, BAI score, and BDI score); (3) and metrics and correlates of global cognitive function (educational attainment, income, and cognitive reflection). Unless otherwise stated, post-hoc tests were conducted as least significant difference (LSD)-corrected estimated marginal means.

Meta-analyses were conducted using random effects models with the R Metafor package[82]. Mean differences were assessed for low versus high paranoia groups in the in laboratory experiment and online version 3. Standardized mean differences (methamphetamine or high paranoia versus saline or low paranoia) were employed to account for the differences in task design between animal and human studies.

The 2-step clustering analysis approach was selected to automatically determine optimal cluster count and cluster group assignment. Clustering variables included paranoia-relevant parameter estimates (μ_3^0, κ, ω_2, and ω_3) from Experiment 1 (block 1); online, version 3 (block 1), and rats (Post-Rx) as continuous variables with a Log-likelihood distance measure, maximum cluster count of 15, and Schwarz's Bayesian Criterion (BIC) clustering criterion. We validated our clustering solution by sorting the data into two halves and running separate cluster analyses. We also compared cluster solutions derived exclusively from rat data versus human data. A Chi-Square test determined the significance of the association between cluster membership and group (methamphetamine/high paranoia versus saline/low paranoia).

Data availability

Data are available on ModelDB[83] (http://modeldb.yale.edu/258631) with accession code **p2c8q74m**. Figures 2, 3, 4, and 5 and Supplemental Figures 1 and 2 have associated raw data.

Code availability

Code for the HGF toolbox v5.3.1 is freely available at https://translationalneuromodeling.github.io/tapas/. Additional instructions are provided in the Supplementary Information. Task code is available by request.

References

1 Freeman, D. & Garety, P. A. Comments on the content of persecutory delusions: does the definition need clarification? *Br J Clin Psychol* **39 (Pt 4)**, 407-414 (2000).

2 Freeman, D. *et al.* Psychological investigation of the structure of paranoia in a non-clinical population. *The British journal of psychiatry : the journal of mental science* **186**, 427-435, doi:10.1192/bjp.186.5.427 (2005).

3 Freeman, D., Pugh, K., Vorontsova, N., Antley, A. & Slater, M. Testing the continuum of delusional beliefs: an experimental study using virtual reality. *J Abnorm Psychol* **119**, 83-92, doi:10.1037/a0017514 (2010).

4 Freeman, D. *et al.* Concomitants of paranoia in the general population. *Psychol Med* **41**, 923-936, doi:10.1017/S0033291710001546 (2011).

5 Bebbington, P. E. *et al.* The structure of paranoia in the general population. *The British journal of psychiatry : the journal of mental science* **202**, 419-427, doi:10.1192/bjp.bp.112.119032 (2013).

6 Freeman, D. Delusions in the nonclinical population. *Curr Psychiatry Rep* **8**, 191-204 (2006).

7 Hofstadter, R. The Paranoid Style in American Politics. *Harper's Magazine*, 77-86 (1964).

8 Freeman, D. Suspicious minds: the psychology of persecutory delusions. *Clin Psychol Rev* **27**, 425-457, doi:10.1016/j.cpr.2006.10.004 (2007).

9 van Os, J. *et al.* Self-reported psychosis-like symptoms and the continuum of psychosis. *Social Psychiatry and Psychiatric Epidemiology*.**34**, pp, doi:10.1007/s001270050220 10541665 (1999).

10 Frangos, E., Athanassenas, G., Tsitourides, S., Psilolignos, P. & Katsanou, N. Psychotic depressive disorder: A separate entity? *Journal of Affective Disorders*.**5**, pp, doi:10.1016/0165-0327%2883%2990049-6 6224837 (1983).

11 Trimble, M. R. The schizophrenia-like psychosis of epilepsy. *Neuropsychiatry, Neuropsychology, & Behavioral Neurology*.**5**, pp (1992).

12 Rubin, E. H., Drevets, W. C. & Burke, W. J. The nature of psychotic symptoms in senile dementia of the Alzheimer type. *Journal of Geriatric Psychiatry and Neurology*.**1**, pp, doi:10.1177/089198878800100104 3266997 (1988).

13 Leamon, M. H. *et al.* Methamphetamine and paranoia: the methamphetamine experience questionnaire. *Am J Addict* **19**, 155-168, doi:10.1111/j.1521-0391.2009.00014.x (2010).

14 Raihani, N. J. & Bell, V. Conflict and cooperation in paranoia: a large-scale behavioural experiment. *Psychol Med* **48**, 1523-1531, doi:10.1017/S0033291717003075 (2018).

15 Raihani, N. J. & Bell, V. Paranoia and the social representation of others: a large-scale game theory approach. *Sci Rep* **7**, 4544, doi:10.1038/s41598-017-04805-3 (2017).

16 Raihani, N. J. & Bell, V. An evolutionary perspective on paranoia. *Nat Hum Behav* **3**, 114-121, doi:10.1038/s41562-018-0495-0 (2019).

17 Fineberg, S. K., Steinfeld, M., Brewer, J. A. & Corlett, P. R. A Computational Account of Borderline Personality Disorder: Impaired Predictive Learning about Self and Others Through Bodily Simulation. *Front Psychiatry* **5**, 111, doi:10.3389/fpsyt.2014.00111 (2014).

18 Behrens, T. E., Hunt, L. T., Woolrich, M. W. & Rushworth, M. F. Associative learning of social value. *Nature* **456**, 245-249, doi:10.1038/nature07538 (2008).

19 Cramer, R. E. *et al.* Human agency and associative learning: Pavlovian principles govern social process in causal relationship detection. *Q J Exp Psychol B* **55**, 241-266, doi:10.1080/02724990143000289 (2002).

20 Heyes, C. & Pearce, J. M. Not-so-social learning strategies. *Proceedings. Biological sciences / The Royal Society* **282**, doi:10.1098/rspb.2014.1709 (2015).

21 Rescorla, R. A., Wagner, A.R. in *Classical conditioning II: Current research and theory* (ed A.H. Black, Prokasy, W.F.) (Appleton-Century-Crofts, 1972).

22 Corlett, P. R. *et al.* Disrupted prediction-error signal in psychosis: evidence for an associative account of delusions. *Brain : a journal of neurology* **130**, 2387-2400, doi:10.1093/brain/awm173 (2007).

23 Dickinson, A. The 28th Bartlett Memorial Lecture. Causal learning: an associative analysis. *Q J Exp Psychol B* **54**, 3-25 (2001).

24 Soltani, A. & Izquierdo, A. Adaptive learning under expected and unexpected uncertainty. *Nat Rev Neurosci*, doi:10.1038/s41583-019-0180-y (2019).

25 Yu, A. J. & Dayan, P. Uncertainty, neuromodulation, and attention. *Neuron* **46**, 681-692, doi:10.1016/j.neuron.2005.04.026 (2005).

26 Payzan-LeNestour, E. & Bossaerts, P. Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. *PLoS Comput Biol* **7**, e1001048, doi:10.1371/journal.pcbi.1001048 (2011).

27 Payzan-LeNestour, E., Dunne, S., Bossaerts, P. & O'Doherty, J. P. The neural representation of unexpected uncertainty during value-based decision making. *Neuron* **79**, 191-201, doi:10.1016/j.neuron.2013.04.037 (2013).

28 Tervo, D. G. *et al.* Behavioral variability through stochastic choice and its gating by anterior cingulate cortex. *Cell* **159**, 21-32, doi:10.1016/j.cell.2014.08.037 (2014).

29 Nour, M. M., McCutcheon, R. & Howes, O. D. The Relationship Between Dopamine Synthesis Capacity and Release: Implications for Psychosis. *Neuropsychopharmacology* **43**, 1195-1196, doi:10.1038/npp.2017.293 (2018).

30 Mathys, C., Daunizeau, J., Friston, K. J. & Stephan, K. E. A bayesian foundation for individual learning under uncertainty. *Frontiers in human neuroscience* **5**, 39, doi:10.3389/fnhum.2011.00039 (2011).

31 Mathys, C. D. *et al.* Uncertainty in perception and the Hierarchical Gaussian Filter. *Frontiers in human neuroscience* **8**, 825, doi:10.3389/fnhum.2014.00825 (2014).

32 Groman, S. M., Rich, K. M., Smith, N. J., Lee, D. & Taylor, J. R. Chronic Exposure to Methamphetamine Disrupts Reinforcement-Based Decision Making in Rats. *Neuropsychopharmacology* **43**, 770-780, doi:10.1038/npp.2017.159 (2018).

33 Ryder, A. G., Costa, P. T. & Bagby, R. M. Evaluation of the SCID-II personality disorder traits for DSM-IV: coherence, discrimination, relations with general personality traits, and functional impairment. *J Pers Disord* **21**, 626-637, doi:10.1521/pedi.2007.21.6.626 (2007).

34 Beck, A. T., Epstein, N., Brown, G. & Steer, R. A. An inventory for measuring clinical anxiety: psychometric properties. *J Consult Clin Psychol* **56**, 893-897 (1988).

35 Beck, A. T., Ward, C. H., Mendelson, M., Mock, J. & Erbaugh, J. An inventory for measuring depression. *Archives of general psychiatry* **4**, 561-571 (1961).

36 Lawson, R. P., Mathys, C. & Rees, G. Adults with autism overestimate the volatility of the sensory environment. *Nat Neurosci* **20**, 1293-1299, doi:10.1038/nn.4615 (2017).

37 Powers, A. R., Mathys, C. & Corlett, P. R. Pavlovian conditioning-induced hallucinations result from overweighting of perceptual priors. *Science* **357**, 596-600, doi:10.1126/science.aan3458 (2017).

38 Sevgi, M., Diaconescu, A. O., Tittgemeyer, M. & Schilbach, L. Social Bayes: Using Bayesian Modeling to Study Autistic Trait-Related Differences in Social Cognition. *Biological psychiatry* **80**, 112-119, doi:10.1016/j.biopsych.2015.11.025 (2016).

39 Kong, X., McEwan, J.S., Bizo, L.A., Foster, T.M. An Analysis of U-Value as a Measure of Variability. *Psychological Rec* **67**, 581-586 (2017).

40 Fung, B. J., Qi, S., Hassabis, D., Daw, N. & Mobbs, D. Slow escape decisions are swayed by trait anxiety. *Nat Hum Behav* **3**, 702-708, doi:10.1038/s41562-019-0595-5 (2019).

41 Humphries, D. A. & Driver, P. M. Protean defence by prey animals. *Oecologia* **5**, 285-302, doi:10.1007/BF00815496 (1970).

42 Broker, F., Marshall, L., Bestmann, S. & Dayan, P. Forget-me-some: General versus special purpose models in a hierarchical probabilistic task. *PLoS One* **13**, e0205974, doi:10.1371/journal.pone.0205974 (2018).

43 Tkaczynski, A. in *Segmentation in Social Marketing* (ed Rundle-Thiele S. Dietrich T., Kubacki K) (2017).

44 Deserno, L. Overestimating environmental volatility increases switching behavior and is linked to activation of dorsolateral prefrontal cortex in schizophrenia. *Bioarxiv* (2018).

45 Webster, D. M. & Kruglanski, A. W. Individual differences in need for cognitive closure. *Journal of personality and social psychology* **67**, 1049-1062, doi:10.1037//0022-3514.67.6.1049 (1994).

46 Holt, A. E. & Albert, M. L. Cognitive neuroscience of delusions in aging. *Neuropsychiatr Dis Treat* **2**, 181-189, doi:10.2147/nedt.2006.2.2.181 (2006).

47 Iacovino, J. M., Jackson, J. J. & Oltmanns, T. F. The relative impact of socioeconomic status and childhood trauma on Black-White differences in paranoid personality disorder symptoms. *J Abnorm Psychol* **123**, 225-230, doi:10.1037/a0035258 (2014).

48 Mahoney, J. J., 3rd, Hawkins, R. Y., De La Garza, R., 2nd, Kalechstein, A. D. & Newton, T. F. Relationship between gender and psychotic symptoms in cocaine-dependent and methamphetamine-dependent participants. *Gend Med* **7**, 414-421, doi:10.1016/j.genm.2010.09.003 (2010).

49 Chakraborty, S., Kolling, N., Walton, M. E. & Mitchell, A. S. Critical role for the mediodorsal thalamus in permitting rapid reward-guided updating in stochastic reward environments. *Elife* **5**, doi:10.7554/eLife.13588 (2016).

50 Groman, S. M. *et al.* Orbitofrontal Circuits Control Multiple Reinforcement-Learning Processes. *Neuron* **103**, 734-746 e733, doi:10.1016/j.neuron.2019.05.042 (2019).

51 Marshall, L. *et al.* Pharmacological Fingerprints of Contextual Uncertainty. *PLoS Biol* **14**, e1002575, doi:10.1371/journal.pbio.1002575 (2016).

52 Kane, G. A. *et al.* Increased locus coeruleus tonic activity causes disengagement from a patch-foraging task. *Cogn Affect Behav Neurosci* **17**, 1073-1083, doi:10.3758/s13415-017-0531-y (2017).

53 Aston-Jones, G. & Cohen, J. D. An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annu Rev Neurosci* **28**, 403-450, doi:10.1146/annurev.neuro.28.061604.135709 (2005).

54 Aston-Jones, G., Rajkowski, J. & Cohen, J. Role of locus coeruleus in attention and behavioral flexibility. *Biol Psychiatry* **46**, 1309-1320 (1999).

55 Eldar, E., Cohen, J. D. & Niv, Y. The effects of neural gain on attention and learning. *Nat Neurosci* **16**, 1146-1153, doi:10.1038/nn.3428 (2013).

56 Borodovitsyna, O., Flamini, M. D. & Chandler, D. J. Acute Stress Persistently Alters Locus Coeruleus Function and Anxiety-like Behavior in Adolescent Rats. *Neuroscience* **373**, 7-19, doi:10.1016/j.neuroscience.2018.01.020 (2018).

57 McCall, J. G. *et al.* CRH Engagement of the Locus Coeruleus Noradrenergic System Mediates Stress-Induced Anxiety. *Neuron* **87**, 605-620, doi:10.1016/j.neuron.2015.07.002 (2015).

58 Atzori, M. *et al.* Locus Ceruleus Norepinephrine Release: A Central Regulator of CNS Spatio-Temporal Activation? *Front Synaptic Neurosci* **8**, 25, doi:10.3389/fnsyn.2016.00025 (2016).

59 Liddell, B. J. *et al.* A direct brainstem-amygdala-cortical 'alarm' system for subliminal signals of fear. *Neuroimage* **24**, 235-243, doi:10.1016/j.neuroimage.2004.08.016 (2005).

60 Rothman, R. B. *et al.* Amphetamine-type central nervous system stimulants release norepinephrine more potently than they release dopamine and serotonin. *Synapse* **39**, 32-41, doi:10.1002/1098-2396(20010101)39:1<32::AID-SYN5>3.0.CO;2-3 (2001).

61 Ferrucci, M., Giorgi, F. S., Bartalucci, A., Busceti, C. L. & Fornai, F. The effects of locus coeruleus and norepinephrine in methamphetamine toxicity. *Curr Neuropharmacol* **11**, 80-94, doi:10.2174/157015913804999522 (2013).

62 Segal, D. S., Kuczenski, R., O'Neil, M. L., Melega, W. P. & Cho, A. K. Escalating dose methamphetamine pretreatment alters the behavioral and neurochemical profiles associated with exposure to a high-dose methamphetamine binge. *Neuropsychopharmacology* **28**, 1730-1740, doi:10.1038/sj.npp.1300247 (2003).

63 Nour, M. M. *et al.* Dopaminergic basis for signaling belief updates, but not surprise, and the link to paranoia. *Proc Natl Acad Sci U S A* **115**, E10167-E10176, doi:10.1073/pnas.1809298115 (2018).

64 Schlagenhauf, F. *et al.* Striatal dysfunction during reversal learning in unmedicated schizophrenia patients. *Neuroimage* **89**, 171-180, doi:10.1016/j.neuroimage.2013.11.034 (2014).

65 Richardson, G., Dickinson, P., Burman, O. H. P. & Pike, T. W. Unpredictable movement as an anti-predator strategy. *Proc Biol Sci* **285**, doi:10.1098/rspb.2018.1112 (2018).

66 Humphries, D. A. & Driver, P. M. Erratic display as a device against predators. *Science* **156**, 1767-1768 (1967).

67 Aston-Jones, G., Rajkowski, J., Kubiak, P. & Alexinsky, T. Locus coeruleus neurons in monkey are selectively activated by attended cues in a vigilance task. *J Neurosci* **14**, 4467-4480 (1994).

68 Rajkowski, J., Kubiak, P. & Aston-Jones, G. Locus coeruleus activity in monkey: phasic and tonic changes are associated with altered vigilance. *Brain Res Bull* **35**, 607-616 (1994).

69 Usher, M., Cohen, J. D., Servan-Schreiber, D., Rajkowski, J. & Aston-Jones, G. The role of locus coeruleus in the regulation of cognitive performance. *Science* **283**, 549-554 (1999).

70 Eldar, E., Niv, Y. & Cohen, J. D. Do You See the Forest or the Tree? Neural Gain and Breadth Versus Focus in Perceptual Processing. *Psychol Sci* **27**, 1632-1643, doi:10.1177/0956797616665578 (2016).

71 Sales, A. C., Friston, K. J., Jones, M. W., Pickering, A. E. & Moran, R. J. Locus Coeruleus tracking of prediction errors optimises cognitive flexibility: An Active Inference model. *PLoS Comput Biol* **15**, e1006267, doi:10.1371/journal.pcbi.1006267 (2019).

72 Corlett, P. R., Fletcher, P.C. Computational Psychiatry: A Rosetta Stone linking the brain to mental illness. *Lancet Psychiatry* (2014).

73 Feeney, E. J., Groman, S. M., Taylor, J. R. & Corlett, P. R. Explaining Delusions: Reducing Uncertainty Through Basic and Computational Neuroscience. *Schizophr Bull*, doi:10.1093/schbul/sbw194 (2017).

74 Gershman, S. J. & Niv, Y. Exploring a latent cause theory of classical conditioning. *Learn Behav* **40**, 255-268, doi:10.3758/s13420-012-0080-8 (2012).

75 Gershman, S. J. & Niv, Y. Learning latent structure: carving nature at its joints. *Curr Opin Neurobiol* **20**, 251-256, doi:10.1016/j.conb.2010.02.008 (2010).

76 Turner, J. C., Oakes, P.J., Haslam, S.A., McGarty, C. Self and Collective: Cognition and Social Context. *Personality and Social Psychology B* **20**, 454-463 (1994).

77 Powers, A. R., 3rd, Bien, C. & Corlett, P. R. Aligning Computational Psychiatry With the Hearing Voices Movement: Hearing Their Voices. *JAMA psychiatry* **75**, 640-641, doi:10.1001/jamapsychiatry.2018.0509 (2018).

78 van Prooijen, J. W. & Douglas, K. M. Conspiracy theories as part of history: The role of societal crisis situations. *Mem Stud* **10**, 323-333, doi:10.1177/1750698017701615 (2017).

79 Han, E., Paulus, M. P., Wittmann, M., Chung, H. & Song, J. M. Hair analysis and self-report of methamphetamine use by methamphetamine dependent individuals. *J Chromatogr B Analyt Technol Biomed Life Sci* **879**, 541-547, doi:10.1016/j.jchromb.2011.01.002 (2011).

80 Bell, V. & O'Driscoll, C. The network structure of paranoia in the general population. *Soc Psychiatry Psychiatr Epidemiol* **53**, 737-744, doi:10.1007/s00127-018-1487-0 (2018).

81 Spitzer, M., Wildenhain, J., Rappsilber, J. & Tyers, M. BoxPlotR: a web tool for generation of box plots. *Nat Methods* **11**, 121-122, doi:10.1038/nmeth.2811 (2014).

82 Viechtbauer, W. Conducting meta-analyses in R with the metafor package. *Journal of statistical software* **36** (2010).

83 McDougal, R. A. *et al.* Twenty years of ModelDB and beyond: building essential modeling tools for the future of neuroscience. *J Comput Neurosci* **42**, 1-10, doi:10.1007/s10827-016-0623-7 (2017).

Acknowledgements

This work was supported by the Yale University Department of Psychiatry, the Connecticut Mental Health Center (CMHC) and Connecticut State Department of Mental Health and Addiction Services (DMHAS). It was funded by an IMHRO / Janssen Rising Star Translational Research Award, an Interacting Minds Center (Aarhus) Pilot Project Award, and NIMH R01MH12887 (P.R.C.). E.J.R. was supported by the NIH Medical Scientist Training Program Training Grant, GM007205; NINDS Neurobiology of Cortical Systems Grant, T32 NS007224; and a Gustavus and Louise Pfeiffer Research Foundation Fellowship. S.U. received funding from NSF Fellowships DGE1122492 and DGE1752134. S.M.G. and J.R.T. were supported by NIDA DA DA041480. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. The authors thank Dr. James Waltz for providing an earlier version of the reversal-learning e-prime code. The authors acknowledge the help, support, and advice of Dr. Sarah Fineberg, Dr. Albert Powers III, and Dr. Pantelis Leptourgos.

Author contributions

E.J.R., S.U., C.D.M., J.R.T., S.M.G., and P.R.C. contributed to the conception and design of the experiment. S.U. and E.J.R. developed the online experiments. C.D.M. advised on modelling analyses. S.M.G. and J.R.T. contributed the rodent dataset. E.J.R. conducted the experiments, collected the data, and analysed the data. E.J.R. and P.R.C. drafted the manuscript. All authors reviewed the manuscript and gave final approval for publication.

Competing interests

The authors declare no competing interests.

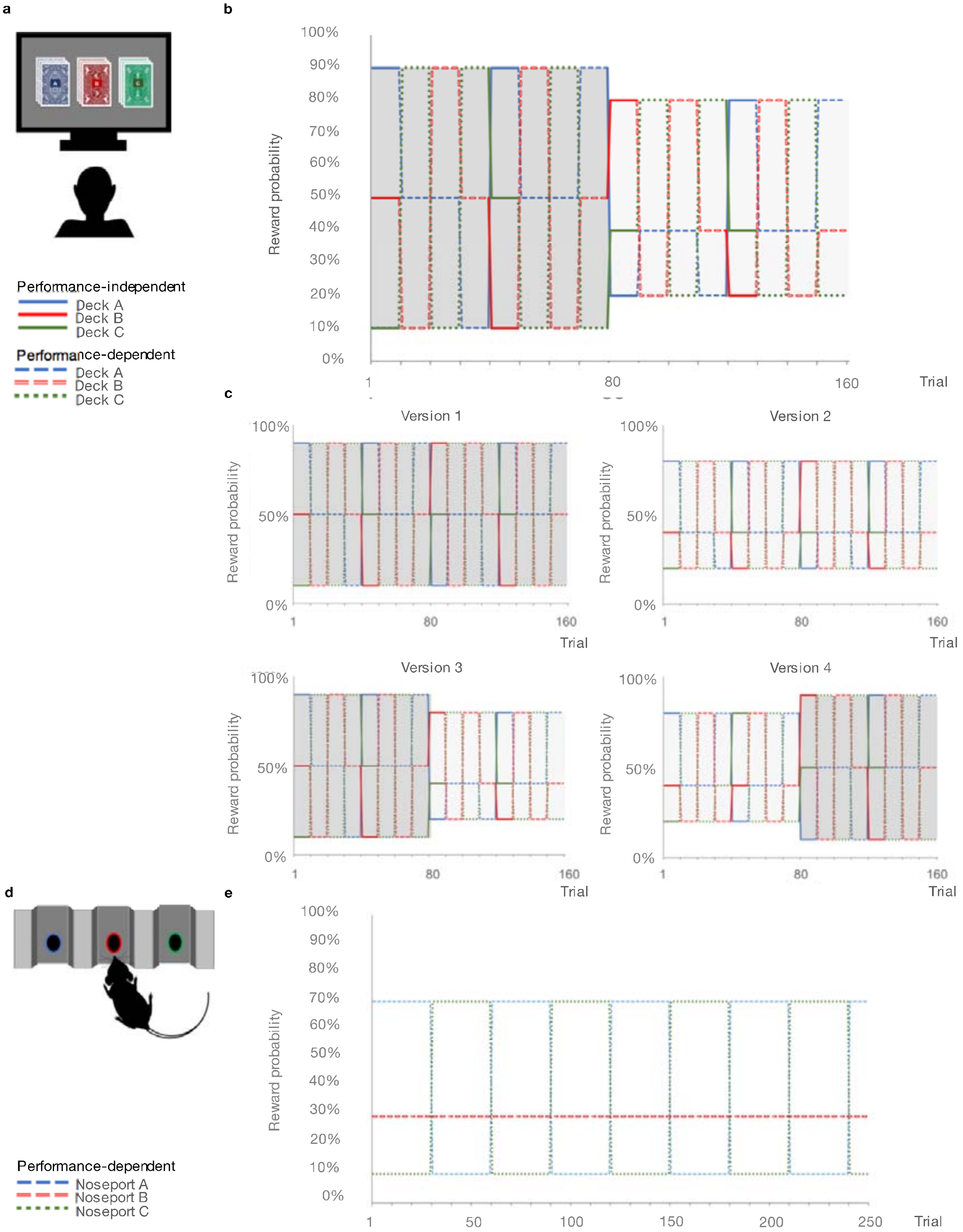
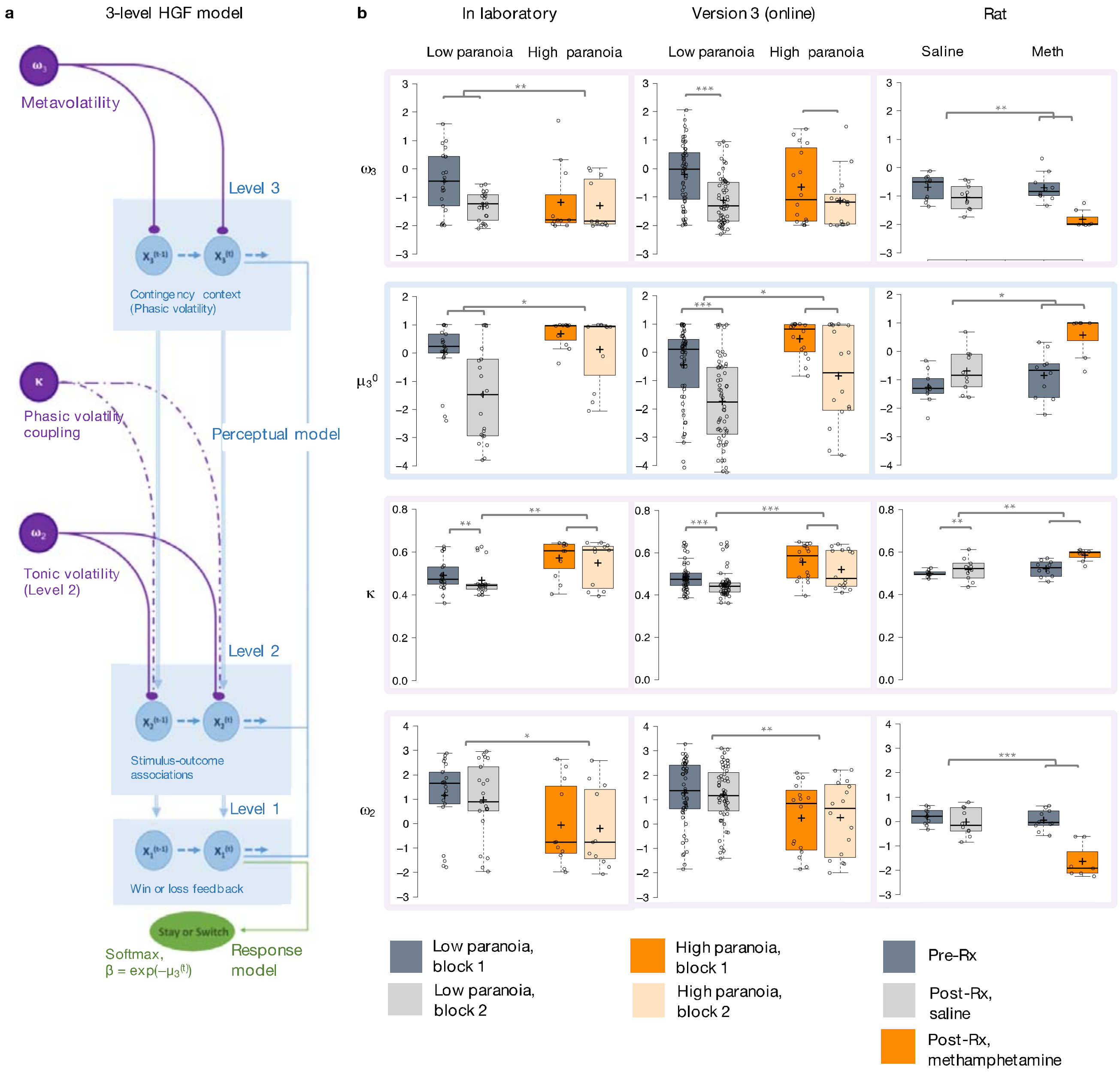
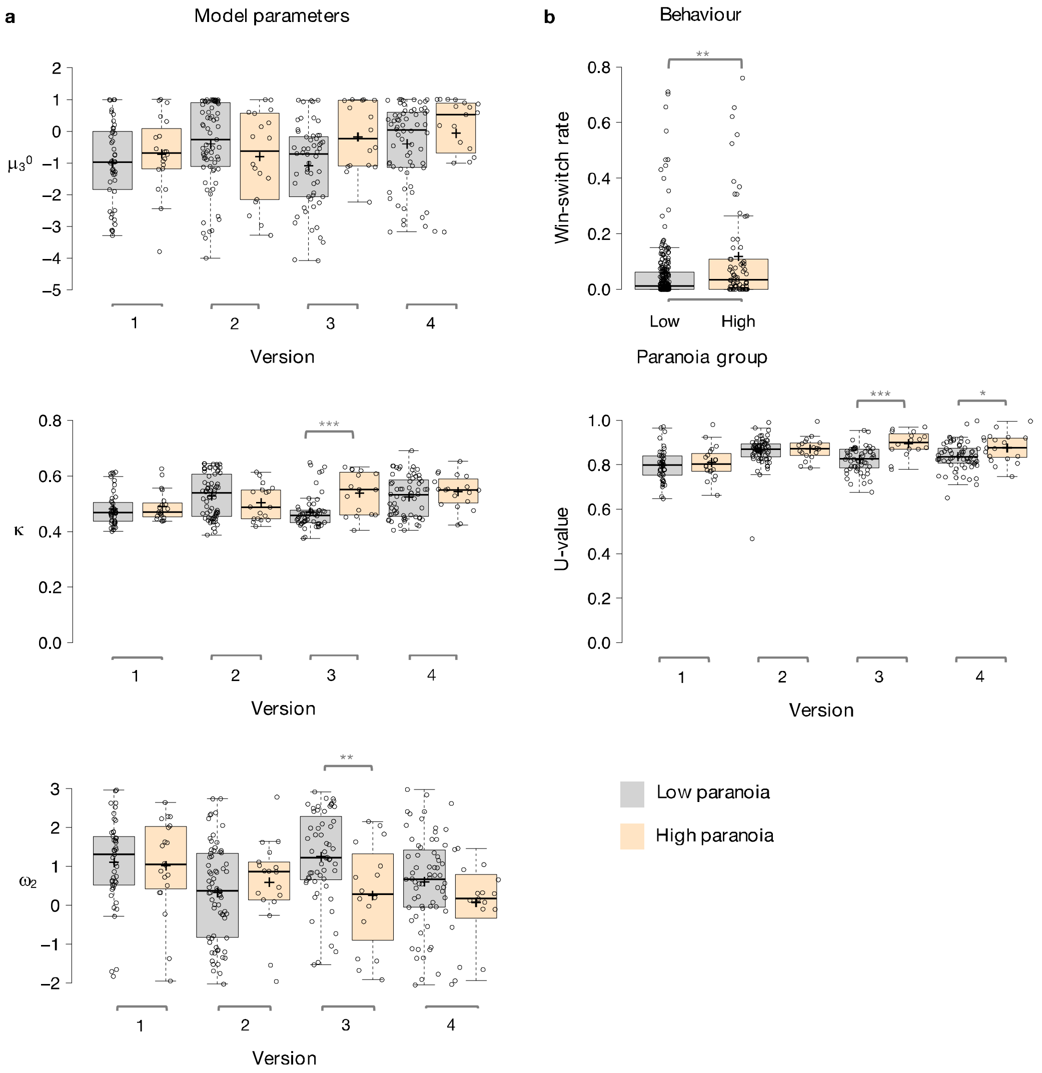


Fig. 1. Probabilistic reversal learning task. a, Human paradigm: participants choose between three decks of cards with different, unknown probabilities of reward and loss. **b,** Reward contingency schedule for in laboratory experiment. On trial 81, the probability context shifts from 90%, 50%, and 10% (dark grey) to 80%, 40%, and 20% without warning (light grey). **c**, Reward contingency schedules for online experiment. **d,** Rat paradigm: subjects choose between three noseports with different probabilities of sucrose pellet reward. **e,** Reward contingency schedule for rat experiment[32].



Fi ı a

sc stimulus-outcome associations (e.g., deck values). Level 3 (x_3): perception of the overall reward contingency context. The impact of phasic volatility upon x_2 is captured by κ (i.e., coupling). Tonic volatility modulates x_3 and x_2 via ω_3 and ω_2, respectively. μ_3^0 is the initial value of the third level volatility belief. **b,** Parameters replicate across high paranoia groups (orange) in the in laboratory experiment (*n*=21 low paranoia, 11 high paranoia); analogous online task (version 3, *n*=56 low paranoia, 16 high paranoia); and rats exposed to chronic, escalating methamphetamine (*n*=10 per group Pre-Rx; *n*=10 saline, 7 methamphetamine Post-Rx). Centre lines depict medians; box limits indicate the 25th and 75th percentiles; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles, outliers are represented by dots; crosses represent sample means; data points are plotted as open circles. *$P \leq 0.05$, **$P \leq 0.01$, ***$P \leq 0.001$.



Fi

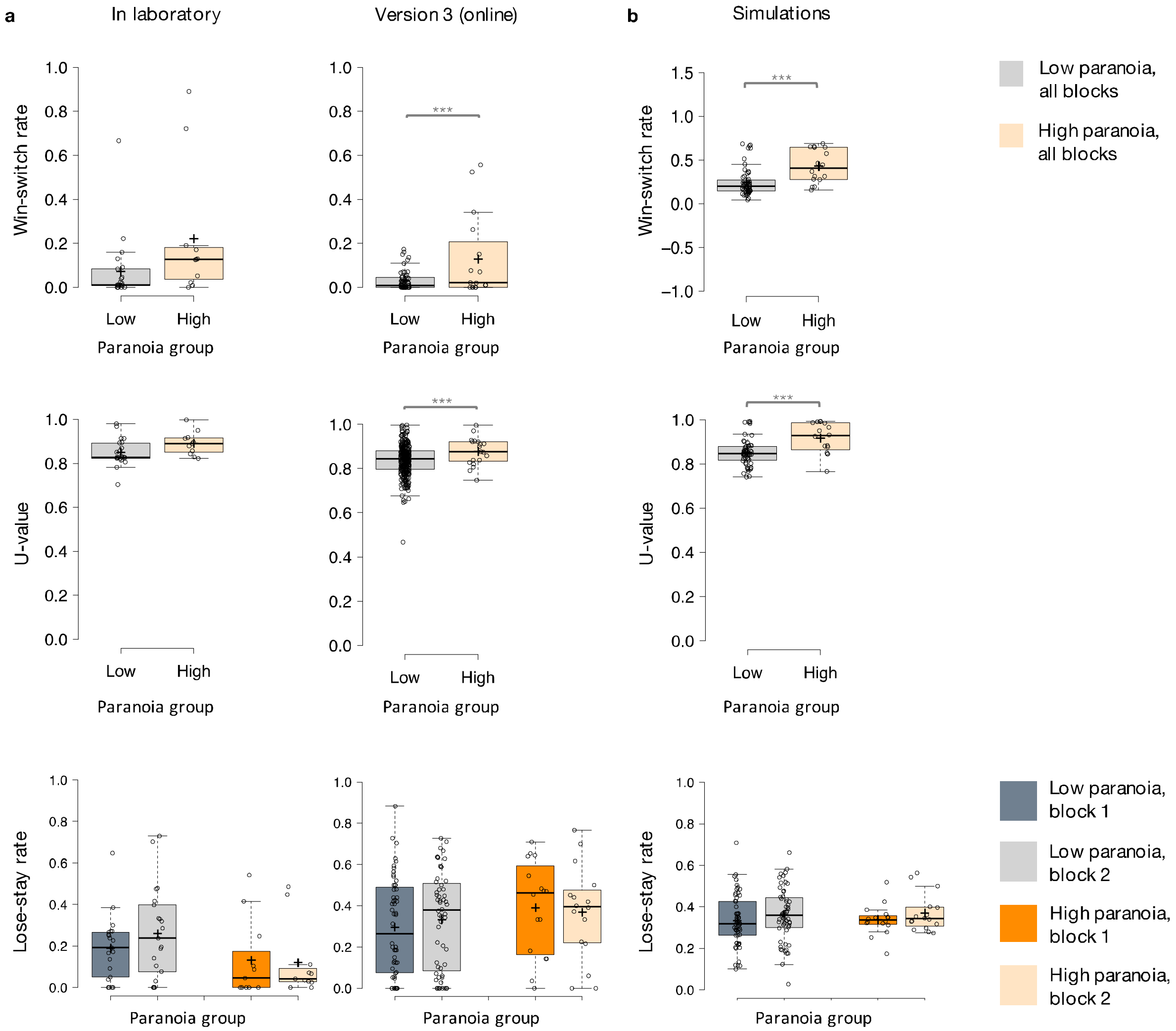
tre

Be

ta

on

perseverative returns to a previously rewarding option. Centre lines show the medians; box limits indicate the 25th and 75th percentiles; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles, outliers are represented by dots; crosses represent sample means; data points are plotted as open circles. *P*-values correspond to estimated marginal means: **P* ≤ 0.05, ***P* ≤ 0.01, ****P* ≤ 0.001.



Fi

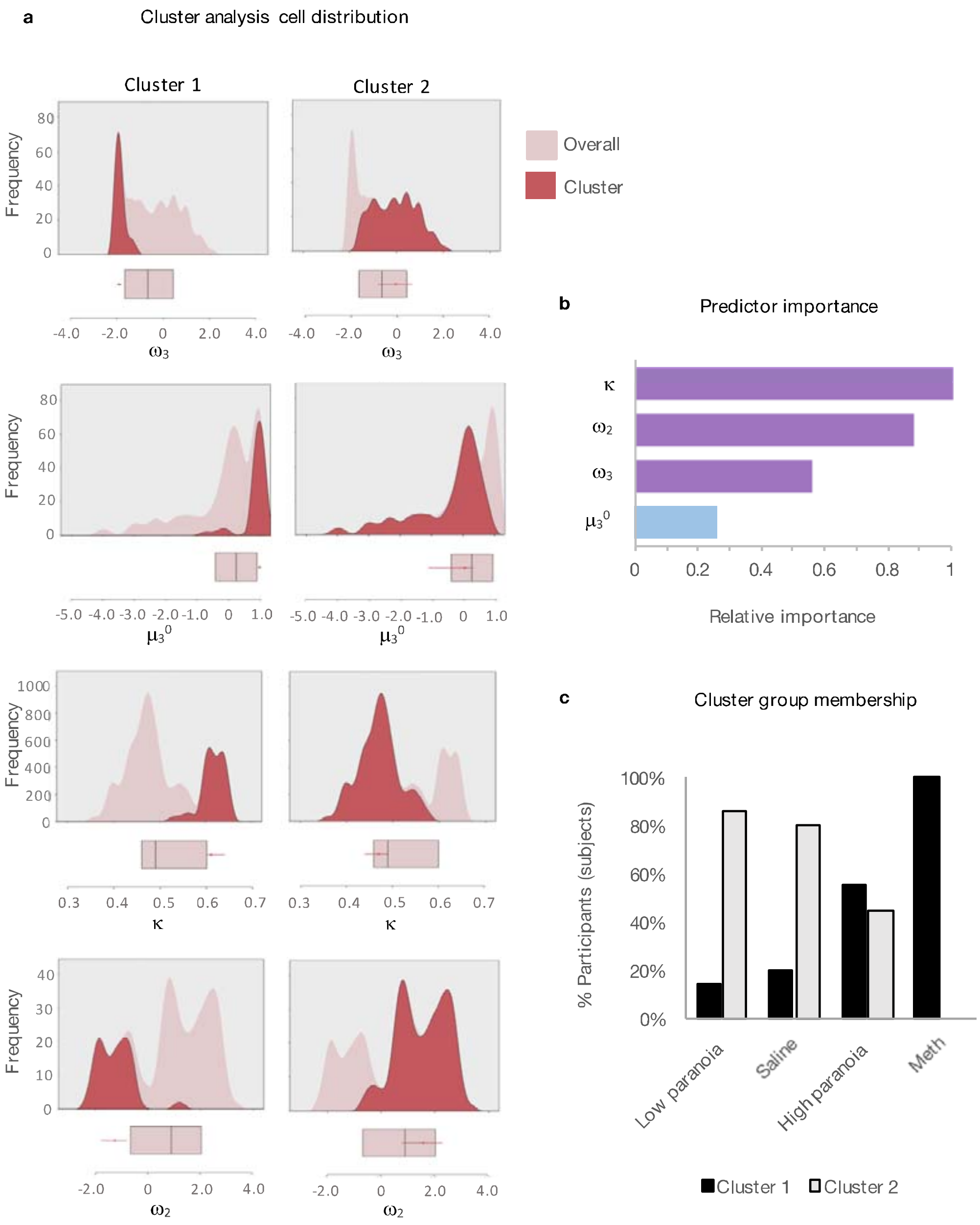
an

sig

pa

pa

medians; box limits indicate the 25th and 75th percentiles; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles, outliers are represented by dots; crosses represent sample means; data points are plotted as open circles; *n*=21 low paranoia, 11 high paranoia (in laboratory); *n*=56 low paranoia, 16 high paranoia (online, version 3); **P* ≤ 0.05, ***P* ≤ 0.01, ****P* ≤ 0.001.



Fi

hu

op

ra

(re er

me

ele

Pr

n= ly

associated with paranoia and methamphetamine groups (χ^2(1, *n*=121)=29.447, *P*=5.75E-8).

Table 1. In laboratory vs. online version 3

	In laboratory				Online version 3			
	Low paranoia (n=21)	High paranoia (n=11)	Statistic	p-value	Low paranoia (n=56)	High paranoia (n=16)	Statistic	p-value
Demographics								
Age (years)	36.0 [3.2]	38.9 [3.9]	-0.531 (27)†	0.6	38.6 [1.6]	32.9 [1.7]	2.441 (41.842)†	**0.019¶**
Gender			0.006 (1)‡	1§			.780 (1)‡	0.410
% Female	71.4%	72.7%	n/a	n/a	50.0%	62.5%	n/a	n/a
% Male	28.6%	27.3%	n/a	n/a	50.0%	37.5%	n/a	n/a
% Other or not specified	0%	0%	n/a	n/a	0%	0%	n/a	n/a
Education			4.972 (6)‡	0.638§			5.351 (6)‡	0.549§
% High school degree or equivalent	19.0%	45.5%	n/a	n/a	16.1%	6.3%	n/a	n/a
% Some college or university, no degree	14.3%	0%	n/a	n/a	17.9%	25.0%	n/a	n/a
% Associate degree	9.5%	9.1%	n/a	n/a	12.5%	12.5%	n/a	n/a
% Bachelor's degree	23.8%	27.3%	n/a	n/a	35.7%	56.3%	n/a	n/a
% Master's degree	9.5%	0%	n/a	n/a	14.3%	0%	n/a	n/a
% Doctorate or professional degree	4.8%	0%	n/a	n/a	1.8%	0%	n/a	n/a
% Completed some postgraduate	0%	0%	n/a	n/a	1.8%	0%	n/a	n/a
% Other / not specified	19.0%	18.2%	n/a	n/a	0%	0%	n/a	n/a
Ethnicity			.134 (1)‡	1§			.117 (1)‡	1§
% Hispanic, Latino, or Spanish origin	23.8%	18.2%	n/a	n/a	8.9%	6.3%	n/a	n/a
% Not of Hispanic, Latino, or Spanish origin	76.2%	81.8%	n/a	n/a	91.1%	93.8%	n/a	n/a
Race			6.250 (4)‡	0.186§			5.368 (4)‡	0.229§
% White	*61.9%*	*36.4%*	n/a	n/a	85.7%	75.0%	n/a	n/a
% Black or African American	*19.0%*	*36.4%*	n/a	n/a	0%	12.5%	n/a	n/a
% Asian	*14.3%*	*9.1%*	n/a	n/a	3.6%	6.3%	n/a	n/a
% American Indian or Alaska Native	*4.8%*	*0%*	n/a	n/a	1.8%	6.3%	n/a	n/a
% Multiracial	*0%*	*0%*	n/a	n/a	3.6%	0%	n/a	n/a
% Other / not specified	*0%*	*18.2%*	n/a	n/a	5.4%	0%	n/a	n/a
Mental health								
Psychiatric diagnosis			12.329 (2)‡	**0.002§**			7.850 (3)‡	**0.039§**
% No psychiatric diagnosis	71.4%	9.1%	adj. residuals	**0.004**	71.4%	50.0%	adj. residuals	0.465
% Schizophrenia spectrum	19.0%	36.4%	adj. residuals	0.546	0%	6.3%	adj. residuals	0.307
% Mood disorder	9.5%	54.5%	adj. residuals	0.020#	21.4%	43.8%	adj. residuals	0.356
% Not specified	0%	0%	adj. residuals	n/a	7.1%	0%	adj. residuals	0.751
% Medicated	23.8%	81.8%	9.871 (1)‡	**0.003§**	7.1%	31.3%	8.730 (2)‡	**0.023§**
Beck's Anxiety Inventory	0.27 [0.08]	0.85 [0.17]	-3.453 (30)†	**0.002**	0.24 [0.04]	0.90 [0.20]	-3.303 (16.179)†	**0.004¶**
Beck's Depression Inventory	0.23 [0.05]	0.66 [0.15]	-2.67 (11.854)†	**0.021¶**	0.25 [0.04]	1.03 [0.19]	-3.951 (16.659)†	**0.001¶**
SCID Paranoia Personality Score	0.09 [0.02]	0.63 [0.04]	-13.476 (30)†	**2.92E-14**	0.1 [0.02]	0.72 [0.04]	-16.551 (70)†	**6.712E-26**
Reversal learning performance								
Total points earned	7061.9 [286.9]	6290.9 [372.2]	1.608 (30)†	0.118	7533.0 [143.8]	6503.1 [340.6]	3.177 (70)†	**0.002**
Total reversals achieved	4.8 [0.7]	2.5 [0.8]	2.145 (30)†	**0.04**	6.3 [0.3]	4.9 [0.8]	1.758 (20.14)†	0.094¶
% Achieving reversals	90.5%	72.7%	1.407 (1)‡	0.327§	100%	87.5%	7.200 (1)‡	**0.047§**
Trials to switch	1.68 [0.22]	1.43 [0.20]	0.671 (24)†	0.509	2.1 [0.2]	2.6 [0.6]	-1.088 (64)†	0.280
Trials to recovery	3.75 [0.51]	4 [0.93]	-0.285 (21)†	0.779	2.9 [0.3]	4.9 [0.8]	-2.694 (60)†	**0.009**
Win-switch rate, block 1 (90-50-10)	0.08 [0.03]	0.24 [0.09]	-1.742 (12.379)†	0.106¶	0.04 [0.01]	0.13 [0.05]	-1.906 (15.762)†	0.075¶
Win-switch rate, block 2 (80-40-20)	0.07 [0.04]	0.21 [0.1]	-1.601 (30)†	0.12	0.02 [0.01]	0.12 [0.05]	-2.02 (15.915)†	0.061¶
Lose-stay rate, block 1 (90-50-10)	0.19 [0.03]	0.13 [0.06]	0.919 (30)†	0.365	0.30 [0.03]	0.39 [0.06]	-1.425 (70)†	0.158
Lose-stay rate, block 2 (80-40-20)	0.26 [0.05]	0.12 [0.05]	1.817 (30)†	0.079	0.33 [0.03]	0.37 [0.06]	-0.554 (70)†	0.581
Null trials	8.5 [2.8]	10.4 [3.7]	-0.391 (30)†	0.699	n/a	n/a	n/a	n/a

Columns display means [standard error] or percentage of participants within the described category, test-statistics, and p-values.
†Independent samples t-test: t-value (df). Two-tailed *P*-values reported.
‡Chi square coefficient (df).
§Fisher's exact test, exact significance (2-sided).
¶Equal variances not assumed.
#Not significant (Bonferonni correction).
††Data presented in Fig. 4; repeated measures ANOVA, paranoia group trend or effect: *F*(df), *P*; estimated marginal means and standard error.
‡‡Data presented in Fig. 2; repeated measures ANOVA, *F*(df), *P*. In laboratory: paranoia x block interactions for ω_3, μ_3^0; paranoia group effects for κ, ω_2. Version 3: paranoia group effects reported. See Supplementary Table 1 for complete ANOVA. results.

Table 2. Online experiment

	Version 1		Version 2		Version 3		Version 4		Version Effect		Paranoia Effect		Interaction	
	Low paranoia (n=45)	High paranoia (n=20)	Low paranoia (n=69)	High paranoia (n=18)	Low paranoia (n=56)	High paranoia (n=16)	Low paranoia (n=64)	High paranoia (n=19)	Statistic	*P*-value	Statistic	*P*-value	Statistic	*P*-value
Demographics														
Age (years)	36.5 [1.5]	35.4 [2.4]	36.2 [1.4]	39.5 [2.8]	38.6 [1.6]	32.9 [1.7]	37.6 [1.3]	30.7 [1.6]	1.12 (3)[††]	0.342	3.20 (1)[††]	0.075	2.62 (3)[††]	0.051
Gender									7.29 (6)‡	0.238[§]	1.37 (2)‡	0.503[§]	n/a	n/a
% Female	44.4%	45.0%	47.8%	50.0%	50.0%	62.5%	57.8%	73.7%	n/a	n/a	n/a	n/a	n/a	n/a
% Male	55.6%	55.0%	50.7%	50.0%	50.0%	37.5%	42.2%	26.3%	n/a	n/a	n/a	n/a	n/a	n/a
% Other or not specified	0.0%	0.0%	1.4%	0.0%	0.0%	0.0%	0.0%	0.0%	n/a	n/a	n/a	n/a	n/a	n/a
Education									15.94 (21)‡	0.812[||]	7.33 (7)‡	0.4[§]	n/a	n/a
% High school degree or equivalent	17.8%	20.0%	13.0%	16.7%	16.1%	6.3%	25.0%	10.5%	n/a	n/a	n/a	n/a	n/a	n/a
% Some college or university, no degree	22.2%	30.0%	24.6%	22.2%	17.9%	25.0%	25.0%	26.3%	n/a	n/a	n/a	n/a	n/a	n/a
% Associate degree	13.3%	15.0%	17.4%	22.2%	12.5%	12.5%	9.4%	21.1%	n/a	n/a	n/a	n/a	n/a	n/a
% Bachelor's degree	33.3%	35.0%	40.6%	22.2%	35.7%	56.3%	28.1%	31.6%	n/a	n/a	n/a	n/a	n/a	n/a
% Master's degree	8.9%	0.0%	2.9%	0.0%	14.3%	0.0%	7.8%	10.5%	n/a	n/a	n/a	n/a	n/a	n/a
% Doctorate or professional degree	4.4%	0.0%	0.0%	5.6%	1.8%	0.0%	1.6%	0.0%	n/a	n/a	n/a	n/a	n/a	n/a
% Completed some postgraduate	0.0%	0.0%	1.4%	5.6%	1.8%	0.0%	3.1%	0.0%	n/a	n/a	n/a	n/a	n/a	n/a
% Other / not specified	0.0%	0.0%	0.0%	5.6%	0.0%	0.0%	0.0%	0.0%	n/a	n/a	n/a	n/a	n/a	n/a
Income									15.0 (18)‡	.671[||]	1.18 (6)‡	0.981[§]	n/a	n/a
Less than $20,000	24.4%	25.0%	24.6%	33.3%	17.9%	37.5%	23.4%	15.8%	n/a	n/a	n/a	n/a	n/a	n/a
$20,000 to $34,999	40.0%	25.0%	20.3%	22.2%	33.9%	31.3%	28.1%	31.6%	n/a	n/a	n/a	n/a	n/a	n/a
$35,000 to $49,999	15.6%	15.0%	18.8%	16.7%	12.5%	6.3%	18.8%	15.8%	n/a	n/a	n/a	n/a	n/a	n/a
$50,000 to $74,999	13.3%	35.0%	20.3%	5.6%	21.4%	12.5%	18.8%	21.1%	n/a	n/a	n/a	n/a	n/a	n/a
$75,000 to $99,999	4.4%	0.0%	7.2%	11.1%	8.9%	6.3%	7.8%	15.8%	n/a	n/a	n/a	n/a	n/a	n/a
Over $100,000	0.0%	0.0%	5.8%	5.6%	3.6%	6.3%	1.6%	0.0%	n/a	n/a	n/a	n/a	n/a	n/a
Not specified	2.2%	0.0%	2.9%	5.6%	1.8%	0.0%	1.6%	0.0%	n/a	n/a	n/a	n/a	n/a	n/a
Cognitive Reflection									11.92 (9)‡	0.223[||]	7.00 (3)‡	0.071[§]	n/a	n/a
Ethnicity									5.16 (3)‡	0.157[§]	3.72 (1)‡	0.069[§]	n/a	n/a
% Hispanic, Latino, or Spanish origin	4.4%	15.0%	1.4%	0.0%	8.9%	6.3%	1.6%	15.8%	n/a	n/a	n/a	n/a	n/a	n/a
% Not of Hispanic, Latino, or Spanish origin	95.6%	85.0%	98.6%	100.0%	91.1%	93.8%	98.4%	84.2%	n/a	n/a	n/a	n/a	n/a	n/a
Race									19.56 (15)‡	.173[||]	9.63 (5)‡	0.084[§]	n/a	n/a
% White	82.2%	75.0%	84.1%	88.9%	85.7%	75.0%	85.9%	73.7%	n/a	n/a	n/a	n/a	n/a	n/a
% Black or African American	6.7%	15.0%	5.8%	11.1%	0.0%	12.5%	4.7%	10.5%	n/a	n/a	n/a	n/a	n/a	n/a
% Asian	8.9%	10.0%	7.2%	0.0%	3.6%	6.3%	7.8%	0.0%	n/a	n/a	n/a	n/a	n/a	n/a
% American Indian or Alaska Native	0.0%	0.0%	0.0%	0.0%	1.8%	6.3%	0.0%	0.0%	n/a	n/a	n/a	n/a	n/a	n/a
% Multiracial	2.2%	0.0%	1.4%	0.0%	3.6%	0.0%	1.6%	15.8%	n/a	n/a	n/a	n/a	n/a	n/a
% Other / not specified	0.0%	0.0%	1.4%	0.0%	5.4%	0.0%	0.0%	0.0%	n/a	n/a	n/a	n/a	n/a	n/a
Mental health														
Psychiatric diagnosis									10.78 (9)‡	0.292[||]	2.96 (3)‡	0.361[§]	n/a	n/a
% No psychiatric diagnosis	73.3%	80.0%	60.9%	55.6%	71.4%	50.0%	65.6%	42.1%	n/a	n/a	n/a	n/a	n/a	n/a
% Schizophrenia spectrum	2.2%	0.0%	0.0%	0.0%	0.0%	6.3%	0.0%	0.0%	n/a	n/a	n/a	n/a	n/a	n/a
% Mood disorder	13.3%	15.0%	27.5%	22.2%	21.4%	43.8%	26.6%	31.6%	n/a	n/a	n/a	n/a	n/a	n/a
% Not specified	11.1%	5.0%	11.6%	22.2%	7.1%	0.0%	7.8%	26.3%	n/a	n/a	n/a	n/a	n/a	n/a
% Medicated	8.9%	10.0%	13.0%	22.2%	7.1%	31.3%	14.1%	10.5%	3.58 (6)‡	0.744[§]	4.164 (2)‡	0.121[§]	n/a	n/a
Beck's Anxiety Inventory	0.34 [0.06]	0.52 [0.14]	0.31 [0.04]	0.6 [0.13]	0.24 [0.04]	0.90 [0.20]	0.33 [0.06]	0.79 [0.18]	1.24 (3)[††]	0.294	38.75 (1)[††]	**1.63E-09**	2.58 (3)[††]	0.054
Beck's Depression Inventory	0.36 [0.07]	0.86 [0.15]	0.32 [0.05]	0.79 [0.13]	0.25 [0.04]	1.03 [0.19]	0.38 [0.07]	1.06 [0.20]	1.02 (3)[††]	0.382	74.53 (1)[††]	**3.62E-16**	1.09 (3)[††]	0.354
SCID Paranoia Personality Score	0.11 [0.02]	0.67 [0.04]	0.11 [0.02]	0.61 [0.03]	0.1 [0.02]	0.72 [0.04]	0.11 [0.02]	0.65 [0.03]	1.30 (3)[††]	0.276	879.38 (1)[††]	**4.81E-91**	2.02 (3)[††]	0.111
Reversal learning performance														
Total points earned	8656.7 [182.9]	8372.5 [405.2]	6045.7 [135.7]	6266.7 [288.0]	7533.0 [143.8]	6503.1 [340.6]	7171.1 [175.6]	6510.5 [403.6]	32.29 (3)[††]	**4.16E-18**	6.18 (1)[††]	**0.0135**	2.26 (3)[††]	0.082
Total reversals achieved	7.2 [0.3]	6.5 [0.5]	5.5 [0.3]	5.7 [0.5]	6.3 [0.3]	4.9 [0.8]	5.9 [0.3]	4.8 [0.6]	4.33 (3)[††]	**0.005**	5.76 (1)[††]	**0.017**	1.10 (3)[††]	0.349
% Achieving reversals	100.0%	100.0%	98.6%	94.4%	100.0%	87.5%	96.9%	94.7%	2.26 (3)‡	0.598[§]	4.40 (1)‡	0.058[§]	n/a	n/a
Win-switch rate, block 1 (90-50-10)	0.09 [0.03]	0.09 [0.04]	0.07 [0.01]	0.11 [0.05]	0.04 [0.01]	0.13 [0.05]	0.1 [0.03]	0.21 [0.06]	2.28 (3)[††]	0.079	7.12 (1)[††]	**0.008**	1.15 (3)[††]	0.329
Win-switch rate, block 2 (80-40-20)	0.05 [0.02]	0.08 [0.03]	0.04 [0.01]	0.05 [0.04]	0.02 [0.01]	0.12 [0.05]	0.06 [0.02]	0.15 [0.05]	2.07 (3)[††]	0.105	9.92 (1)[††]	**0.002**	1.17 (3)[††]	0.32
Lose-stay rate, block 1 (90-50-10)	0.27 [0.03]	0.34 [0.05]	0.37 [0.03]	0.34 [0.04]	0.3 [0.03]	0.39 [0.06]	0.32 [0.03]	0.34 [0.04]	0.56 (3)[††]	0.641	1.83 (1)[††]	0.177	0.75 (3)[††]	0.521
Lose-stay rate, block 2 (80-40-20)	0.28 [0.03]	0.23 [0.05]	0.4 [0.03]	0.32 [0.05]	0.33 [0.03]	0.37 [0.06]	0.29 [0.03]	0.33 [0.06]	2.47 (3)[††]	0.062	0.18 (1)[††]	0.674	0.83 (3)[††]	0.476
Reaction time, block 1	433.6 [28.8]	789.3 [282.7]	548.1 [77.8]	365.6 [26.4]	448 [60.1]	442.1 [59.5]	557.2 [108.2]	530 [130.2]	0.79 (3)[††]	0.499	0.16 (1)[††]	0.689	1.73 (3)[††]	0.161
Reaction time, block 2	370.7 [23.3]	494.3 [88.6]	465.3 [61.6]	331.4 [22.9]	391.7 [52.3]	555.9 [121.2]	385.4 [29.2]	504.1 [82.7]	0.39 (3)[††]	0.757	1.92 (1)[††]	0.167	1.95 (3)[††]	0.122
U-value‡‡	0.798 [0.009]	0.81 [0.01]	0.868 [0.007]	0.871 [0.01]	0.824 [0.008]	0.894 [0.02]	0.837 [0.007]	0.877 [0.01]	13.61 (3)	**2.42E-08**	15.28 (1)	**0.0001**	3.44 (3)	**0.017**
Model parameters‡‡														
ω_3	-0.537 [0.12]	-0.736 [0.17]	-1.04 [0.93]	-0.821 [0.18]	-0.663 [0.10]	-0.898 [0.19]	-0.912 [0.10]	-0.993 [0.18]	2.06 (3)	0.105	0.50 (1)	0.481	1.01 (3)	0.391
$\propto_3^0$	-1.001 [0.19]	-0.721 [0.29]	-0.402 [0.16]	-0.804 [0.30]	-1.089 [0.17]	-0.180 [0.32]	-0.401 [0.16]	-0.067 [0.30]	2.32 (3)	0.075	45.08 (1)	0.108	2.33 (3)	0.075
κ	0.480 [0.010]	0.490 [0.015]	0.528 [0.008]	0.503 [0.016]	0.470 [0.009]	0.538 [0.017]	0.525 [0.009]	0.543 [0.016]	5.06 (3)	**0.002**	3.60 (1)	0.059	4.18 (3)	**0.006**
ω_2	1.102 [0.177]	1.017 [0.265]	0.330 [0.143]	0.590 [0.280]	1.246 [0.158]	0.252 [0.296]	0.603 [0.148]	0.074 [0.272]	4.16 (3)	**0.007**	4.44 (1)	**0.036**	2.81 (3)	**0.04**

Version columns display means [standard error] or percentage of participants within the described category.
[††] Univariate analysis, F(df).
[‡] Exact test, chi-square coefficient (df).
[§] Exact significance (2-sided).
[||] Monte Carlo significance (2-sided).
[‡‡] Data presented in Fig. 3; repeated measures ANOVA, *F*(df), *P*. Mean values collapsed across blocks.

Table 3. Summary of Paranoia / Methamphetamine Effects on Belief-Updating

	In lab	Online	Rats	Meta-analysis[†]
ω_3	↓[‡]	⇣	⬇	⬇
μ_3^0	⬆	⬆[§¶]	⬆	⬆
κ	⬆	⬆[§]	⬆	⬆
ω_2	⬇	⬇[§#]	⬇	⬇

⇡ ⇣ Non-significant increase/decrease in high paranoia or meth, relative to low paranoia or saline

↑ ↓ Trend-level increase/decrease in high paranoia or meth, relative to low paranoia or saline

⬆ ⬇ Significantly higher/lower in high paranoia or meth, relative to low paranoia or saline

- - No significant findings or trends

[†] Random effects, consistent findings for mean difference (In lab and Online) and standardized mean difference (In lab, Online, and Rat)

[‡] Baseline trend; parameter decreases in second block for low but not high paranoia

[§] Version 3 only

[¶] Trend-level significance disappears with inclusion of demographic covariates

[#] Significance reduced to trend with inclusion of demographic covariates

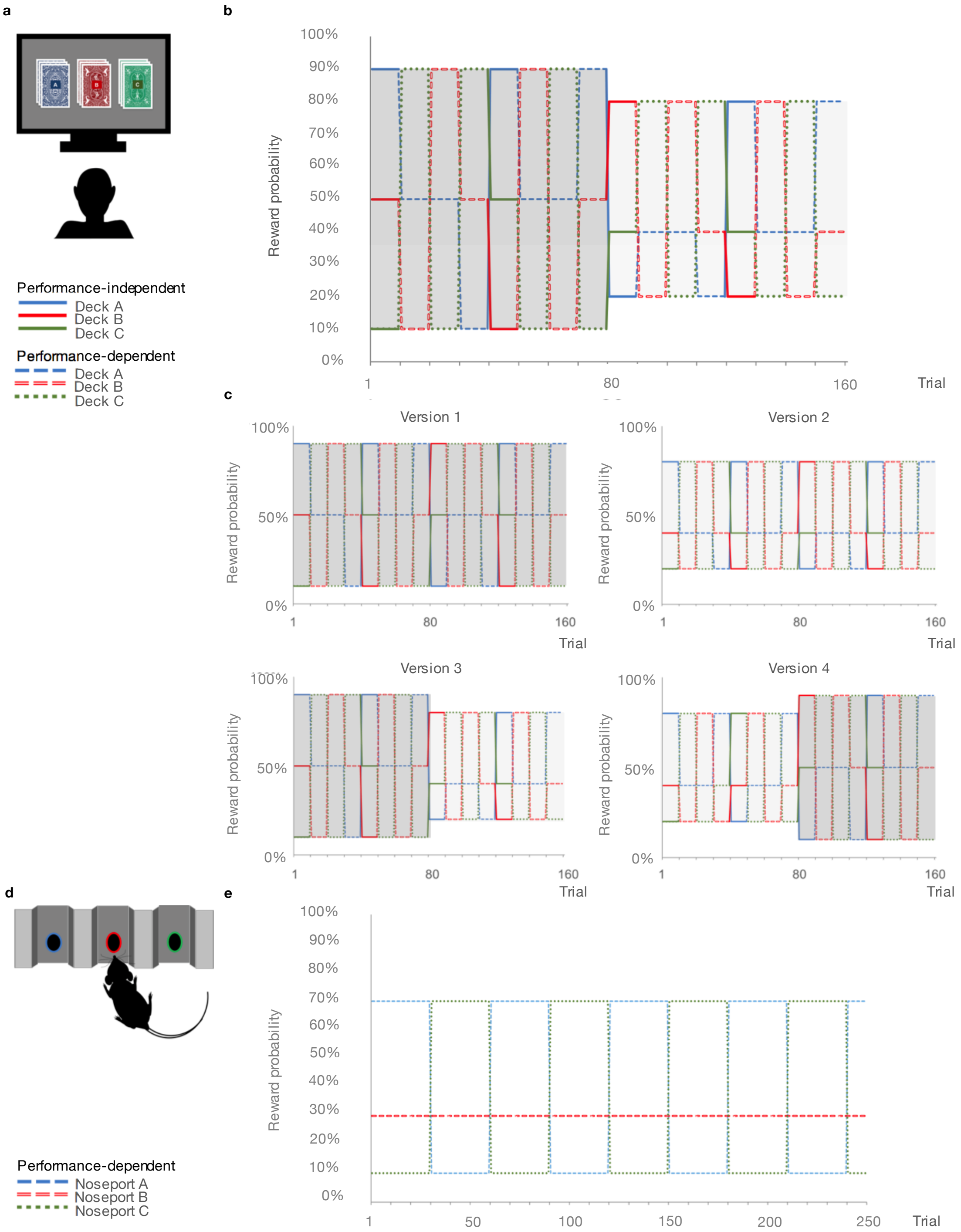


Fig. 1. Probabilistic reversal learning task. a, Human paradigm: participants choose between three ecks of cards with different, unknown probabilities of reward and loss. **b,** Reward contingency schedule for in laboratory experiment. On trial 81, the probability context shifts from 90%, 50%, and 10% (dark grey) to 80%, 40%, and 20% without warning (light grey). **c**, Reward contingency schedules for online experiment. **d,** Rat paradigm: subjects choose between three noseports with different probabilities of sucrose pellet reward. **e,** Reward contingency schedule for rat experiment[39].

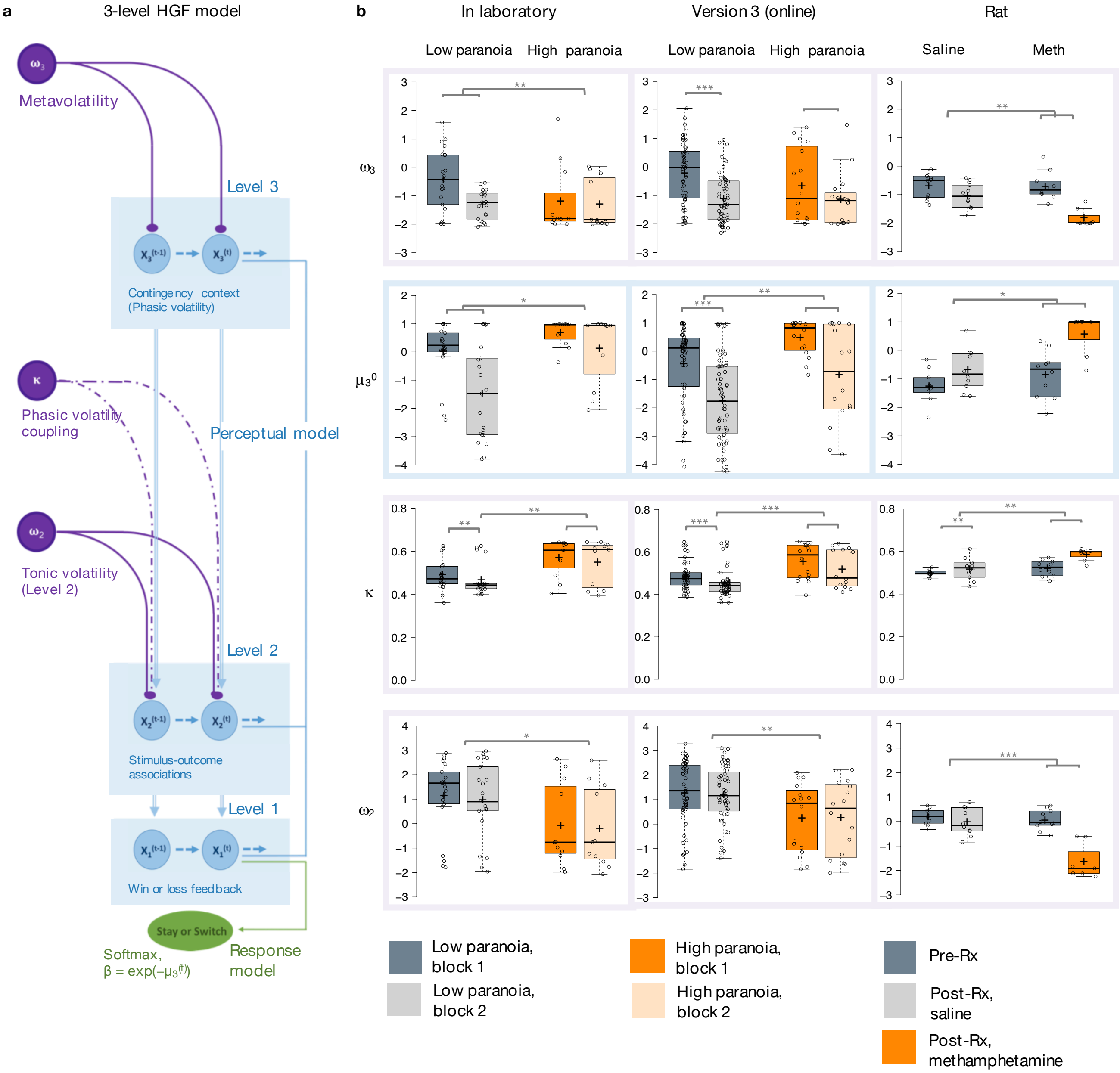


Fig. 2. Hierarchical Gaussian Filter (HGF) model parameters. a, 3-level HGF perceptual model (blue) with a softmax decision model (green). Level 1 (x_1) corresponds to trial-by-trial perception of win or loss feedback. Level 2 (x_2) represents stimulus-outcome associations (e.g., deck values). Level 3 (x_3) models perception of the overall reward contingency context. The impact of phasic volatility upon x_2 is captured by κ, the coupling parameter. Tonic volatility modulates x_3 and x_2 via ω_3 and ω_2, respectively. μ_3^0 is the initial value of the third level volatility belief. **b,** Estimated parameters replicate across high paranoia groups (orange) in the in laboratory experiment (n=21 low paranoia, 11 high paranoia); analogous online task (version 3, n=56 low paranoia, 16 high paranoia); and rats exposed to chronic, escalating methamphetamine (n=10 per group Pre-Rx; n=10 saline, 7 methamphetamine Post-Rx). Error bars denote standard error (SEM); $^*p \leq 0.05$, $^{**}p \leq 0.01$, $^{***}p \leq 0.001$.

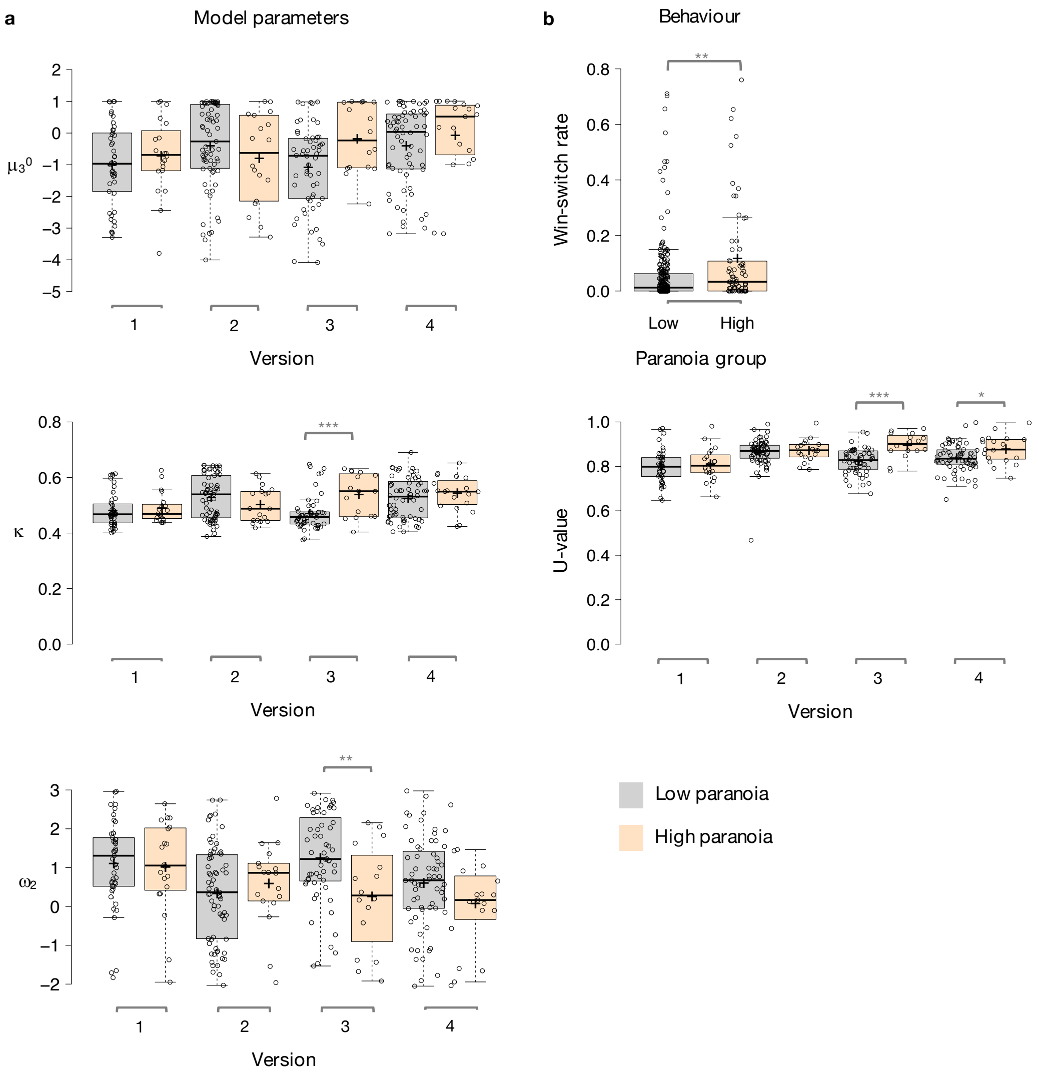


Fig. 3. Paranoia effects across task versions. a, HGF parameters μ_3^0, κ, and ω_2 show version 3 specific trends and effects of paranoia group membership (Experiment 2, n=234 low paranoia, 73 high paranoia, collapsed across task versions). **b,** Behaviourally, paranoid participants switched between decks more frequently after positive feedback, across all task versions and blocks (paranoia group effect; version trend, p=0.081). In versions 3 and 4 only, paranoid participants showed higher U-values, suggesting increasingly stochastic switching rather than perseverative returns to a previously rewarding option. Error bars denote standard error (SEM); *p ≤ 0.05, **p ≤ 0.01, ***p ≤ 0.001.

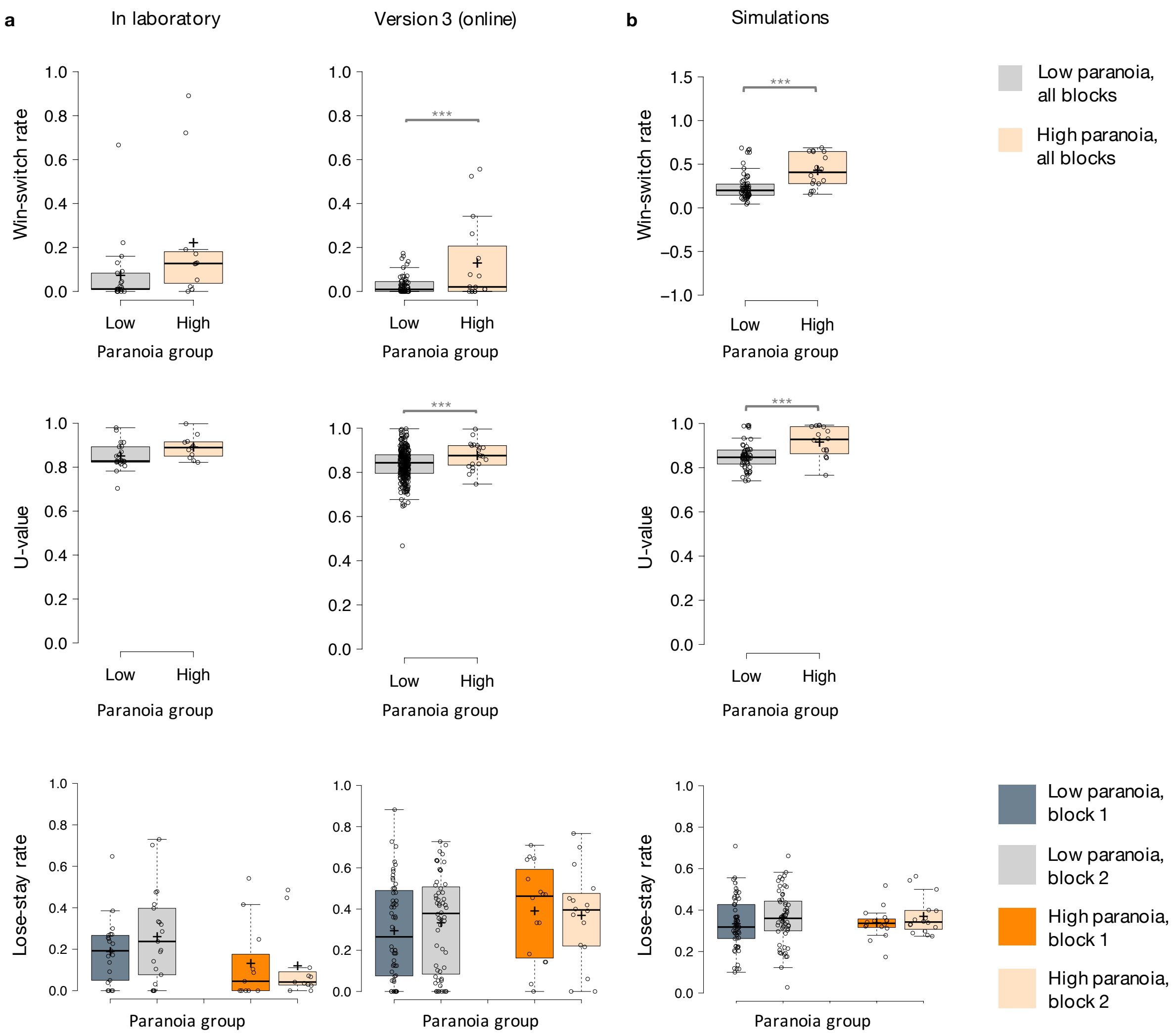


Fig. 4. Behavioural data and simulations. a, Behavioural switching patterns replicate across in laboratory and online version 3 experiments (win-switch: in laboratory paranoia group trend, p=0.068; version 3 paranoia effect; U-value: in laboratory paranoia group trend, p=0.079; version 3 paranoia effect). Perseveration after negative feedback (lose-stay behaviour) did not significantly differ between paranoia groups or task block. **b,** Simulated data generated from HGF perceptual parameters (version 3) replicates win-switch and U-value behaviours (win-switch paranoia effect; U-value paranoia effect). Ten simulations were performed per subject. Rates and U-values were averaged across simulations. Error bars denote standard error (SEM); n=21 low paranoia, 11 high paranoia (in laboratory); n=56 low paranoia, 16 high paranoia (online, version 3); *p ≤ 0.05, **p ≤ 0.01, ***p ≤ 0.001.

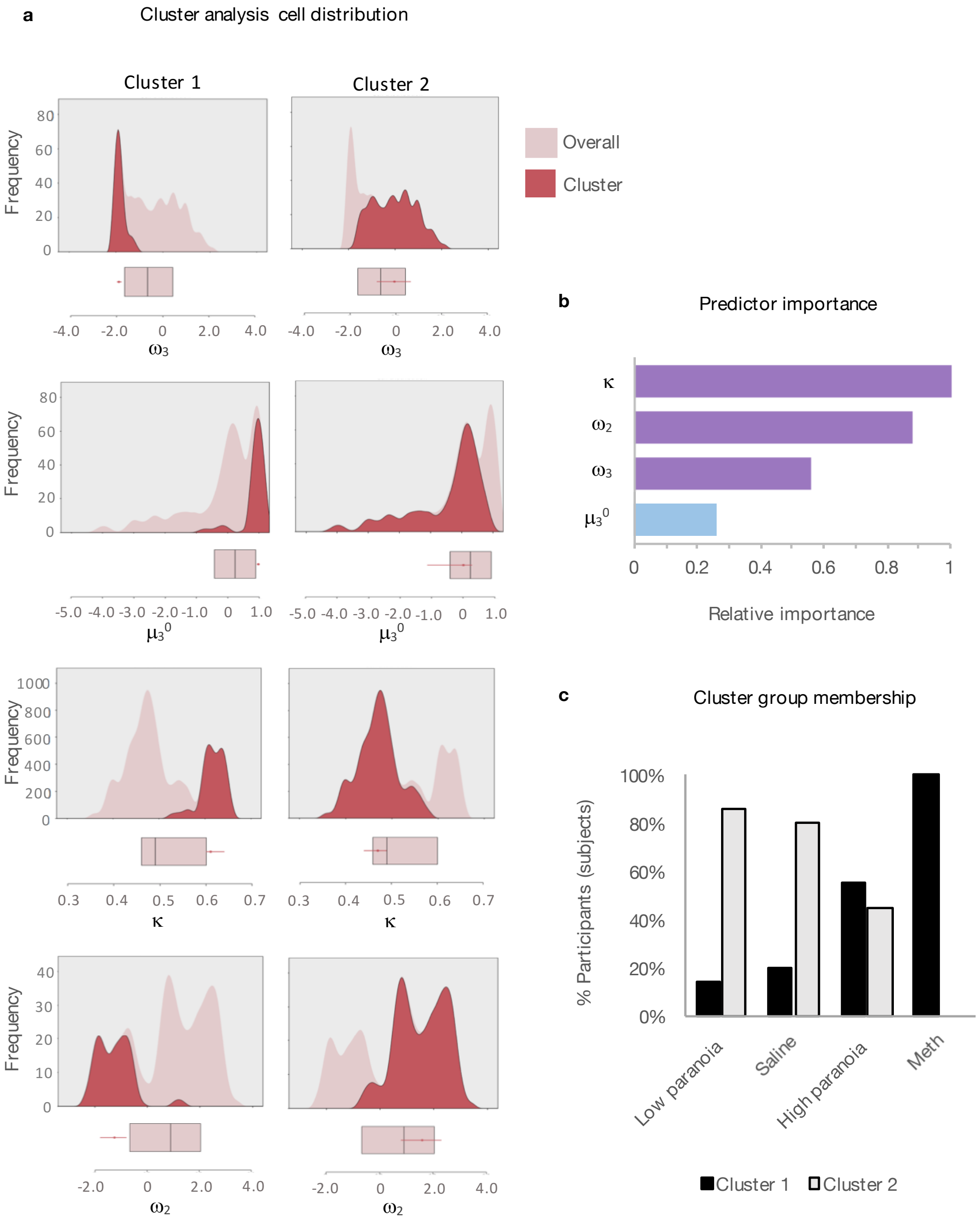
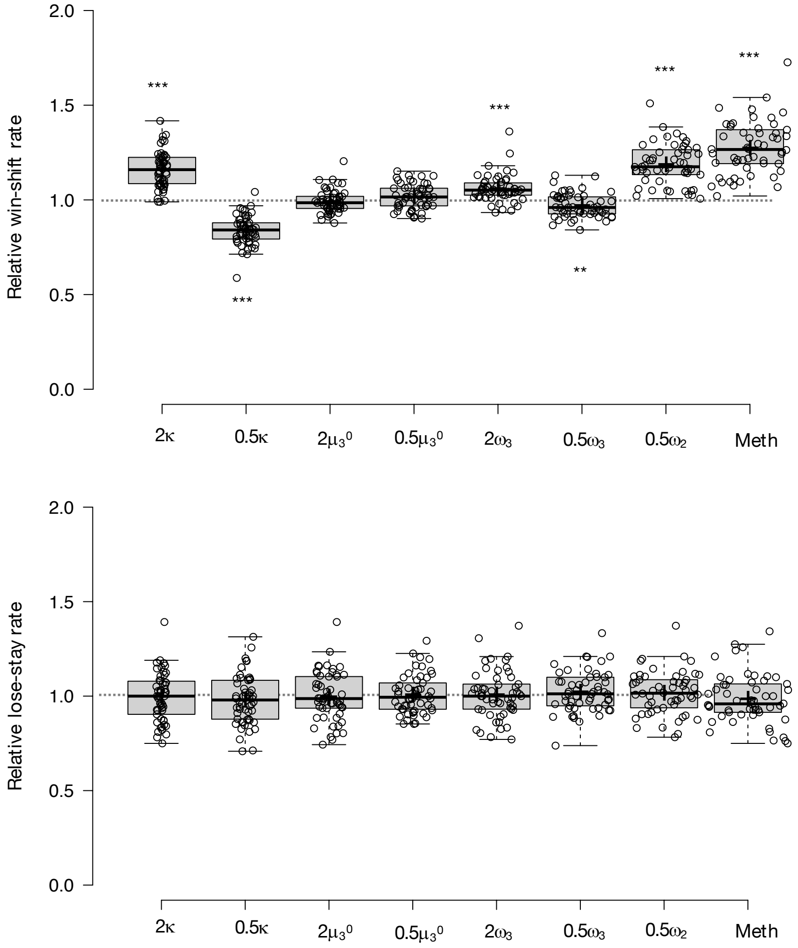
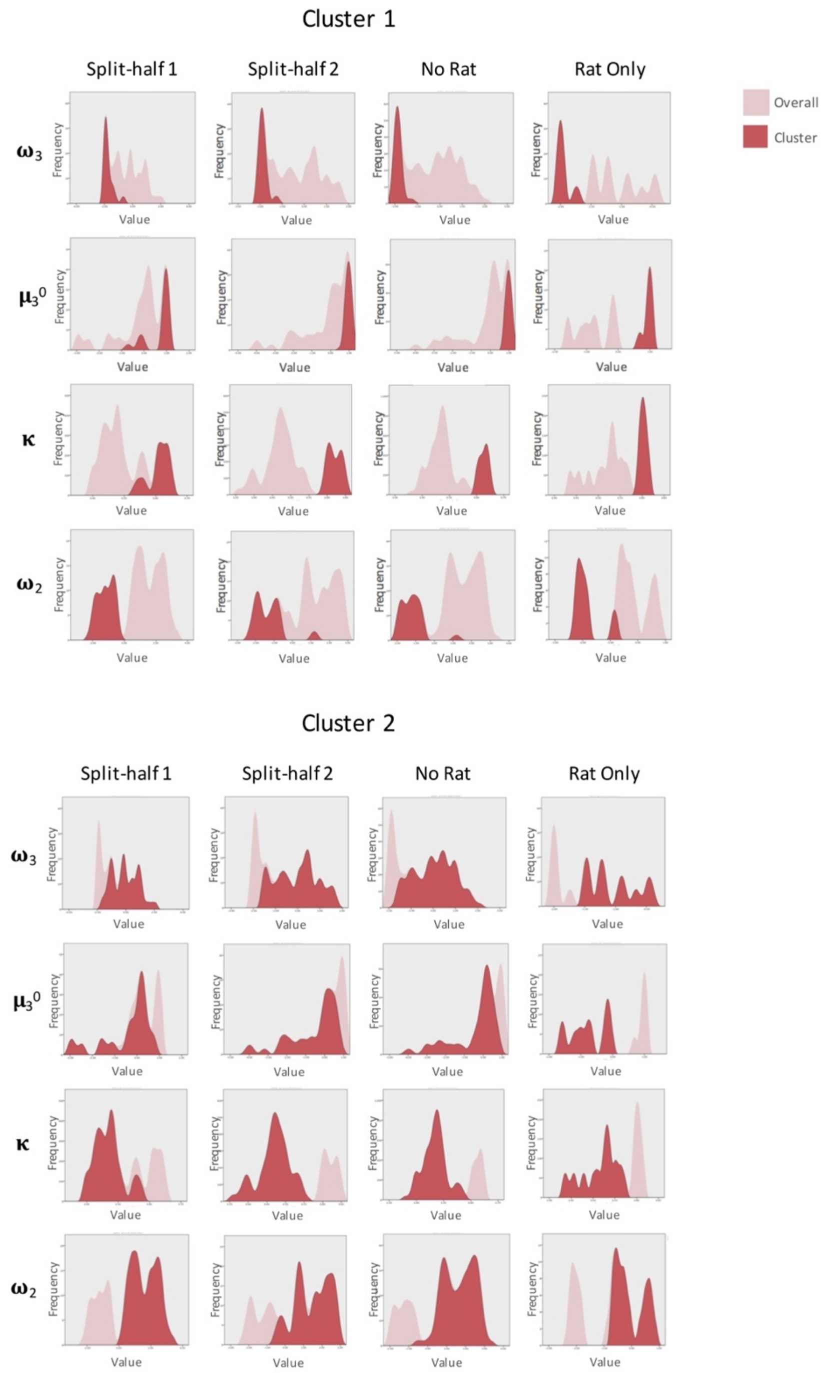


Fig. 5. Cluster analysis of HGF parameters. Two-step cluster analysis of model parameters across rat and human data sets (rat, post-Rx; in laboratory and online version 3, block 1). Automated clustering yielded an optimal two clusters with good cohesion and separation (average silhouette coefficient=0.7; cluster size ratio=2.46). **a,** Parameter density plots for overall distributions (light pink) and cluster-specific distributions (red). Box-plots of overall median, 25th quartile, and 75th quartile are aligned below each plot (pink), with cluster medians and quartiles superimposed (red). Relative to the overall distribution, Cluster 1 (n=35) medians are elevated for μ_3^0 and κ, decreased for ω_2 and ω_3. Cluster 2 (n=86) falls within each overall distribution. **b,** Predictor importance of included parameters. **c,** Distribution of cluster identities within groups (low paranoia: n=77; high paranoia: n=27; rat-saline: n=10; rat-methamphetamine: n=7). Cluster 1 membership is significantly associated with paranoia and methamphetamine groups (χ^2(1, n=121)=29.447, p=5.75E-8).



Supplementary Fig. 1. Parameter effects on simulated task performance. We simulated behaviour from low paranoia participants (online Version 3, n=54) to evaluate the effects of $\boldsymbol{\kappa}$, $\boldsymbol{\mu}_3^0$, $\boldsymbol{\omega}_2$, and $\boldsymbol{\omega}_3$ on win-shift and lose-stay rates. Estimated perceptual parameters were averaged across subjects to create a single set of baseline parameters. Additional parameter sets were created by doubling or halving one parameter at a time (e.g., 2 $\boldsymbol{\kappa}$ or 0.5 $\boldsymbol{\kappa}$), while the others were held constant (n.b., 2 $\boldsymbol{\omega}_2$ violated model assumptions and was excluded from analysis). We also included the average parameter values of rats exposed to methamphetamine (Meth). Ten simulations were run per subject for each condition (i.e., parameter set). Win-shift and lose-stay rates were calculated, then averaged across simulations and subjects. Rates from each condition were divided by the baseline condition rate to generate relative win-shift and lose-stay rates. We compared relative rates for each condition to the baseline (relative rate of 1; paired t-tests, Bonferroni-corrected p-values). Baseline parameters were positive for $\boldsymbol{\kappa}$ and $\boldsymbol{\omega}_2$, and negative for $\boldsymbol{\mu}_3^0$ and $\boldsymbol{\omega}_3$. Consequently, the doubled (2x) condition makes $\boldsymbol{\mu}_3^0$ and $\boldsymbol{\omega}_3$ more negative (lower). (n=54). Error bars denote standard error (SEM); $*p \leq 0.05$, $**p \leq 0.01$, $***p \leq 0.001$.



Supplementary Fig. 2. Cluster validation. We replicated our 2-cluster solution (Fig. 4) by independently running two-step cluster analyses on separate halves of the data (Split-half 1, Split-half 2), removing the rat data and running the human data only (No Rat), and running the rat data alone (Rat Only). In each condition, we identified two clusters with good cohesion and separation (Split-half 1, n=19 cluster 1, 42 cluster 2: silhouette coefficient = 0.6; Split-half 2, n = 17 cluster 1, 43 cluster 2: silhouette coefficient = 0.7; No Rat, n=26 cluster 1, 78 cluster 2: silhouette coefficient = 0.7; Rat Only, n=6 cluster 1, 11 cluster 2: silhouette coefficient = 0.7). All variables showed predictor importance above 0.2 with some variation in order of importance (Split-half 1, $\boldsymbol{\omega_2} > \boldsymbol{\kappa} > \boldsymbol{\omega_3} > \boldsymbol{\mu_3}^0$; Split-half 2, $\boldsymbol{\kappa} > \boldsymbol{\omega_2} > \boldsymbol{\omega_3} > \boldsymbol{\mu_3}^0$; No Rat, $\boldsymbol{\kappa} > \boldsymbol{\omega_2} > \boldsymbol{\omega_3} > \boldsymbol{\mu_3}^0$; Rat Only, $\boldsymbol{\mu_3}^0 > \boldsymbol{\omega_2} > \boldsymbol{\omega_3} > \boldsymbol{\kappa}$). Predictor importance was weighted more evenly across variables in the Rat Only condition; all variables showed predictor importance above