





Visualizing threat and trustworthiness prior beliefs in face perception in high versus low paranoia

Antonia Bott¹  , Hanna C. Steer¹, Julian L. Faße¹ and Tania M. Lincoln¹ 

Predictive processing accounts of psychosis conceptualize delusions as overly strong learned expectations (prior beliefs) that shape cognition and perception. Paranoia, the most prevalent form of delusions, involves threat prior beliefs that are inherently social. Here, we investigated whether paranoia is related to overly strong threat prior beliefs in face perception. Participants with subclinical levels of high ($n = 109$) versus low ($n = 111$) paranoia viewed face stimuli paired with written descriptions of threatening versus trustworthy behaviors, thereby activating their threat versus trustworthiness prior beliefs. Subsequently, they completed an established social-psychological reverse correlation image classification (RCIC) paradigm. This paradigm used participants' responses to randomly varying face stimuli to generate individual classification images (ICIs) that intend to visualize either facial prior belief (threat vs. trust). An independent sample ($n = 76$) rated these ICIs as more threatening in the threat compared to the trust condition, validating the causal effect of prior beliefs on face perception. Contrary to expectations derived from predictive processing accounts, there was no evidence for a main effect of paranoia. This finding suggests that paranoia was not related to stronger threat prior beliefs that directly affected face perception, challenging the assumption that paranoid beliefs operate on a perceptual level.


Schizophrenia (2024)10:40; <https://doi.org/10.1038/s41537-024-00459-z>

INTRODUCTION

Delusions, characterized as fixed beliefs that persist despite lacking or conflicting evidence¹, represent a core symptom of schizophrenia and other psychotic disorders. They predominantly revolve around social themes², with the most prevalent type, paranoid delusions, involving the belief that others intend to persecute or harm oneself^{3,4}. While manifest paranoid delusions are typically associated with significant distress, they exist on a continuum with milder forms of paranoid beliefs found in the general population^{5–7}. Over the past decades, a multitude of theoretical models have proposed various risk factors contributing to the formation of paranoid beliefs, including cognitive reasoning biases^{8–10} and social risk factors^{8,9}. Specifically, paranoid beliefs have been consistently associated with premorbid social adversity experiences (e.g., interpersonal childhood trauma)^{11,12}, mediated by learned negative beliefs about others¹³. In recent years, our mechanistic understanding of delusions has significantly progressed through the lens of predictive processing accounts. According to this framework, the brain generates predictions about upcoming sensory inputs (*prior beliefs*) and integrates them with the observed sensory inputs to refine future predictions^{14–17}. This integration is weighted by the relative certainty (*precision*¹⁸) assigned to both components. The greater the relative precision assigned to the prior belief, the less the observed inputs impact the final percept (*posterior belief*), and vice versa¹⁶. A compelling illustration is face pareidolia, where the perception of faces in inanimate objects could be evoked by highly precise prior beliefs for facial features^{19,20}. Within the predictive processing framework, delusions are proposed to arise as overly precise prior beliefs formed to resolve the chronic uncertainty associated with sensory inputs^{21–24}. Similarly, paranoid delusions can be reconceptualized as precise *threat prior beliefs*, sculpting the individual's perception as if viewing the world – including other people – through “threat-colored glasses”.

Supporting this conceptualization, individuals with paranoid beliefs have been found to misclassify faces with neutral emotional expressions as angry^{25–27} and rate them as less trustworthy^{28–30} compared to those without such beliefs (but see^{31–33}). However, these findings are based on explicit ratings derived from individuals' *percepts*, thus failing to disentangle the relative impact of prior beliefs and sensory inputs on perception. Previous studies attempting to isolate the impact of prior beliefs on perception in individuals with psychotic symptoms and delusion-proneness utilized various signal detection paradigms^{34–39}. These paradigms typically involved the detection of a specific signal within ambiguous non-social stimuli, leveraging experimentally induced prior beliefs to resolve sensory ambiguity. For instance, participants could rely on cues previously associated with leftward versus rightward rotation to detect the ambiguous rotation direction of dot spheres³⁶. The severity of psychotic symptoms positively correlated with the reliance on these prior beliefs in determining rotation directions³⁶, consistent with the concept of overly precise prior beliefs. However, despite the predominantly social nature of delusional beliefs, investigations into the imbalanced integration during perceptual inference in the social domain are scarce^{40–42}. Moreover, existing studies focused on detecting the mere presence of hidden social stimuli, such as a person in a two-tone image⁴¹ or faces within visual noise patterns⁴². Thus, it remains unclear whether paranoid beliefs condense in overly precise threat prior beliefs, shaping the perception of ambiguous social sensory inputs.

We addressed this objective within the domain of face perception, utilizing the established social-psychological reverse correlation image classification paradigm (RCIC^{43,44}; for reviews, see^{45,46}). This data-driven signal detection technique enables the visualization of individuals' mental representations of a face with specific emotional states or traits (e.g., anger, trustworthiness). In the standard RCIC paradigm, participants view pairs of ambiguous

¹Clinical Psychology and Psychotherapy, Faculty of Psychology and Human Movement Science, Universität Hamburg, Hamburg, Germany. email: antonia.bott@uni-hamburg.de

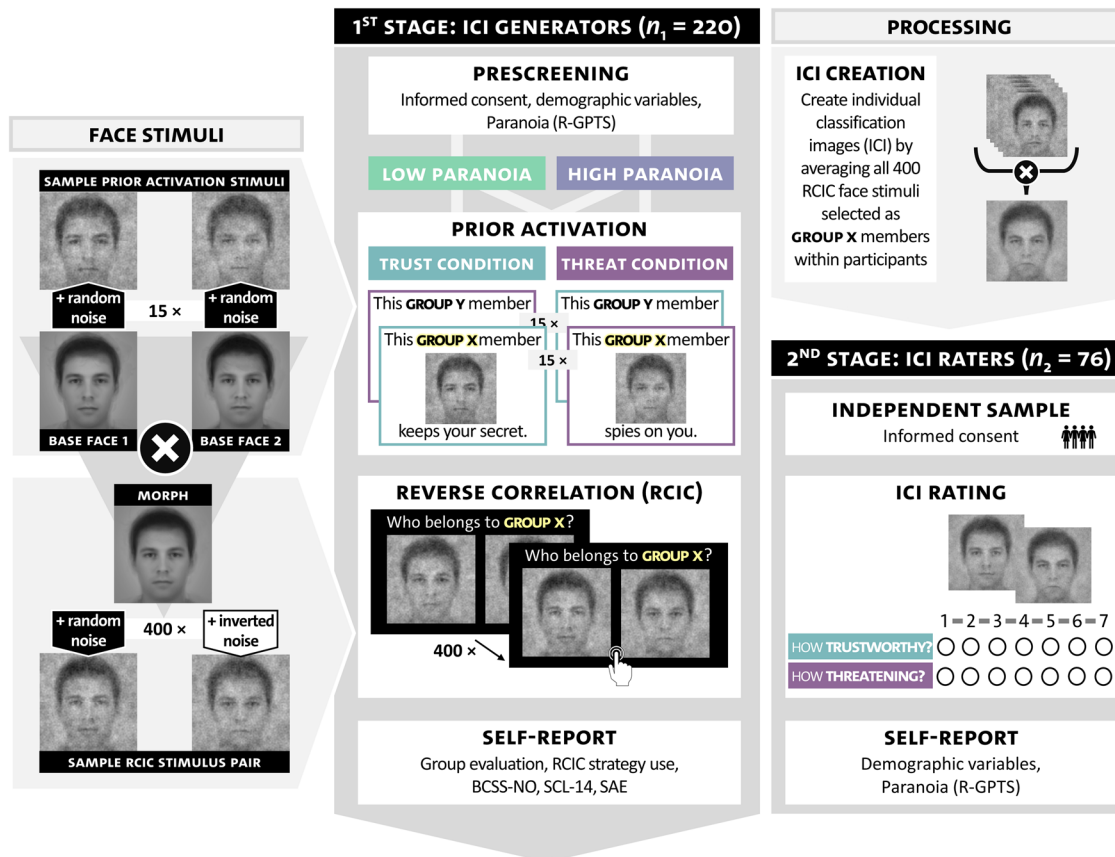


Fig. 1 Procedure flow chart and face stimuli used during prior activation and reverse correlation image classification (RCIC). *BCSS-NO* Negative beliefs about others, *ICI* individual classification image, *R-GPTS* Revised Green Paranoid Thoughts Scale, *SAE* Social adversity experiences, *SCL-GSI* Symptom Checklist.

faces in a series of trials, selecting the face they deem to best represent a designated state or trait (e.g., “Who is more trustworthy?”). In contrast to traditional signal detection paradigms, all sensory inputs vary randomly (i.e., one constant ‘base face’ is superimposed with two unique random visual noise patterns in each trial). This randomness ensures that stimulus selection is guided by participants’ individual mental representations, such that averaging all selected stimuli produces individual classification images (ICIs) that distill the features guiding their selections⁴⁶. In other words, RCIC offers a nuanced visualization of facial prior beliefs (e.g., of a trustworthy face).

In the present study, we applied a variant of this paradigm⁴⁷ to investigate whether the impact of threat prior beliefs on the perception of ambiguous faces is stronger in individuals with high versus low paranoia. Furthermore, we examined whether this effect is specific to threat prior beliefs or extends to trustworthiness prior beliefs (i.e., a primary dimension in face perception and evaluation with contrasting valence^{46,48}). In this variant, participants initially viewed face stimuli labeled as ‘members of two fictitious groups’ (*Group X* and *Group Y*), paired with written descriptions of threatening versus trustworthy behaviors (e.g., “This Group X member spies on you” vs. “This Group Y member keeps a secret you told them”). In a subsequent RCIC paradigm, they selected those faces they deemed most likely to depict a member of one of these groups (similar to⁴⁷). As all participants viewed identical face stimuli, any systematic variation in the appearance of the resulting ICIs can be attributed to individual differences in the extent to which facial prior beliefs, activated by the provided behavioral descriptions, affected the perception of ambiguous faces. Consistent with previous findings⁴⁷, we expected the ICIs to appear more threatening following the threat

prior activation and more trustworthy following the trustworthiness prior activation (H1). Building on the conceptualization of paranoid beliefs as overly precise threat prior beliefs, we expected individuals with high paranoia to generate ICIs that appear more threatening and less trustworthy as compared to individuals with low paranoia (H2). Additionally, we explored whether social adversity experiences and generalized negative beliefs about others moderated the impact of prior beliefs on face perception. We expected a larger impact of threat prior beliefs on face perception in individuals with high versus low levels of social adversity experiences (H3) and negative beliefs about others (H4).

METHODS

This preregistered study (<https://osf.io/epbw3>) was approved by the Local Ethics Committee of Universität Hamburg and was conducted in accordance with the Declaration of Helsinki.

Participants and procedure

We conducted this study as multi-stage online experiment (see Fig. 1 for an overview). Upon interest, individuals filled out a short online prescreening including demographic variables and self-reported paranoia. Eligible individuals were forwarded to the first stage, comprising a quasi-experimental factorial design with *Paranoia level* (low vs. high) and *Prior activation condition* (trust vs. threat) as between-subjects factors. Following an affective state assessment and a prior activation phase, participants engaged in a RCIC paradigm. Afterwards, they completed their participation by filling out self-report measures. In an interspersed processing stage, we created ICIs from the stimuli selected in the RCIC

paradigm. In the second stage of the study, an external sample rated the ICIs on threat and trustworthiness, constituting the primary outcome. Finally, participants from this external sample provided self-reports of paranoia and demographic variables.

Participant recruitment included social media posts, flyers, and advertisements via *Prolific* (www.prolific.com). General inclusion criteria were informed consent, minimum age of 18 years, sufficient self-reported German language skills, normal or corrected eyesight, and participation using a desktop PC or laptop. A lifetime diagnosis of neurological disorders (e.g., prosopagnosia), participation via mobile devices or failing attention checks twice led to exclusion from participation. For sample assignment in the first stage, we used the persecution subscale of the revised Green et al. Paranoid Thoughts Scales (R-GPTS^{49,50}). Individuals with a score of ≥ 11 (i.e., *moderately severe*) and ≤ 5 (i.e., *within average range*) were assigned to the high paranoia (HP) and low paranoia (LP) sample, respectively. Individuals scoring between 6 and 10 were excluded. An a priori power analysis (G*Power, Version 3.1.9.4⁵¹) based on $\alpha = 0.05$ and $1 - \beta = 0.80$ in a fixed-effects analysis of covariance (ANCOVA) indicated a required sample size of 220 participants to detect a small to moderate between-subjects effect (Cohen's $f = 0.19$). Of 279 eligible individuals who started the study, 44 failed to complete it and 11 failed the attention checks. We excluded four additional participants from all analyses (see below). The final sample thus consisted of 220 participants: 109 HP and 111 LP participants. Participants were compensated with EUR 5.80 (€5.00) for participation via Prolific (47.3%) or invited to take part in a lottery (10 online vouchers à EUR 5.00) if they did not wish to participate via Prolific (52.7%). For the second stage of the study, we recruited an external sample of 76 university students in exchange for partial course credit. In addition to the general eligibility criteria above, participants of the first stage were excluded from participation to ensure naivety regarding ICI generation including the design of the first stage.

Materials

Prior activation. We used a counterbalanced block-wise prior activation phase to activate participants' mental representations (i.e., prior beliefs) of threatening versus trustworthy faces. To this end, we exposed participants to what we called members of two fictitious groups (labeled *Group X* and *Group Y*) by presenting them with differentiable sets of face stimuli paired with unique written behavioral descriptions (15 stimulus pairs per group). For half of the participants, the behavioral descriptions paired with Group X implied the trait *threatening* (e.g., "This Group X member stares at you"), whereas the behavioral descriptions paired with Group Y implied the trait *trustworthy* (e.g., "This Group Y member respects your privacy"). For the other half, this group-trait association was reversed ($X = \text{trustworthy}$, $Y = \text{threatening}$). The face stimuli consisted of one 'base face' per group (i.e., morphs of 10 human face images each, randomly selected from the Chicago Face Database⁶⁹) which we converted to grayscale and smoothed with a Gaussian blur. Finally, we superimposed each group-specific base face with 15 unique patterns of visual noise to generate subtle variations of the same underlying face (see Fig. 1). The images did not differ with respect to relevant traits (e.g., threat, trustworthiness, masculinity, attractiveness; see Supplementary Information S1). The assignment of base faces to group, block order (Group X vs. Group Y first), and stimulus order within blocks were randomized. Participants were instructed to carefully read and memorize the presented materials. Task completion was self-paced ($Mdn = 8.53$ min, $SD = 8.08$), with a minimum presentation duration of 5 s per stimulus pair.

Reverse correlation image classification (RCIC) paradigm. Following prior activation, participants completed a two-image forced choice RCIC paradigm^{43,44,47} with 400 trials. For this paradigm, we created a perfect morph of the two group-specific base faces to

create one group-ambiguous base face (i.e., reflecting both base faces to the same extent; see Fig. 1). Next, we superimposed this morphed base face with both 400 unique random visual noise patterns and their mathematical inverses (i.e., a white pixel in the original noise pattern was black in the inverted pattern and vice versa, see Supplementary Information S1) by using the *rcicr*⁵² package for R⁵³. Thus, each of the 400 stimulus pairs reflected *random and very subtle* variations of the same underlying base face, with anti-correlated variation within stimulus pairs. In each trial, participants were presented with one of these stimulus pairs presented side-by-side against a black background (512×512 pixels) and were instructed to select the face they spontaneously deemed most likely to *depict a Group X* member (with "Who belongs to Group X" presented above the stimuli). Note that Group X was either paired with threat-implicating behavioral descriptions (*threat* condition) or trustworthiness-implicating behavioral descriptions (*trust* condition) during prior activation. Task completion was self-paced ($Mdn = 16.59$ min, $SD = 13.16$). After blocks of 100 trials, participants were offered a short break (30 s).

Group evaluation and RCIC strategy use. Following the RCIC paradigm, we measured explicit evaluations of both Group X and Group Y on separate 7-point rating scales (ranging from -3 : *very negative* via 0 : *neutral* to $+3$: *very positive*). This manipulation check was intended to assess whether the behavioral descriptions induced differently valenced general perceptions of the two fictitious groups. This would be particularly important in the absence of a main effect of condition on ICI appearance (e.g., no threatening appearance in the threat condition), ensuring the strength of the manipulation. Furthermore, we asked participants to report their RCIC selection strategies (see Supplementary Information S2).

Self-report measures

Paranoia. We measured paranoia using the 10-item self-report persecution subscale of the R-GPTS⁴⁹. The R-GPTS assesses to what extent participants have experienced paranoid thoughts during the last month (e.g., "People wanted me to feel threatened, so they stared at me") on 5-point scales (ranging from 0 : *not at all* to 4 : *totally*). The R-GPTS has shown good psychometric properties^{49,54,55} and achieved satisfactory internal consistencies within the present samples (Cronbach's $\alpha = 0.93$ and 0.89 for the first and second stage, respectively).

Negative affective states. We asked participants to report current feelings of happiness, sadness, anger, shame, and guilt on 5-point scales (ranging from 0 : *not at all* to 4 : *very*; based on⁵⁶) before prior activation to control for negative affective states.

Negative beliefs about others. We measured beliefs about the self and others with the Brief Core Schema Scales (BCSS⁵⁷). The BCSS is a 24-item self-report instrument with four 6-item subscales assessing positive and negative beliefs about the self (e.g., "I am talented" vs. "I am unloved") and others (e.g., "Others are fair" vs. "Others are hostile") on 5-point scales (ranging from 0 : *no, don't believe it* to 4 : *yes, believe it totally*). Only the Negative Others subscale (BCSS-NO) was included in the present analyses. The BCSS demonstrated good psychometric properties⁵⁷.

Social adversity. We measured participants' social adversity experiences with four items assessing the prevalence of emotional, psychological, physical, and sexual abuse prior to their 18th birthday rated on 6-point scales (ranging from 0 : *never* to 5 : *very often*; based on⁵⁸).

General psychopathology. We administered a 14-item short version of the Symptom-Checklist (SCL-GSI⁵⁹) to control for participants' general psychopathology (i.e., the severity of phobic,

Table 1. Sample characteristics.

	First stage			Second stage
	Low paranoia ($n = 111$)	High paranoia ($n = 109$)	Test statistic	Raters ($n = 76$)
Age, M (SD)	32.76 (14.26)	26.4 (7.64)	$t(168.94) = 4.13, p < 0.001^b$	25.45 (6.61)
Gender (female/male/diverse)	65/46/0	52/53/4	$\chi^2(1) = 1.43, p = 0.232^c$	55/21/0
Education (low/medium/high)	1/9/100 ^a	5/21/83	$\chi^2(2) = 9.04, p = 0.011$	0/2/74
R-GPTS, M (SD)	0.88 (1.28)	17.21 (5.63)	-	4.88 (6.29)
Number of blocks rated (one/two/three)	-	-	-	58/5/13

R-GPTS Revised Green et al. Paranoid Thoughts scale, persecution scale.
^a $n = 1$ missing due to technical error.
^bWelch two sample t -test.
^cDiverse gender was omitted from this test due to a limited number of cases.

depressive, and somatic symptoms) during the last seven days on 5-point scales (ranging from 0: *not at all* to 4: *very strong*).

ICI creation and ICI rating. We created one ICI per participant by averaging all noise patterns selected during RCIC and by superimposing this average onto the morphed base face, using the `rcicr` package⁵² for R (version 4.1.0⁵³). These ICIs can be interpreted as visual proxies of participants' mental representations of a typical Group X member and reflect the extent to which the behavioral descriptions informed their facial prior beliefs. In the second stage of the study, an external sample rated the ICIs on threat and trustworthiness in random order using separate 7-point scales (ranging from 1: *not at all* to 7: *very*). Due to the large number of stimuli, ICIs were presented in three approximately equal-sized blocks (i.e., two blocks with 73 and one block with 74 ICIs) and raters could decide how many blocks they wished to rate. ICI order and trait rating order were randomized. Each ICI was rated by 26 participants who were blind to all procedures related to the generation of the ICIs.

Statistical analysis

We combined participants' separate explicit group evaluations into a difference score (i.e., positive values = Group X was rated more positively than Group Y; negative values = Group X was rated more negatively than Group Y) and applied non-parametric Wilcoxon rank-sum tests to account for the non-normality of this metric. Due to the high inter-rater reliability of the ICI ratings ($ICC(1,k) = 0.97$ for threat and trustworthiness ratings), we averaged them across raters to obtain one mean threat and one mean trustworthiness value per ICI. Because these values were highly correlated ($r > 0.80$), we subtracted the mean trustworthiness from the mean threat ratings (i.e., ICI threat score; positive score = ICI was rated as more threatening than trustworthy; negative score = ICI was rated as more trustworthy than threatening). We excluded four participants from the analyses because they did not comply with the RCIC instructions ($n = 3$) or their ICI was not rated due to a technical error ($n = 1$; see Supplementary Information S3). ICI threat scores were submitted to a 2 (Paranoia: low vs. high) \times 2 (Condition: trust vs. threat) between-subjects analysis of variance (ANOVA) to test for main and interaction effects (H1 & H2). In a second step, we added BCSS-NO and social adversity experiences as well as their first- and second-order interactions with the between-subjects factors as covariates (ANCOVA) to explore whether ICI appearance covaried with these variables (H3 & H4). We complemented the frequentist analyses with Bayesian analysis counterparts performed with JASP⁶⁰ according to guidelines^{61,62}, and report Bayes Factors (BF) along with the p -values. BF hypothesis testing directly and continuously compares two competing statistical models, with BF_{10} quantifying the amount of evidence for the alternative over the null hypothesis and BF_{incl} quantifying the amount of evidence

for including a predictor in a model (e.g., ANOVA) over excluding it. A widely accepted rule of thumb⁶³ distinguishes 'anecdotal' ($1 < BF < 3$) from 'moderate' ($3 < BF < 10$) and 'strong' evidence ($BF > 10$; see Supplementary Information S4 for a more details).

RESULTS

Descriptive statistics and group classification images

Socio-demographic characteristics are presented in Table 1. Participants with high versus low paranoia differed significantly with respect to mean age and educational level. Descriptive statistics of ICI threat scores and self-report measures are shown in Table 2. Participants with high versus low paranoia differed significantly with respect to all self-report measures (see Supplementary Information S5). For a visualization of averaged classification images, see Fig. 2.

Manipulation check

As expected, there was strong evidence that participants in the threat condition explicitly evaluated Group X significantly more negatively than Group Y ($M = -3.77, SD = 2.03$), while the opposite was true for participants in the trust condition ($M = 3.60, SD = 2.34; W = 411.00, p < 0.001, BF_{10} = 3.48 \times 10^8$). Moreover, explicit evaluations did not differ across paranoia levels (HP: $M = -0.14, SD = 4.48$; LP: $M = -0.04, SD = 4.13; W = 5951.00, p = 0.834, BF_{10} = 0.15$), indicating that the prior activation was equally effective across samples.

ICI ratings

The two-way ANOVA revealed a significant main effect of condition on ICI ratings, indicating a higher ICI threat score in the threat condition ($M = 1.28, SD = 1.42$) relative to the trust condition ($M = -0.80, SD = 1.36; F(1, 216) = 123.05, p < 0.001, \eta_p^2 = 0.36, 95\%CI [0.26, 0.46], BF_{incl} = 2.44 \times 10^{14}$; see Fig. 3A). Average ratings of the ICIs generated by participants in the threat condition were significantly above zero (i.e., threatening appearance, $t(109) = 9.44, p < .001, d = 0.90, 95\%CI [0.68, 1.12]$), whereas average ratings of the ICIs generated by participants in the trust condition were significantly below zero (i.e., trustworthy appearance, $t(109) = -6.12, p < .001, d = -0.58, 95\%CI [-0.78, -0.38]$). However, we found no evidence for a main effect of paranoia level ($F(1, 216) = 3.18, p = 0.076, \eta_p^2 = 0.01, 95\%CI [0.00, 0.06], BF_{incl} = 0.56$) or an interaction effect between both factors ($F(1, 216) = 0.23, p = 0.632, \eta_p^2 = 0.00, 95\%CI [0.00, 0.03], BF_{incl} = 0.92$).

Adding the BCSS-NO and social adversity scores as well as all first- and second-order interactions as covariates to the model did not affect the main and interaction effects. However, the ANCOVA revealed a significant three-way interaction of condition, paranoia, and BCSS-NO ($F(1,202) = 10.66, p = 0.001, \eta_p^2 = 0.05, 95\%CI [0.01, 0.12]$; see Fig. 3B). Specifically, BCSS-NO and ICI threat scores

Table 2. Means and standard deviations per paranoia level and condition.

Variable	α	Low paranoia				High paranoia			
		Trust ($n = 57$)		Threat ($n = 54$)		Trust ($n = 53$)		Threat ($n = 56$)	
		<i>M</i>	(<i>SD</i>)	<i>M</i>	(<i>SD</i>)	<i>M</i>	(<i>SD</i>)	<i>M</i>	(<i>SD</i>)
ICI threat score ^a	-	-0.68	(1.41)	1.50	(1.52)	-0.92	(1.32)	1.07	(1.30)
Explicit group evaluation ^b	-	3.33	(2.32)	-3.59	(2.12)	3.89	(2.34)	-3.95	(1.94)
BCSS-NO	0.87	4.81	(3.75)	4.65	(3.15)	10.00	(4.66)	9.25	(4.28)
SAE	0.76	2.93	(3.35)	2.94	(3.19)	5.40	(4.52)	5.68	(4.06)
State negative affect	0.80	0.33	(0.37)	0.37	(0.43)	0.98	(0.85)	0.71	(0.70)
SCL-GSI	0.91	7.75	(7.50)	7.31	(5.99)	17.91	(10.79)	14.27	(10.70)

ICI Individual classification image, BCSS-NO Brief Core Schema Scale-Negative Others, SAE Social Adversity Experiences, SCL-GSI Symptom Checklist (general severity index), α Cronbach's α .

^aDifference score (mean threat rating – mean trustworthiness rating).

^bDifference score (explicit Group X evaluation – explicit Group Y evaluation).



Fig. 2 Averaged classification images per condition, paranoia level, and condition \times paranoia combinations. Due to an increased type I error rate associated with averaged CI ratings⁶⁸, we restricted our inferential analyses to the ICI ratings. The morphed base face is presented for reference. HP high paranoia, LP low paranoia.

correlated positively only for HP participants in the threat condition ($r = 0.32$, 95%CI [0.07, 0.54], $t(54) = 2.51$, $p = 0.015$). All results remained the same when controlling for participant age, education, state negative affect, general psychopathology, completion duration, and a metric quantifying the signal in each ICI⁶⁴ (see Supplementary Information S6, S7).

Exploratory Analyses

We repeated the main analyses including only those HP participants with R-GPTS values ≥ 18 ($n = 44$) to examine whether the expected main and interaction effects would emerge only in participants with at least severe paranoia⁴⁹. However, this was not the case (see Supplementary Information S8).

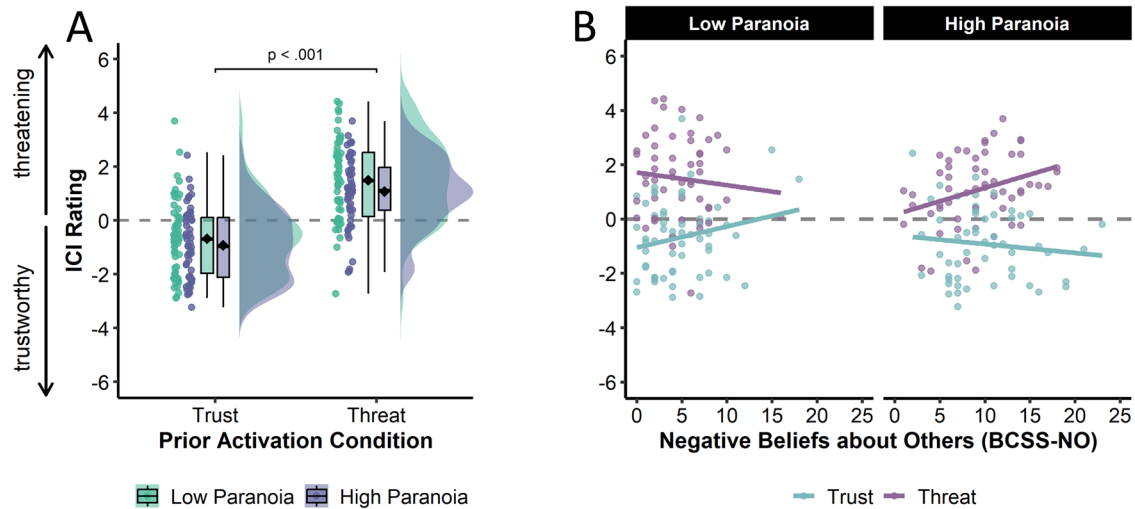


Fig. 3 Individual composite image ratings (ICI threat scores). **A** Raincloud plots by condition and paranoia level. Center lines represent medians, diamonds represent mean values, boxes represent the first and third quartiles, and whiskers represent $1.5 \times$ interquartile range. **B** Scatter plots of ICI threat scores as a function of negative beliefs about others. Fitted regression lines are displayed over the data.

DISCUSSION

We investigated whether the impact of threat prior beliefs on face perception is stronger in individuals with high versus low paranoia, and whether this effect generalizes to trustworthiness prior beliefs. As expected, we observed a substantial main effect of prior activation condition on ICI appearance, such that ICIs were rated as more threatening in the threat condition and more trustworthy in the trustworthiness condition. However, we neither observed a main effect of paranoia nor an interaction between paranoia level and condition. A more nuanced analysis revealed a significant three-way interaction, indicating that in the threat condition, ICIs generated by participants with high paranoia were rated the more threatening the more strongly they held negative beliefs about others. Social adversity experiences did not affect ICI ratings in either condition.

Our findings conceptually replicate previous evidence documenting a causal effect of behavioral descriptions on ICI appearance⁴⁷. This effect is reflected in the averaged classification images, where the threat condition corresponded to an angry appearance and the trustworthiness condition was associated with a happy appearance, aligning with the established link between emotional valence and social attributions⁴⁸. It is essential to keep in mind that participants were presented with random variations of a group-ambiguous base face and asked to select the faces they believed to best represent a *Group X member*, without explicit assessment of threatening or trustworthy appearances. Consequently, participants could have relied on the face stimuli presented along with the behavioral descriptions during prior activation to inform their mental representations of the groups and, ultimately, stimulus selection. However, our results are consistent with the idea that individuals drew spontaneous trait inferences from the behavioral descriptions (a robust phenomenon; see⁶⁵ for a meta-analysis) and used their mental representations associated with these traits to inform their prior beliefs, guiding subsequent stimulus selection. This aligns with both previous research documenting the effect of several top-down biases (e.g., gender, ethnicity, personality traits) on face perception⁴⁶ and with the fundamental principles of predictive processing, positing that prior beliefs shape the perception of ambiguous sensory inputs^{16,46}.

Building on predictive processing accounts of psychosis, which propose an imbalanced integration during perceptual inference as a candidate mechanism underlying delusions^{22–24}, we investigated the impact of threat versus trustworthiness prior beliefs on

face perception in individuals with high versus low paranoia. As such, we expanded on previous signal detection research utilizing non-social sensory inputs (e.g., rotating dot spheres), acknowledging the predominantly social nature of delusions in general and the specific threat-related social valence in paranoia. Contrary to expectations, ICIs generated by participants with high paranoia were not rated as more threatening or less trustworthy than those generated by participants with low paranoia. Thus, our findings do not support the notion that paranoia relates to overly strong threat prior beliefs that shape the perception of ambiguous facial inputs. A plausible interpretation of this null result could be that paranoid beliefs typically transcend observable behaviors, such as facial expressions, and instead involve the assumption of hidden harmful intentions. Consequently, the threat prior beliefs relevant to paranoia might operate on a higher cognitive level rather than directly impacting low-level perceptual processing^{23,24}. These higher-level threat prior beliefs could hinder neutral or even trustworthy facial appearances from being interpreted as evidence against harmful intentions (e.g., ‘Others are dangerous, regardless of their facial appearance’ or even ‘I know they have it in for me because they smile, luring me into believing that everything is fine’). In this case, threat prior beliefs might be more evident in the overall evaluation of others rather than in the expectation of threatening facial appearances. Importantly, this alternative explanation remains untested in the present study, offering an intriguing avenue for exploration in future research.

In light of the absence of a paranoia main effect, our findings challenge the notion that delusional beliefs are rooted in a *domain-general* aberrant perceptual inference process. Recent extension to predictive processing accounts of delusion formation suggested that the predominantly social nature might be accommodated through exposure to early social stressors and adversity⁶⁶. Consistent with this idea, ICIs generated by individuals with high paranoia in the threat condition were rated the more threatening the more strongly these individuals held generalized negative beliefs about others. Thus, negative beliefs about others formed throughout life may sensitize individuals with high paranoia to potential threats, intensifying their reliance on threat information and ultimately sculpting their percept into conformity with these threat prior beliefs. While this interpretation is in line with the association between social adversity and psychotic experiences^{11,12} via learned negative beliefs about others¹³, it should be taken with a grain of salt given both the small effect size and the fact that there was no effect of social adversity

experiences on ICI threat ratings. Nevertheless, our findings underscore the importance of considering potential influencing factors in investigations of aberrant perceptual inference processes in future research, contributing to a nuanced understanding of the mechanisms underlying delusions.

An alternative explanation relates to the paranoia severity reported by our high paranoia sample, which may not have been elevated enough to observe *overly strong* threat prior beliefs. However, we included only participants who reported at least *moderately severe* paranoia during the last month – a severity that was found optimal to discriminate patients with paranoid delusions from a non-clinical group⁴⁹. Moreover, an exploratory analysis including only participants with *severe* levels of paranoia did not support this interpretation. Therefore, we are confident that our results are not attributable to a lack of paranoia severity in the present sample.

To our knowledge, this study is the first to investigate whether paranoia is linked to strong threat prior beliefs shaping the perception of ambiguous social sensory inputs. We leveraged a well-established social-psychological paradigm to address the challenge of accessing individuals' prior beliefs while accounting for the social valence inherent in most delusional beliefs. It is noteworthy that participants in this paradigm are typically unaware of the criteria guiding their face selections⁴⁶. The main contribution of this study thus lies in paving the way for investigations into *socially meaningful and potentially incommunicable* prior beliefs, which we believe is crucial when aiming to uncover the mechanisms underlying delusions. Nonetheless, several limitations merit consideration. Firstly, we compared two behavioral description conditions with opposite valences. Therefore, it remains unclear if our results reflect specific traits (threat vs. trustworthiness) or mere valence (negative vs. positive). Importantly, the proposed mechanisms and their interpretations pertain to both scenarios. Secondly, the ICIs represent an approximation of participants' mental representations, constrained by stimulus materials and performance⁴⁶ as well as the traits rated by the external sample. Future research could employ novel reverse correlation techniques⁶⁷ to test whether our results replicate with photorealistic portraits, potentially incorporating additional personality dimensions of interest. Thirdly, while prior activation may have diminished during RCIC, controlling for individual differences in completion duration revealed no influence of timing on the results. Fourthly, the online setting may have affected performance, although controlling for negative affective states in the beginning of the study did not alter the results. Future studies could repeat the experiment in a more controlled setting such as a laboratory. Lastly, paranoia level was a quasi-experimental factor and the samples significantly differed in age and educational level, potentially limiting the generalizability of the results; however, our results remained robust to controlling for these socio-demographic variables.

In conclusion, our findings suggest that behavioral descriptions inform individuals' facial prior beliefs, shaping the subsequent perception of others' faces. Contrary to expectations, our study does not support the idea of a generally stronger impact of threat prior beliefs on face perception in individuals with high paranoia compared to low paranoia. This challenges the assumption that paranoid beliefs operate on a perceptual level. Future research should further investigate the nuanced interplay between threat prior beliefs at different hierarchical levels.

DATA AVAILABILITY

The data that support the findings of this study are available at the OSF (<https://osf.io/385sy>).

CODE AVAILABILITY

The R code that reproduces the present analyses is publicly available at the OSF (<https://osf.io/385sy>).

Received: 15 November 2023; Accepted: 12 March 2024;
Published online: 20 March 2024

REFERENCES

- American Psychiatric Association. *Diagnostic and statistical manual of mental disorders*. <https://doi.org/10.1176/appi.books.9780890425596> (2013).
- Collin, S., Rowse, G., Martinez, A. P. & Bentall, R. P. Delusions and the dilemmas of life: a systematic review and meta-analyses of the global literature on the prevalence of delusional themes in clinical groups. *Clin. Psychol. Rev.* **104**, 102303 (2023).
- Freeman, D. Suspicious minds: the psychology of persecutory delusions. *Clin. Psychol. Rev.* **27**, 425–457 (2007).
- Bell, V., Raihani, N. & Wilkinson, S. Derationalizing delusions. *Clin. Psychol. Sci.* **9**, 24–37 (2021).
- Freeman, D. et al. Psychological investigation of the structure of paranoia in a non-clinical population. *Br. J. Psychiatry* **186**, 427–435 (2005).
- Bebbington, P. E. et al. The structure of paranoia in the general population. *Br. J. Psychiatry* **202**, 419–427 (2013).
- Elahi, A., Perez Algorta, G., Varese, F., McIntyre, J. C. & Bentall, R. P. Do paranoid delusions exist on a continuum with subclinical paranoia? A multi-method taxometric study. *Schizophr. Res.* **190**, 77–81 (2017).
- Bentall, R. P., Corcoran, R., Howard, R., Blackwood, N. & Kindermann, P. Persecutory delusions: a review and theoretical integration. *Clin. Psychol. Rev.* **21**, 1143–1192 (2001).
- Freeman, D. Persecutory delusions: a cognitive perspective on understanding and treatment. *The Lancet Psychiatry* **3**, 685–692 (2016).
- Broyd, A., Balzan, R. P., Woodward, T. S. & Allen, P. Dopamine, cognitive biases and assessment of certainty: A neurocognitive model of delusions. *Clin. Psychol. Rev.* **54**, 96–106 (2017).
- Varese, F. et al. Childhood adversities increase the risk of psychosis: A meta-analysis of patient-control, prospective-and cross-sectional cohort studies. *Schizophr. Bull.* **38**, 661–671 (2012).
- Bentall, R. P. et al. From adversity to psychosis: pathways and mechanisms from specific adversities to specific symptoms. *Soc. Psychiatry Psychiatr. Epidemiol.* **49**, 1011–1022 (2014).
- Alameda, L. et al. A systematic review on mediators between adversity and psychosis: Potential targets for treatment. *Psychol. Med.* **50**, 1966–1976 (2020).
- Rao, R. P. N. & Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2**, 79–87 (1999).
- Knill, D. C. & Pouget, A. The Bayesian brain: the role of uncertainty in neural coding and computation. *Opin. TRENDS Neurosci.* **27**, 712–719 (2004).
- Friston, K. A theory of cortical responses. *Philos. Trans. R. Soc. B Biol. Sci.* **360**, 815–836 (2005).
- Friston, K. Hierarchical models in the brain. *PLoS Comput. Biol.* **4**, e1000211 (2008).
- Yon, D. & Frith, C. D. Precision and the Bayesian brain. *Curr. Biol.* **31**, R1026–R1032 (2021).
- Zhou, L. F. & Meng, M. Do you see the “face”? Individual differences in face pareidolia. *J. Pacific Rim Psychol.* **14**, e2 (2020).
- Salge, J. H., Pollmann, S. & Reeder, R. R. Anomalous visual experience is linked to perceptual uncertainty and visual imagery vividness. *Psychol. Res.* **85**, 1848–1865 (2021).
- Corlett, P. R., Taylor, J. R., Wang, X.-J., Fletcher, P. C. & Krystal, J. H. Toward a neurobiology of delusions. *Prog. Neurobiol.* **92**, 345–369 (2010).
- Adams, R. A., Stephan, K. E., Brown, H. R., Frith, C. D. & Friston, K. J. The computational anatomy of psychosis. *Front. Psychiatry* **4**, 47 (2013).
- Sterzer, P. et al. The predictive coding account of psychosis. *Biol. Psychiatry* **84**, 634–643 (2018).
- Diaconescu, A. O., Hauke, D. J. & Borgwardt, S. Models of persecutory delusions: a mechanistic insight into the early stages of psychosis. *Mol. Psychiatry* **24**, 1258–1267 (2019).
- Premkumar, P. et al. Misattribution bias of threat-related facial expressions is related to a longer duration of illness and poor executive function in schizophrenia and schizoaffective disorder. *Eur. Psychiatry* **23**, 14–19 (2008).
- Pinkham, A. E., Brensinger, C., Kohler, C., Gur, R. E. & Gur, R. C. Actively paranoid patients with schizophrenia over attribute anger to neutral faces. *Schizophr. Res.* **125**, 174–178 (2011).
- Mitrovic, M., Ristic, M., Dimitrijevic, B. & Pesic, M. H. Facial emotion recognition and persecutory ideation in paranoid schizophrenia. *Ment. Phys. Heal. Psychol. Reports* **123**, 1099–1116 (2020).
- Pinkham, A. E., Hopfinger, J. B., Pelphrey, K. A., Piven, J. & Penn, D. L. Neural bases for impaired social cognition in schizophrenia and autism spectrum disorders. *Schizophr. Res.* **99**, 164–175 (2008).

29. Pinkham, A. E., Harvey, P. D. & Penn, D. L. Paranoid individuals with schizophrenia show greater social cognitive bias and worse social functioning than non-paranoid individuals with schizophrenia. *Schizophr. Res. Cogn.* **3**, 33–38 (2016).
30. Kirk, H., Gilmour, A., Dudley, R. & Riby, D. M. Paranoid ideation and assessments of trust. *J. Exp. Psychopathol.* **4**, 360–367 (2013).
31. Hillmann, T. E., Ascone, L., Kempkensteffen, J. & Lincoln, T. M. Scanning to conclusions? Visual attention to neutral faces under stress in individuals with and without subclinical paranoia. *J. Behav. Ther. Exp. Psychiatry* **56**, 137–143 (2017).
32. McIntosh, L. G. & Park, S. Social trait judgment and affect recognition from static faces and video vignettes in schizophrenia. *Schizophr. Res.* **158**, 170–175 (2014).
33. Haut, K. M. & MacDonald III, A. W. Persecutory delusions and the perception of trustworthiness in unfamiliar faces in schizophrenia. <https://doi.org/10.1016/j.psychres.2010.04.015> (2010).
34. Schmack, K. et al. Delusions and the role of beliefs in perceptual inference. *J. Neurosci.* **33**, 13701–13712 (2013).
35. Schmack, K., Schnack, A., Priller, J. & Sterzer, P. Perceptual instability in schizophrenia: probing predictive coding accounts of delusions with ambiguous stimuli. *Schizophr. Res. Cogn.* **2**, 72–77 (2015).
36. Schmack, K., Rothkirch, M., Priller, J. & Sterzer, P. Enhanced predictive signalling in schizophrenia. *Hum. Brain Mapp.* **38**, 1767–1779 (2017).
37. Weilhhammer, V. et al. Psychotic experiences in schizophrenia and sensitivity to sensory evidence. *Schizophr. Bull.* **46**, 927–936 (2020).
38. Stuke, H., Weilhhammer, V. A., Sterzer, P. & Schmack, K. Delusion proneness is linked to a reduced usage of prior beliefs in perceptual decisions. *Schizophr. Bull.* **45**, 80–86 (2018).
39. Bansal, S. et al. Association Between Failures in Perceptual Updating and the Severity of Psychosis in Schizophrenia. *JAMA Psychiatry* **79**, 169–177 (2022).
40. Davies, D. J., Teufel, C. & Fletcher, P. C. Anomalous perceptions and beliefs are associated with shifts toward different types of prior knowledge in perceptual inference. *Schizophr. Bull.* **44**, 1245–1253 (2018).
41. Teufel, C. et al. Shift toward prior knowledge confers a perceptual advantage in early psychosis and psychosis-prone healthy individuals. *Proc. Natl. Acad. Sci. USA* **112**, 13401–13406 (2015).
42. Stuke, H., Kress, E., Weilhhammer, V. A., Sterzer, P. & Schmack, K. Overly strong priors for socially meaningful visual signals are linked to psychosis proneness in healthy individuals. *Front. Psychol* **12**, 583637 (2021).
43. Dotsch, R., Wigboldus, D. H. J., Langner, O. & Van Knippenberg, A. Ethnic out-group faces are biased in the prejudiced mind. *Psychol. Sci.* **19**, 978–980 (2008).
44. Mangini, M. C. & Biederman, I. Making the ineffable explicit: estimating the information employed for face classifications. *Cogn. Sci.* **28**, 209–226 (2004).
45. Jack, R. E. & Schyns, P. G. Toward a social psychophysics of face communication. *Annu. Rev. Psychol.* **68**, 269–297 (2017).
46. Brinkman, L., Todorov, A. & Dotsch, R. Visualising mental representations: a primer on noise-based reverse correlation in social psychology. *Eur. Rev. Soc. Psychol.* **28**, 333–361 (2017).
47. Dotsch, R., Wigboldus, D. H. J. & Van Knippenberg, A. Behavioral information biases the expected facial appearance of members of novel groups. *Eur. J. Soc. Psychol.* **43**, 116–125 (2013).
48. Todorov, A., Olivola, C. Y., Dotsch, R. & Mende-Siedlecki, P. Social attributions from faces: determinants, consequences, accuracy, and functional significance. *Annu. Rev. Psychol.* **66**, 519–545 (2015).
49. Freeman, D. et al. The revised Green et al., Paranoid thoughts scale (R-GPTS): psychometric properties, severity ranges, and clinical cut-offs. *Psychol. Med.* **51**, 244–253 (2021).
50. Green, C. E. L. et al. Measuring ideas of persecution and social reference: the Green et al. paranoid thought scales (GPTS). *Psychol. Med.* **38**, 101–111 (2008).
51. Faul, F., Erdfelder, E., Buchner, A. & Lang, A.-G. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. *Behav. Res. Methods* **41**, 1149–1160 (2009).
52. Dotsch, R. rcicr: Reverse correlation image classification toolbox. R package version 0.4.1. <https://rdrr.io/cran/rcicr/> (2017).
53. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/> (2021).
54. Statham, V., Emerson, L. M. & Rowse, G. A systematic review of self-report measures of paranoia. *Psychol. Assess.* **31**, 139–158 (2019).
55. Williams, T. F. et al. The reliability and validity of the revised Green et al. paranoid thoughts scale in individuals at clinical high-risk for psychosis. *Acta Psychiatr. Scand.* **147**, 623–633 (2023).
56. Stemmler, G., Heldmann, M., Pauls, C. A. & Scherer, T. Constraints for emotion specificity in fear and anger: the context counts. *Psychophysiology* **38**, 275–291 (2001).
57. Fowler, D. et al. The brief core schema scales (BCSS): psychometric properties and associations with paranoia and grandiosity in non-clinical and psychosis samples. *Psychol. Med.* **36**, 749–759 (2006).
58. Jaya, E. S., Ascone, L. & Lincoln, T. M. Social adversity and psychosis: the mediating role of cognitive vulnerability. *Schizophr. Bull.* **43**, 557–565 (2017).
59. Prinz, U. et al. Die symptom-checkliste-90-R und ihre kurzversionen: psychometrische analysen bei patienten mit psychischen erkrankungen. *Phys. Medizin Rehabil. Kurortmedizin* **18**, 337–343 (2008).
60. JASP Team. JASP (Version 0.16.1) [Computer Software]. <https://jasp-stats.org/> (2022).
61. van Doorn, J. et al. The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychon. Bull. Rev.* **28**, 813–826 (2021).
62. Van Den Bergh, D. et al. A tutorial on conducting and interpreting a bayesian ANOVA in JASP. *Annee Psychol* **120**, 73–96 (2020).
63. Wagenmakers, E.-J. et al. Bayesian inference for psychology. Part II: Example applications with JASP. *Psychon Bull Rev* **25**, 58–76 (2018).
64. Brinkman, L. et al. Quantifying the informational value of classification images. *Behav. Res. Methods* **51**, 2059–2073 (2019).
65. Bott, A. et al. Spontaneous Trait Inferences From Behavior: A Systematic Meta-Analysis. *Personal. Soc. Psychol. Bull.* **50**, 78–102 (2024).
66. Williams, D. & Montagnese, M. Bayesian psychiatry and the social focus of delusions. in *Expected Experiences* (eds. Cheng, T., Sato, R. & Hohwy, J.) 257–282 (Routledge, New York, 2023).
67. Albohn, D. N., Uddenberg, S. & Todorov, A. A data-driven, hyper-realistic method for visualizing individual mental representations of faces. *Front. Psychol.* **13**, 997498 (2022).
68. Cone, J., Brown-Iannuzzi, J. L., Lei, R. & Dotsch, R. Type I error is inflated in the two-phase reverse correlation procedure. *Soc. Psychol. Personal. Sci.* **12**, 760–768 (2021).
69. Ma, D. S., Correll, J. & Wittenbrink, B. The Chicago face database: A free stimulus set of faces and norming data. *Behav. Res. Methods* **47**, 1122–1135 (2015).

ACKNOWLEDGEMENTS

We acknowledge financial support from the Open Access Publication Fund of Universität Hamburg. The authors gratefully thank Dr. Jürgen Kempkensteffen for his technical support, Linda Branecka for her help with the data collection, as well as Saskia Denecke and Sven N. Schöning for their feedback on a previous draft.

AUTHOR CONTRIBUTIONS

Conceptualization: A.B., T.M.L. Methodology: A.B., H.C.S. Formal Analysis: A.B., J.L.F. Investigation: A.B., H.C.S., T.M.L. Writing – Original Draft: A.B. Writing – Review & Editing: A.B., H.C.S., J.L.F., and T.M.L.

FUNDING

Open Access funding enabled and organized by Projekt DEAL.

COMPETING INTERESTS

The authors declare no competing interests.


ADDITIONAL INFORMATION

Supplementary information The online version contains Supplementary Material available at <https://doi.org/10.1038/s41537-024-00459-z>.

Correspondence and requests for materials should be addressed to Antonia Bott.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024