Current Psychology (2023) 42:29090–29099
https://doi.org/10.1007/s12144-022-04003-z

Interpretation inflexibility and negative social impression in paranoia

Wisteria Deng[1] · **Jutta Joormann**[1] · **Tyrone Cannon**[1]

Accepted: 2 November 2022 / Published online: 12 November 2022

Abstract

Paranoia is associated with difficulties in revising initial interpretations of social situations, a phenomenon that may help explain resistance of paranoid thoughts to modification by experience. But what leads individuals with paranoia to become suspicious of newly encountered others? The present study examined the relationship between paranoia and social impression formation for characters who varied in degree of familiarity and according to emotional valence of prior encounters. The study utilized the Interpretation Inflexibility Task (IIT) to measure social interpretation bias and inflexibility, followed by examining the social impression of the IIT characters and the novel ones. Participants from the general population ($N = 213$) overall responded more favorably to previously seen characters – with or without recognizing them as having appeared before – and to characters from scenarios with positive outcomes. However, these effects were absent in participants with higher paranoia, who rated characters more negatively regardless of prior exposure, recognition, or valence of scenario outcome. This pattern suggests that a pronounced negativity bias among individuals with paranoia likely plays an integral role in initial impression formation.

Keywords Cognitive inflexibility · Belief revisioning · Psychosis · Memory · Bias · Implicit

Paranoia is a form of delusional belief that reflects excessive distrust and suspicion of others (Raihani & Bell, 2017). Like other types of delusions, paranoia is extremely resistant to modification by experience. One potential contributor to this incorrigibility is an elevated bias against disconfirmatory evidence, as demonstrated in tasks involving ambiguous social scenarios that gradually reveal information discrediting initial impressions (Bronstein et al., 2019; Hurley et al., 2018). Among individuals with paranoia, interpretation inflexibility in social scenarios exists above and beyond the effects of initial interpretation bias and regardless of the emotional valence of the scenarios (Deng et al., 2022). Although a generalized difficulty with revising interpretations helps to explain the persistence of paranoid beliefs, it remains unclear how newly-encountered individuals may come to be held with suspicion by those with paranoia. Understanding the mechanisms by which paranoid individuals gauge the trustworthiness of newly-encountered individuals is critical for identifying therapeutic targets that could help in encouraging initial impressions that do not involve excessive distrust or suspicion, thereby improving interpersonal functioning (Freeman et al., 2010; Hooker et al., 2011).

Relevant research has noted that people with paranoia show an exaggerated negativity bias in forming social impressions, especially in judging the trustworthiness of others (Hooker et al., 2011). It is thus possible that paranoid beliefs represent highly potent negative "priors" that dominate over the normative positivity biases characteristic of the general population. One such bias is the increased positive impression of social characters who are familiar to an individual, defined as the "mere exposure effect" (de Zilva et al., 2016; Montoya et al., 2017). Mere exposure effects exist with and without one's conscious awareness (Dijksterhuis & Smith, 2002) – for example, one may react more favorably to a character previously seen, both when they retain an explicit memory of the prior

Public health statement Our findings highlight the importance of modifying negative social interpretations in paranoia by connecting the scenario-based belief inflexibility with a negativity bias in impression formation. The blunted mere exposure effect in individuals with high paranoia accentuates the importance of social skills training, especially in early interventions that aim at improving social functional outcomes.

✉ Wisteria Deng
wisteria.deng@yale.edu

[1] Department of Psychology, Yale University, 2 Hillhouse Avenue, New Haven, CT 06520, USA

encounter and when they fail to recognize the character (Montoya et al., 2017). However, findings on the relationship between mere exposure effects and explicit awareness of prior encounters are mixed – some studies have shown a boosted mere exposure effect when individuals do not recall prior encounters (i.e., increased liking of the previous stimuli without the awareness of experiencing them before); whereas others observed an inhibited mere exposure effect when prior encounters were not remembered (Zilva et al., 2013). Multiple models are theorized to explain this relationship between familiarity and liking. The representation-matching model, for example, proposes that individuals attempt to match novel stimuli with mental representations in memory, to guide their judgment of the new stimuli (Montoya et al., 2017). Considering the mere exposure effect, it is reasonable to predict that at *lower* levels of paranoia, individuals will judge previously seen characters more favorably, with or without accurately recognizing them as having appeared before.

The mere exposure effect has not yet been examined among paranoid individuals in regard to impression formation in real-world contexts, such as the process of interpreting social interactions and forming impressions of others. However, individuals higher in paranoia show reduced memory performance, particularly in relation to explicit and episodic forms of memory (Dorofeikova et al., 2018). Given the dependency of the mere exposure effect in creating positive social evaluation priors on memory processes, it seems likely that those *higher* in paranoia would show less of a positivity bias in evaluating the trustworthiness of previously encountered characters (i.e., less susceptibility to the mere exposure effect).

The present study aimed to determine the relationship between paranoia and social impressions of characters encountered during an interpretation flexibility task. The task paradigm involved initially ambiguous social situations that were gradually resolved toward either positive or negative outcomes. Subjects were asked to rate their impressions of characters drawn from the positive and negative outcome scenarios as well as foils not previously encountered, and to indicate whether they remembered seeing the characters in the initial phase. Overall, we hypothesized that people with lesser degrees of paranoia are more sensitive to prior exposure effects and to valence of scenario outcomes, in the direction that prompts more positive social impressions of previously seen characters and characters from positive outcome scenarios. In contrast, we predicted that individuals with greater degrees of paranoia will show overall less positivity in social impressions and less of an increase in positive social impression of characters previously shown compared with novel characters, irrespective of whether they explicitly remember the characters or the valence of the scenario outcome.

Method

Participants

Power analysis was performed using G*Power (Faul et al., 2007) to determine the ideal sample size that would achieve a power of 80% at α-level of 0.05 (two-sided). Participants ($N = 213$) were recruited via Amazon's Mechanical Turk (MTurk), an online crowdsourcing platform that provides access to a large and diverse sample for mental health research studies (demographics: see Table 1). Participation in this study was restricted to MTurk users who were 18 years or older and lived in the United States.

Data quality

Following recommendations for research using crowdsourced samples (Chandler & Shapiro, 2016), the study only recruited MTurk users who had a history of providing good-quality responses. Participants were required to have completed at least 500 MTurk studies and to have had their work approved (vs. rejected) in 98% of the studies they completed previously on MTurk. In addition, three questions were included to discriminate attentive from inattentive participants. These questions were presented at random intervals and participants were required to answer all three correctly. Consistent with previous research, one participant was excluded from all analyses as they completed the survey

Table 1 Demographic characteristics

Characteristic	*n*
Gender	
Male	111
Female	102
Race	
White	179
Black or African American	22
Asian	12
Latino/Latina	18
Native American or Alaska native	5
Other	3
Education	
Less than high school	3
High school graduate	27
Some college	48
Two-year college graduate	25
Four-year college graduate	89
Master's degree graduate	19
Doctorate degree graduate	2

Mean age = 39.61 years ($SD = 10.41$); age range = 23–74

in less than 60% of the projected time (< 27 min). Research using a similar approach (e.g., requiring a history of good quality response) has demonstrated that MTurk data are comparable to those collected in the laboratory (Chandler & Shapiro, 2016).

Measures

Interpretation inflexibility task

The Interpretation Inflexibility Task (IIT) is a validated, picture-based task assessing individuals' belief inflexibility and interpretation bias in social and emotional scenarios (Deng et al., 2022). The IIT contains 24 scenarios about interpersonal situations, each is gradually revealed to the respondents in three stages: with 80% of the photo blurred, 20% of the photo blurred and the original photo with no blurring effects. The blurred region was selected with the goal of obscuring the emotional valence of a given scenario. By gradually reducing the percentage of the photo that is blurred, the respondent receives more information that may help to resolve the initially ambiguous situation.

Two types of scenarios were included in the IIT to examine whether interpretation inflexibility differed according to the valence of the new evidence in relation to the initial interpretation that was most strongly suggested by ambiguous stimuli. An example positive outcome scenario is shown in Fig. 1 (top). In this example, each of the three pictures that make up the scenario (panels 1a-1c) reveal more information about the scenario, culminating in panel 1c, which encourages a positive interpretation. An example negative outcome scenario of the IIT is also shown in Fig. 1 (panels 2a-2c). Each of the three pictures (panels 2a-2c) reveals more information about the scenario, culminating in panel 2c, which prompts a negative interpretation. For each of the three pictures, four interpretations of the depicted scenario were presented to participants, who were asked to rate the plausibility of each interpretation. Interpretations include 1 Absurd, 2 Lures, and 1 True interpretation for each scenario. For example, the IIT scenario depicted in Fig. 1a and b, and 1c is followed by an absurd interpretation ("People discuss the smell of the field"), two lure interpretations ("People stop you from starting a fight" and "People are making fun of you"), and a true interpretation ("People celebrate what a great player you are"). Participants were instructed to imagine each scenario as if they were an observer in the situation and could see it through their own eyes. The order of the scenario types (positive vs. negative) was randomized to minimize the learning effect.

Social impression formation

Follow-up questions on post-IIT social impression updating were given to the participants 30 min after completion of the IIT. Participants were shown a total of 24 faces, 18 of which were previously shown IIT characters whose faces were

Fig. 1 Example IIT scenarios, with gradual unblurring to achieve the disambiguation of emotional information

cropped from their individual IIT scenarios (see Fig. 2), 9 from each emotional valence (i.e., positive vs. negative outcome scenarios). In order to test the overall impression about the character, instead of any particular memory of the revised beliefs, the IIT characters were chosen from the portion of the picture that had not been blurred in the early stages of each scenario, or did not carry the information that prompted the belief revisioning. To control for baseline levels of positive belief about others, 6 foils were included showing faces of characters participants had never seen before.

For each of the faces, participants were asked to imagine being in a social interaction with the depicted character, and rate on a 7-point Likert scale (7 = very likely) how likely they anticipate (1) receiving a compliment from the character; (2) trusting the character; (3) the character disliking them and (4) the character plotting against them. A total score representing levels of positive social impression was calculated by taking the sum across all four ratings (with items 3 and 4 reverse-coded).

Following the belief ratings, participants were asked to recall if the 24 characters (6 foils plus 18 IIT faces) have appeared in the IIT. Based on the emotional valence of the IIT scenarios that these characters were taken from, and the explicit memory participants had about these characters, two within-subject dimensions of the belief ratings were formed. First, impression ratings for the 9 characters taken from positive outcome scenarios and the 9 from negative outcome scenarios were averaged to form two scores, representing positive and negative scenario outcomes, respectively. Second, impression ratings were averaged across the characters that participants correctly recalled as having appeared in the IIT before (i.e., true positives), and those that participants incorrectly rejected as not in the IIT (i.e., false negatives). Ratings for the 6 foils were averaged to form a "baseline" social impression score regardless of whether participants correctly indicated not having seen those faces in the IIT (i.e., including both true negatives and false positives).

The memory judgements were used to calculate d' as an index for general neurocognitive ability with the formula: d' = z(H) - z(F), where z(H) and z(F) are the z transforms of hit rate (i.e., true positives) and false alarm (i.e., false positives), respectively.

Revised Green et al. Paranoid Thoughts Scale (R-GPTS)

The R-GPTS (Freeman et al., 2021) is a 18-item self-report measure of dispositional persecutory ideation (example item: "People have intended me harm") and self-referential thoughts (example item: "I often heard people referring to me"). Participants rate on a five-point scale (0 = "Not at all," 4 = "Totally") how well each item corresponds to their thoughts and feelings about others over the past month. The R-GPTS is sensitive to clinically significant changes in paranoia and has excellent internal consistency in community samples ($\omega_{total} = 0.96$ in the present study). This measure was included because previous research suggests that persecutory ideation is related to inflexibility of interpretations above and beyond interpretation bias (Bronstein et al., 2019; Deng et al., 2022).

Procedure

All participants gave informed consent in accordance with the Yale University Institutional Review Board. Participants completed a survey which began with demographic questions followed by the IIT, which were presented in randomized order. Participants then completed the R-GPTS and other individual differences measures, before giving ratings based on their impression of social characters. Upon completion of the survey, participants were debriefed and received remuneration (6 USD).

Data analysis

Based on participants' performance during the IIT, the following computational approach was employed to capture (a) initial interpretation bias and (b) within-person revision of prior interpretations. An interpretation bias score was calculated at each of the three stages for the 24 IIT scenarios, by

Fig. 2 Example post-IIT social impression item

taking the additive inverse of the true statement plausibility rating divided by the average of the endorsement ratings for the two lures. In particular, the initial interpretation bias was represented by the bias score at stage 1 (i.e., the first picture of each IIT scenario). Using the interpretation bias scores at each stage, the interpretation flexibility index was calculated by taking the Root Mean Square of Successive Differences (RMSSD) that captures moment-to-moment fluctuations in interpretation bias as a proxy for positive or negative interpretation inflexibility (as derived from scenarios with positive or negative emotional valences). Higher RMSSD values represent high moment-to-moment variability and suggest flexibility in revising an initial (biased) interpretation. By contrast, low RMSSD values represent low moment-to-moment variability and reflect inflexibility in revising an initial interpretation bias based on novel information provided by the IIT scenarios. The following formula was applied to calculate the interpretation flexibility index for each scenario (Deng et al., 2022).

$$Interpretation\ flexibility\ index = \sqrt{\frac{(bias\ score\ stage3 - bias\ score\ stage\ 2)^2 + (bias\ score\ stage\ 2 - bias\ score\ stage\ 1)^2}{2}}$$

Follow-up questions assessed participants' impression of social characters involved in the IIT scenarios. Participants were further asked to determine if they have viewed the characters before. Depending on their answers, ratings on the IIT characters that received a response of "definitely yes" to the question of "Is this face among the 24 scenarios you viewed at the beginning of the study?" were grouped into "true positives"; whereas ratings on the faces that received any other responses ("probably yes", "probably not" and "definitely not") were grouped into "false negatives", representing participants' inability to recall these characters from previous social interactions. A response of "probably yes" was not counted as "true positive" because a more stringent criterion for recognition is needed to better ensure the memory of specific IIT scenario outcomes (an independent variable hypothesized to explain positive impression ratings).

The analytic plan aimed to address two primary objectives: (a) determine the relationship between paranoia and social impressions of characters overall and as a function of whether the characters were in positive or negative outcome scenarios and were remembered or not remembered (or were foils); (b) replicate previous findings on inflexibility in paranoia. First, ratings for characters previously shown in the IIT scenarios were categorized according to memory (faces of which participants report explicit memories vs. those participants do not recall as having seen before vs. foils), as well as by emotional valence (faces taken from positive vs. negative outcome scenarios of the IIT). A mixed effects regression model was built to investigate the main effects of paranoia, memory, and emotional valence, as well as all of interaction effects of these three factors on the impression ratings of the characters (18 of which from the IIT scenarios and 6 foils). Specifically, paranoia severity scores, valence, and incidental recognition memory performance were included as fixed effects, along with d' included to control for the effect of general cognitive ability on impression ratings. Second, to determine whether previously observed relations between interpretation inflexibility and paranoia replicate in this sample, multiple regression models were built testing whether interpretation bias and inflexibility uniquely explained variation in delusion-proneness. For each task and scenario type, regression models were tested separately with paranoia as the dependent variable. In each model, interpretation bias and inflexibility (negative or positive, depending on the scenario outcomes) were simultaneously entered into the regression equation as the independent variables. Assumptions of homoscedasticity and normality of residuals were met for all analyses. Collinearity statistics were within acceptable limits (VIF's < 1.21, Tolerance's > 0.83).

Results

Interpretation inflexibility in paranoia

Paranoia severity is significantly correlated with interpretation bias and inflexibility for both positive and negative outcome scenarios of the IIT (Table 2; $ps < 0.05$). Paranoia severity explains a significant amount of variance in interpretation inflexibility, regardless of the emotional valence of the scenario outcome, above and beyond the effects of interpretation bias (Supplemental Table 2).

Negative social impression in paranoia

Given that no significant three-way interaction was found among paranoia severity, incidental recognition memory performance and valence of IIT scenario outcomes ($p = .678$; Supplemental Table 1), a mixed effects model was built to examine main effects of each independent variable, as well as all of the two-way interactions. The model revealed main effects of paranoia (t = -4.54, $SE = 0.04$, $p < .001$), incidental recognition memory performance (t = -6.90, $SE = 0.03$, $p < .001$), and IIT scenarios' emotional valence (t = -8.74, $SE = 0.48$, $p < .001$) on social impression ratings (Table 3). In general, individuals with more severe paranoia symptoms reported more negative impressions of the characters of the presented faces (Fig. 3a). Overall, participants gave more positive impression ratings to false negatives (i.e.,

Table 2 Correlation matrix between paranoia and inflexibility indices

	2	3	4	5	6	7	8	9	10
1. Paranoia	−0.30***	−0.32***	−0.14*	0.16*	−0.17*	−0.32***	−0.29***	−0.27***	−0.15*
2. Positive IFI	-	0.54***	0.02	−0.15*	0.97	0.37***	0.30***	0.13	0.01
3. Negative IFI		-	−0.12	−0.08	−0.04*	0.22***	0.09	<0.001	−0.02
4. PIB			-	−0.37***	0.21**	0.24***	0.16*	0.26***	0.16*
5. NIB				-	−0.18**	−0.19**	−0.18**	−0.19**	−0.14*
6. Foil ratings					-	0.65***	0.52***	0.69***	0.59***
7. P_Miss						-	0.48***	0.67***	0.48***
8. P_Hit							-	0.60***	0.35***
9. N_Miss								-	0.54***
10. N_Hit									-

* $p<.05$; ** $p<.01$; *** $p<.001$. PIB = positive interpretation bias, NIB = negative interpretation bias, IFI = interpretation flexibility index. P/N_Miss = false negatives (i.e., rating for character from positive/negative outcome scenarios that participants do not recall). P/N_Hit = true positives (i.e., rating for character from positive/negative outcome scenarios that participants recall)

IIT characters that participants did not recognize), as compared to true positives (i.e., IIT characters that participants recognized) and foils (i.e., non-IIT characters) (Fig. 3c). Participants also responded more favorably (giving higher impression ratings) to characters from positive outcome scenarios in the IIT, as compared to those from negative outcomes scenarios or the non-IIT characters (Fig. 3b).

In addition, there was a significant interaction effect between paranoia severity and recognition memory performance on social impression ratings ($t=2.43$, $SE=0.01$, $p=.015$; Fig. 4). For individuals with low paranoia symptoms, prior exposure to the IIT characters was associated with more positive social impressions compared with impressions of unfamiliar characters (i.e., foil ratings), with a tendency for more positive impressions of IIT characters

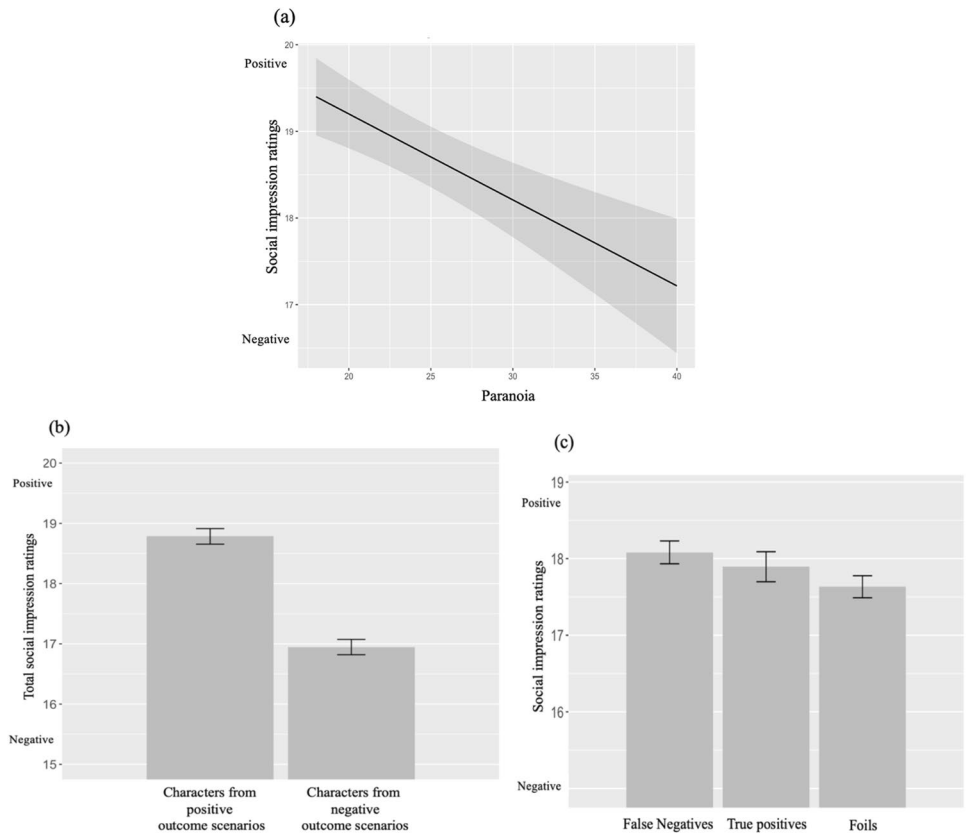

Fig. 3 Main effects of paranoia severity (**a**), valence of the IIT scenario outcomes (**b**), and incidental recognition memory performance (**c**) on positive social impression ratings

Table 3 Mixed effects model with emotional valence (of the IIT scenario outcome), paranoia (winsorized score), memory performance and all the two-way interaction terms predicting positive social impression ratings

	ratings		
Predictors	Estimates	CI	*p*
(Intercept)	27.87	25.87–29.86	**< 0.001**
d'	0.04	-0.22–0.30	0.751
Paranoia	-0.16	-0.23 – -0.09	**< 0.001**
Memory	-2.19	-2.81 – -1.57	**< 0.001**
Valence	-4.20	-5.14 – -3.26	**< 0.001**
Paranoia * Memory	0.02	0.00–0.04	**0.015**
Paranoia * Valence	0.02	-0.01–0.05	0.182
Memory * Valence	0.95	0.66–1.23	**< 0.001**
Random effects			
σ^2	4.39		
$\tau_{00\ subject}$	5.21		
ICC	0.54		
$N_{\ subject}$	213		
Observations	1245		
Marginal R^2 / Conditional R^2	0.138 / 0.606		

Bold indicates $p < .05$

that they did not recognize compared with those that they did (t = -1.80, p = .074). However, these effects of familiarity were blunted among those with high paranoia symptoms, for whom the social impressions of IIT characters were substantially more negative and did not differ between false negatives, true positives, or foil ratings ($ps > 0.05$). Further, there was a significant interaction effect between memory performance and valence of the IIT scenario outcomes ($t = 6.58$, $SE = 0.14$, $p < .001$; Fig. 4) – participants reported more positive ratings of IIT characters taken from the positive outcome scenarios as compared to the ratings of foils, whereas such effect diminished for characters taken from negative outcome scenarios. No significant interaction effect was found between paranoia severity and emotional valence of the IIT scenario outcomes.

The effects just summarized remained significant controlling for memory accuracy (measured by d') and interpretation inflexibility (measured by interpretation flexibility index from IIT), with greater interpretation flexibility in endorsing positive outcomes during the IIT predicting more positive post-IIT social impressions (Table 4).

Discussion

In examining the process of positive impression formation, the study found main effects of paranoia severity, incidental recognition memory performance and emotional valence of prior encounters on positive impressions of social characters. In addition, significant interaction effects were found between paranoia and memory performance, as well as memory and emotional valence. In particular, participants responded more favorably to previously seen characters – with or without accurately recognizing them as having appeared before. However, such mere exposure effects were dampened in participants with more severe paranoid beliefs.

The main effect of paranoia severity on positive impression formation highlights that the association between paranoia and negative impressions of others is global and pervasive, above and beyond the emotional valence of prior encounters and the ability to recognize others as familiar. Specifically, individuals with high paranoia experience difficulty not only in revising their initial interpretation about interpersonal scenarios, but also in making positive judgments about the characters involved, even when prior interactions ended in a positive outcome. In contrast to the

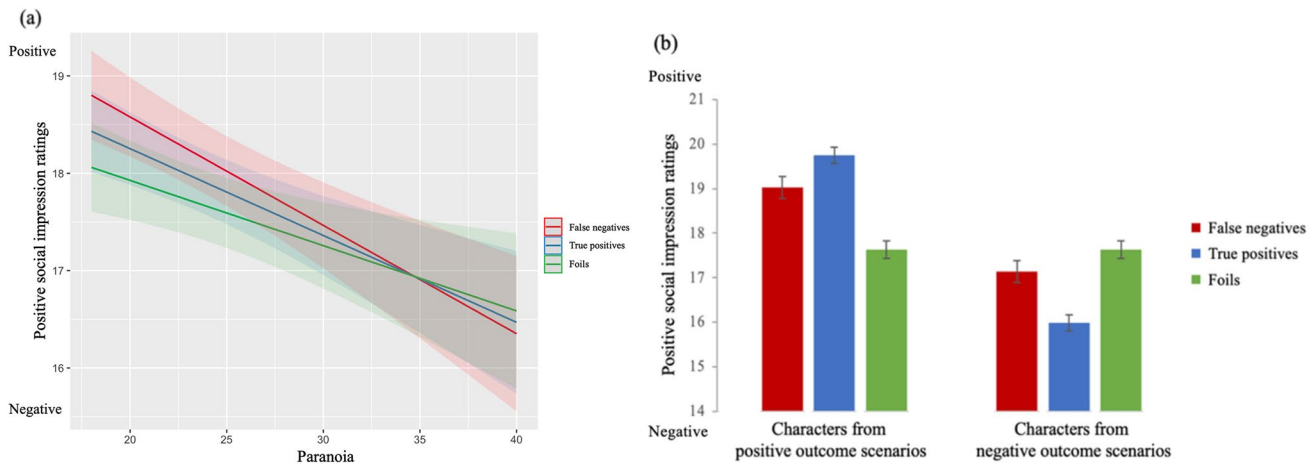

Fig. 4 Two-way interaction effects of paranoia and memory (**a**) and memory and valence (**b**) on positive social impression ratings

Table 4 Regression models with paranoia predicting positive social impression ratings

Dependent variables	Component	*b*	SE_b	*β*	*t*	*p*-value
Positive outcome scenarios						
False negatives	Constant	18.90	0.67		28.35	< 0.001
	Paranoia	− 0.06	0.02	− 0.23	-3.39	< 0.001
	IFI	1.89	0.42	0.30	4.50	< 0.001
	d'	0.03	0.15	0.01	0.22	0.825
True positives	Constant	19.92	0.77		25.99	< 0.001
	Paranoia	− 0.06	0.02	− 0.22	-3.07	0.002
	IFI	1.56	0.48	0.23	3.24	0.001
	d'	0.11	0.18	0.04	0.64	0.522
Negative outcome scenarios						
False negatives	Constant	19.39	0.66		29.52	< 0.001
	Paranoia	− 0.07	0.02	− 0.30	-4.19	< 0.001
	IFI	− 0.46	0.33	− 0.10	-1.38	0.169
	d'	0.05	0.15	0.02	0.31	0.760
True positives	Constant	17.74	0.87		20.32	< 0.001
	Paranoia	− 0.05	0.02	− 0.18	-2.36	0.020
	IFI	− 0.39	0.44	− 0.07	− 0.89	0.374
	d'	− 0.13	0.20	− 0.05	− 0.64	0.520

IFI = interpretation flexibility index

scenario-based interpretation inflexibility associated with paranoia, which is insensitive to the emotional valence of the scenario outcomes, individuals higher in paranoia show a strong negativity bias in forming general impressions about social characters. Importantly, the impression formation questions directly asked about one's expectation of how oneself would be treated in a social interaction with each character, thus capturing more self-referential biases than the scenario-based interpretations. Given that people rely heavily on these social impressions to guide their subsequent behaviors – for example, negative impressions of others may discourage people from future social interactions (Van Kleef, 2010) – the negativity bias in impression formation among individuals with paranoia likely plays an integral role in diminished interests in participating future social activities (i.e., social withdrawal), difficulties fulfilling goal-directed interpersonal tasks (i.e., avolition), and functioning deficits. Modifying negative social impressions could therefore serve as an intervention target to alleviate negative symptoms (such as anhedonia and avolition) that often exist alongside paranoia and contribute to poor functional outcome.

In line with the existing literature on mere exposure effects (Montoya et al., 2017), overall, participants responded more favorably to the IIT characters (as compared to the foils) with or without explicitly recognizing them as familiar. Forming more positive impressions about familiar characters has important implications in intergroup contexts (Crisp et al., 2009), as it helps establish one's identity, foster a sense of belonging and promote in-group cohesion. However, as hypothesized, such mere exposure effects were dampened in individuals with paranoia. Among those with high paranoia, a negativity bias in impression formation persists despite prior encounters with the characters. Considering the blunted response to social interactions, interventions that merely encourage more social involvement (e.g., behavioral activation) may not be as helpful for individuals with paranoia. Instead, social skills training may be necessary for encouraging paranoid individuals to reflect on positive social feedback and notice positive interpersonal experiences that challenge their impressions of others as harboring ill intentions.

In addition to the mere exposure effect, participants overall formed more positive social impressions about characters previously involved in the IIT scenarios that had a positive outcome, compared to those taken from the negative outcome IIT scenarios. The main effect of emotional valence reveals that the general impressions about characters are mostly consistent with the valence of prior encounters, even without one's recognition of these characters as taken from previous scenarios. This adds to the existing literature and demonstrates that not only mere exposure effects exist without conscious recognition of familiarity (Hansen & Wänke, 2009), the emotional context of prior encounters also influences impression formation below the threshold of recognition. Our finding highlights that emotionally charged biases can be passed on from prior social interactions and persist with or without incidental recognition memory. Therefore, it is especially crucial to capture and challenge inflexible negative interpretations of social scenarios, as they exert long-lasting and generalizable impacts on social impression formation beyond scenario-specific beliefs.

The study is not without limitations. The incidental recognition memory assessment took place *after* the impression ratings. While this sequencing minimized potential biases during the impression ratings by presenting the explicit instruction to recall prior encounters later, participants were exposed to the IIT characters twice – during the IIT and when giving the impression ratings – before having to recall prior encounters with these characters. Future studies may consider counterbalancing between presenting questions about impression ratings and the recognition memory task first, or to adopt a between-subject design that minimizes the confounding effects of task sequencing. Further, the current study only examined incidental recognition memory performance with a brief time lapse of 30 min following the completion of the IIT. Future work may evaluate the relation between inflexible biases in impression formation and memory with a longer time interval that allows for greater variance in memory performance and examination of the longer-term persistence of effects of prior exposure on social impressions. It is noteworthy that in the mixed effects model that added the three-way interaction among paranoia, memory and valence, the interaction effect between paranoia severity and memory on positive impression formation was no longer significant. This pattern suggests that some of the variance associated with the paranoia severity by memory interaction is accounted for by the (non-significant) three-way interaction between paranoia, memory, and valence. Nevertheless, it is plausible that, even if at larger sample sizes, the three-way interaction term became significant, it may not further our understanding of the impression formation process among individuals with higher paranoia, given that for these individuals, the impression ratings converge at a low level (i.e., negative impressions) regardless of memory or valence groupings. That said, future work is needed to examine the consistency of the interaction effect between paranoia severity and memory performance in relation to positive impression formation. Finally, it is important to note that our work examined paranoia as presented in the general population. While paranoia is widely regarded as existing on a continuum, ranging from low-grade, non-impairing paranoid beliefs to frank psychosis (Bebbington et al., 2013; Elahi et al., 2017), more work is needed to show whether our findings would generalize to a fully psychotic population.

While this study has started to investigate the relation between emotional valence and memory, more work is needed to expand our knowledge in both aspects of paranoia research. First, specific affective mechanisms (e.g., emotion recognition, emotion regulation) that may influence social interpretation and impression updating remain to be explored. Previous work has noted an impaired emotion recognition ability among individuals with paranoia (Combs et al., 2006; Seo et al., 2020). Research is needed to examine whether the negative social impression ratings in individuals with paranoia is driven by inflexible negative beliefs, or an impaired ability to recognize positive emotions. Second, existing cognitive schemas – beyond the scope of interpretation and impression formation in this study – may play an integral role in individuals' understanding of social situations, as well as their judgment of other characters involved. Research has associated early life adversity and subsequent negative cognitive schema with paranoia (Humphrey et al., 2021). Given that such cognitive schema can be a form of inflexible beliefs that serves as an important intervention target, its relationship with other types of inflexibility (e.g., scenario-based interpretation inflexibility and persistent negative social impressions) deserves further exploration.

In sum, our findings highlight the importance of modifying negative social interpretations in paranoia by connecting the scenario-based belief inflexibility with a negativity bias in impression formation. Such negative impressions about others are reflective of the anticipation of being harmed in social situation, consistent with paranoid beliefs. Further, the blunted mere exposure effect in individuals with high paranoia highlights the importance of social skills training, especially in early interventions that aim at improving social functional outcomes.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12144-022-04003-z.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Bebbington, P. E., McBride, O., Steel, C., Kuipers, E., Radovanovic, M., Brugha, T., Jenkins, R., Meltzer, H. I., & Freeman, D. (2013). The structure of paranoia in the general population. *British Journal of Psychiatry, 202*, 419–427. https://doi.org/10.1192/bjp.bp.112.119032

Bronstein, M. V., Pennycook, G., Joormann, J., Corlett, P. R., & Cannon, T. D. (2019). Dual-process theory, conflict processing, and delusional belief. *Clinical Psychology Review, 72*, 101748. https://doi.org/10.1016/j.cpr.2019.101748

Chandler, J., & Shapiro, D. (2016). Conducting clinical research using crowdsourced convenience samples. *Annual Review of Clinical Psychology, 12*, 53–81. https://doi.org/10.1146/annurev-clinpsy-021815-093623

Combs, D. R., Michael, C. O., & Penn, D. L. (2006). Paranoia and emotion perception across the continuum. *British Journal of Clinical Psychology, 45*(Pt 1), 19–31. https://doi.org/10.1348/014466505X29099

Crisp, R. J., Hutter, R. R., & Young, B. (2009). When mere exposure leads to less liking: the incremental threat effect in intergroup contexts. *British Journal of Psychology, 100*(Pt 1), 133–149. https://doi.org/10.1348/000712608X318635

de Zilva, D., Newell, B. R., & Mitchell, C. J. (2016). Multiple context mere exposure: Examining the limits of liking. *Quarterly Journal*

of Experimental Psychology (Hove), 69(3), 521–534. https://doi.org/10.1080/17470218.2015.1057188

De Zilva, D., Vu, L., Newell, B. R., & Pearson, J. (2013). Exposure is not enough: Suppressing stimuli from awareness can abolish the mere exposure effect. *PLoS One, 8*(10), e77726.

Deng, W., Everaert, J., Creighton, M., Bronstein, M. V., Cannon, T., & Joormann, J. (2022). Developing a novel assessment of interpretation flexibility: Reliability, validity and clinical implications. *Personality and Individual Differences, 190*, 111548.

Dijksterhuis, A., & Smith, P. K. (2002). Affective habituation: subliminal exposure to extreme stimuli decreases their extremity. *Emotion, 2*(3), 203–214. https://doi.org/10.1037/1528-3542.2.3.203

Dorofeikova, M., Neznanov, N., & Petrova, N. (2018). Cognitive deficit in patients with paranoid schizophrenia: Its clinical and laboratory correlates. *Psychiatry Research, 262*, 542–548. https://doi.org/10.1016/j.psychres.2017.09.041

Elahi, A., Perez Algorta, G., Varese, F., McIntyre, J. C., & Bentall, R. P. (2017). Do paranoid delusions exist on a continuum with subclinical paranoia? A multi-method taxometric study. *Schizophrenia Research, 190*, 77–81. https://doi.org/10.1016/j.schres.2017.03.022

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/bf03193146

Freeman, D., Loe, B. S., Kingdon, D., Startup, H., Molodynski, A., Rosebrock, L., Brown, P., Sheaves, B., Waite, F., & Bird, J. C. (2021). The revised Green et al., Paranoid Thoughts Scale (R-GPTS): psychometric properties, severity ranges, and clinical cut-offs. *Psychological Medicine*, *51*(2), 244–253. https://doi.org/10.1017/S0033291719003155

Freeman, D., Pugh, K., Vorontsova, N., Antley, A., & Slater, M. (2010). Testing the continuum of delusional beliefs: an experimental study using virtual reality. *Journal of Abnormal Psychology, 119*(1), 83–92. https://doi.org/10.1037/a0017514

Hansen, J., & Wänke, M. (2009). Liking what's familiar: The importance of unconscious familiarity in the mere-exposure effect. *Social Cognition, 27*(2), 161–182.

Hooker, C. I., Tully, L. M., Verosky, S. C., Fisher, M., Holland, C., & Vinogradov, S. (2011). Can I trust you? Negative affective priming influences social judgments in schizophrenia. *Journal of Abnormal Psychology, 120*(1), 98–107. https://doi.org/10.1037/a0020630

Humphrey, C., Bucci, S., Varese, F., Degnan, A., & Berry, K. (2021). Paranoia and negative schema about the self and others: A systematic review and meta-analysis. *Clinical Psychology Review, 90*, 102081. https://doi.org/10.1016/j.cpr.2021.102081

Hurley, J., Hodgekins, J., Coker, S., & Fowler, D. (2018). Persecutory delusions: effects of Cognitive Bias Modification for Interpretation and the Maudsley Review Training Programme on social anxiety, jumping to conclusions, belief inflexibility and paranoia. *Journal of Behavior Therapy and Experimental Psychiatry, 61*, 14–23. https://doi.org/10.1016/j.jbtep.2018.05.003

Montoya, R. M., Horton, R. S., Vevea, J. L., Citkowicz, M., & Lauber, E. A. (2017). A re-examination of the mere exposure effect: The influence of repeated exposure on recognition, familiarity, and liking. *Psychological Bulletin, 143*(5), 459–498. https://doi.org/10.1037/bul0000085

Raihani, N. J., & Bell, V. (2017). Paranoia and the social representation of others: a large-scale game theory approach. *Scientific Reports, 7*(1), 4544. https://doi.org/10.1038/s41598-017-04805-3

Seo, E., Park, H. Y., Park, K., Koo, S. J., Lee, S. Y., Min, J. E., Lee, E., & An, S. K. (2020). Impaired facial emotion recognition in individuals at ultra-high risk for psychosis and associations with schizotypy and paranoia level. *Frontiers in Psychiatry, 11*, 577. https://doi.org/10.3389/fpsyt.2020.00577

Van Kleef, G. A. (2010). The emerging view of emotion as social information. *Social and Personality Psychology Compass, 4*(5), 331–343.

Data transparency statement None of the results reported in this manuscript have been published in any other journal, and none are under review by any journal. The data have not been used in prior published or in press manuscripts. The treatment of human participants was in accordance with the ethical standards of the American Psychological Association. All authors agree to the authorship order and content of the manuscript. We have no conflicts of interest to disclose.