

Are You Safe or Should I Go? How Perceived Trustworthiness and Probability of a Sexual Transmittable Infection Impact Activation of the Salience Network

 Alexander Wolber,  Stephanie N. L. Schmidt, Brigitte Rockstroh, and  Daniela Mier

Department of Psychology, University of Konstanz, Konstanz 78457, Germany

Abstract

Functional imaging studies indicate that both the assessment of a person as untrustworthy and the assumption that a person has a sexually transmitted infection are associated with activation in regions of the salience network. However, studies are missing that combine these aspects and investigate the perceived trustworthiness of individuals previously assessed with high or low probability of a sexually transmitted infection. During fMRI measurements, 25 participants viewed photographs of people preclassified as having high or low HIV probability and judged their trustworthiness. In a post-rating, stimuli were rated for trustworthiness, attractiveness, and HIV probability. Persons preclassified as HIV⁻ in contrast to those preclassified as HIV⁺ were rated more trustworthy and with lower HIV probability. Activation in medial orbitofrontal cortex was higher for those rated and preclassified as HIV⁻ than HIV⁺. Based on the individual ratings, but not the preclassification, there was significantly higher activation in the insula, amygdala, anterior cingulate cortex, and nucleus accumbens in response to untrustworthy than to trustworthy faces. Activation of the salience network occurred when a person was judged as untrustworthy, but not according to a preclassification. Activation in the medial orbitofrontal cortex, a structure associated with reward, was enhanced when a person was perceived as trustworthy and also when a person was preclassified with low HIV probability. Our findings suggest that trustworthiness and HIV⁻ perception have consistency across samples, while the perception of risk and associated activation of the salience network has restricted cross-sample consistency.

Key words: fMRI; HIV; risk; salience network; trustworthiness

Received June 11, 2024; revised Dec. 4, 2024; accepted Dec. 20, 2024.

The authors declare no competing financial interests.

Author contributions: S.N.L.S., B.R., and D.M. designed research; A.W. performed research; S.N.L.S. contributed unpublished reagents/analytic tools; A.W. and S.N.L.S. analyzed data; A.W., S.N.L.S., B.R., and D.M. wrote the paper.

We thank Alexander Sahm, Sarah Tholl, and Tara Zerbe for their assistance in data collection. This preregistered work was supported by Deutsche Forschungsgemeinschaft (GZ: MI 1975/7-1).

Continued on next page.

Significance Statement

Whether a person is trustworthy or might pose a risk to one's own health must be decided in a few moments and based on limited characteristics. The salience network as an "alarm system" should be involved in these evaluative processes. This paper reports the results of neural activation in trustworthiness judgments of naturalistic stimuli of persons precategorized as HIV⁺ or HIV⁻. We find activation in medial orbitofrontal cortex for people evaluated as trustworthy and for people precategorized as HIV⁻. For people judged as untrustworthy, activation in the insula, amygdala, anterior cingulate cortex, and nucleus accumbens is revealed. These findings suggest a safety signal in the medial orbitofrontal cortex and an involvement of the salience network in risk detection.

Introduction

Individuals' motivation to engage with another person is influenced by implicit assumptions about their interaction-associated risks, based on more general assumptions about

the counterpart's traits, such as their overall trustworthiness (Van't Wout and Sanfey, 2008). According to Adolphs and Tusche (2017), such automatic perception of a counterpart as trustworthy results from the automatic comparison of visual, primarily facial cues with one's own internalized sense of a facial expression's trustworthiness and the similarity with one's own face.

The impact of trustworthiness on risk perception and behavior has become a target of research in the context of sexually transmitted infections (STI) such as HIV infection, as the implicit perception of another person as trustworthy (or not) affects the engagement in unprotected sexual interaction (Renner et al., 2012). In light of the high prevalence of risky sexual interactions despite awareness and knowledge of the risks, identifying factors (such as perceived trustworthiness) that modify risk behavior should inform preventive and instructive strategies. With this aim, the present study addressed the relationship between trustworthiness and risk appraisal, more precisely, how ratings of portrayed persons as trustworthy (or not) relate to their evaluation as being HIV carriers. Building on a negative correlation between these two ratings reported by Renner et al. (2012), Häcker et al. (2015), and Schmälzle et al. (2011), the present study used facial stimuli categorized by participants as either HIV carrier (HIV+) or not (HIV−). Whether or to what extent trustworthiness appraisal is linked to shared appraisal of faces reflecting HIV risk or rather related to individual facial appraisal of HIV risk was examined by presenting the series of facial stimuli twice. First, during the fMRI scan with the task of a yes/no decision on the trustworthiness, and secondly, after the fMRI scan, with the instruction to rate the likelihood of the portrayed individual to be an HIV carrier, their trustworthiness and attractiveness.

Hemodynamic neuroimaging (fMRI) served to substantiate the neural correlates of trustworthiness evaluation, with an emphasis on cortical and subcortical structures and circuits mediating risk processing. The salience network (SN) is a large-scale brain network responding to emotional and social cues, including threatening events, such as potential health risks (Ince et al., 2023). The SN as functional integration of dorsal anterior cingulate cortex (dACC) and fronto-insular cortex (FIC) with its connections to medial prefrontal cortex (mPFC), amygdala and substantia nigra, and ventral tegmental area (Bressler and Menon, 2010; Menon, 2011) should be involved in trustworthiness and risk evaluation. Indeed, Häcker et al. (2015) reported that the appraisal of persons as being HIV+ modified activity in, and connectivity between, SN structures, in particular of dorsal insula and mPFC. Following up on this evidence, as well as further studies revealing a prominent role of the amygdala for trustworthiness perception (Sladky et al., 2021), the present study focused on SN involvement and connectivity in trustworthiness appraisal of facial pictures selected to reflect HIV+/- . Moreover, we focused on activation of the posterior superior temporal sulcus (pSTS), a central region in facial expression processing and social cognition (Bruce and Young, 1986; Haxby and Gobbini, 2011; Said et al., 2011) which is also involved in evaluations in trust (Bzdok et al., 2011; Winston et al., 2013).

In addition to trustworthiness, the evaluation of the portrayed individuals as attractive was assessed in a postmeasurement, because attractiveness has been shown to influence perceived trust (Wilson and Eckel, 2006), and both, attractiveness and trust, are related to activation of the nucleus accumbens (Nacc) and the medial orbitofrontal cortex (mOFC; Cloutier et al., 2008; Sladky et al., 2021).

Since individual trust and HIV risk perceptions seem to be related (Barth et al., 2015), we aimed on investigating the association on the basis of a previous study that allowed precategorizing persons as (perceived) HIV− or HIV+ (Häcker et al., 2015). To this end, we used a facial risk perception task (FaRiP) to examine whether we can replicate previous findings (Renner et al., 2012; Schmälzle et al., 2019a) of an association of perceived HIV probability and trustworthiness in behavioral ratings, as well as in activation and connectivity of the salience network, according to a precategorization of the stimulus materials, and based on individual ratings. Renner et al. (2012) demonstrated high inter-rater agreement and high test-retest reliability of HIV probability ratings of the stimulus materials, suggesting the replicability.

We analyzed whether the trustworthiness ratings in the MRI session are associated with the postratings of HIV probability, as well as with attractiveness and trustworthiness, and how these postratings relate to each other. For the fMRI measurements, we expected that HIV+ compared with HIV− stimuli are related to higher activation in amygdala, insula, and mPFC and higher connectivity between insula and mPFC. The same pattern was expected for the categorization according to the individual trustworthiness ratings. Further, we

Correspondence should be addressed to Alexander Wolber at alexander.wolber@uni-konstanz.de.

Copyright © 2025 Wolber et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license, which permits unrestricted use, distribution and reproduction in any medium provided that the original work is properly attributed.

examined how activation in the pSTS and ACC relates to activity during trustworthiness perception. With respect to an association between postratings and activation during the task, we assumed HIV probability ratings to correlate positively with activation in amygdala, insula, and mPFC, trustworthiness ratings to correlate negatively with activation in amygdala, insula and mPFC, and attractiveness ratings to correlate positively with activation in Nacc and mOFC.

Materials and Methods

Study design

The present study was part of a larger project on neural correlates of risk perception and risk behavior, involving two fMRI tasks (the here presented FaRiP task and a Balloon-Analogue-Risk Task, BART), two clinical samples (patients diagnosed with schizophrenia and patients diagnosed with alcohol use disorder), and a healthy control group in a repeated-measurement intervention/replication design. The present report focused on the results of trustworthiness appraisal by an additional sample of healthy individuals as template of average, unaffected behavioral, and fMRI responses. The study was preregistered under <https://osf.io/pkbt6/> and approved by the local ethics board of the University of Konstanz.

Participants

Thirty-four volunteers were recruited via a university platform (SONA), flyer, and word-of-mouth recommendation. Participants meeting the criteria right handedness, eligibility for MRI measurement, and no history of neurological or mental disorders (screened with a telephone version of the SCID-5-interview; [Beesdo-Baum et al., 2019](#)) were included. All participants reported high school or university degree. From initially $N=34$ participants, one participant had to be excluded due to anatomical findings and two cancelled the second appointment. Interestingly, additional $n=6$ had to be excluded due to an acquiescence bias; i.e., they rated all persons as trustworthy in at least one of the two MRI sessions. Thus, the final sample that was included for all analyses consisted of 25 participants (7 male, age $M=23.16$; $SD=2.84$ years). Participants received verbal and written information about study aims and design and signed written informed consent before participating. Participants received a compensation of €20 or course credits for each assessment, as well as their gain in the BART that was completed after the FaRiP task.

Material and procedure

Participants accomplished two identical appointments scheduled 4–5 weeks apart from each other. At each appointment, participants signed written consent, received instruction, practiced the FaRiP task's trustworthiness ratings in five practice trials, and completed 2–3 practice blocks of the BART. The MRI measurements started with a structural scan and were followed by the FaRiP task and the BART. After the MRI sessions, participants completed postscan ratings of trustworthiness, attractiveness, and HIV probability. The present report refers to the FaRiP task and according postratings. The BART results are presented in [Schmidt et al. \(2024\)](#), and the findings on tasks reliability will be presented elsewhere.

The stimulus set had 100 pictures, with 50 pictures of women and 50 pictures of men. The pictures were selected from a previous study, in which they were rated for perceived HIV+ risk (in the original study, male/female participants evaluated only individuals of opposite gender; [Häcker et al., 2015](#)) and were taken from an online database (www.flickr.com). For both, men and women, 25 pictures were used that were classified in the previous study with HIV+ probability and 25 with HIV– probability. Importantly, this categorization relies completely on the previous study, and no information about the real health status of the stimulus persons is available. All pictures were color pictures of Caucasian individuals between 18 and 35 years of age; portraits included face and partly upper body, including clothing, socioeconomic status cues, or situational context ([Häcker et al., 2015](#)). In the present study, hetero- and bisexual participants were shown the pictures of individuals of the opposite gender, and homosexual participants were shown the pictures of individuals of their own gender. After the second appointment, we made explicit to the participants that we are not aware of an association of the appearance of a person and an actual HIV infection.

In an event-related design, across the series of 50 stimuli, each face was presented for 1,000 ms, followed by a jittered interstimulus interval with a fixation cross on a blank screen for 1,000–3,000 ms. Thereafter, the question “is this person trustworthy” appeared on the screen for 2,000 ms together with the response option “yes/no.” Participants were asked to rate each face with respect to high or low trustworthiness by pressing the left button of the response box for “yes,” and the right button for “no,” on a diamond-shaped four-button response pad (Lumina 3G Response System for fMRI – Diamond; www.cedrus.com). The FaRiP was programmed and presented using presentation (Version 21.1 Build 09.05.19, Neurobehavioral Systems; www.neurobs.com). After a 3,000–5,000 ms intertrial interval, the next face appeared (see [Fig. 1](#) for schematic trial).

After the FaRiP task, which lasted ~10 min, participants completed the BART and returned to a separate room for the post-scan ratings of the identical series of facial stimuli, now presented on a laptop. They were instructed to judge trustworthiness, the risk of the portrayed person to be HIV carrier, and attractiveness on a 7-point Likert scale at self-paced speed.

fMRI data acquisition

We acquired fMRI data using a 3 T Siemens Magnetom Skyra (Siemens Healthcare) with syngo MR D13 software. An EPI-sequence with the following parameters was applied: TR/TE, 2,500/30 ms; flip angle, 80°; 38 interleaved axial slices

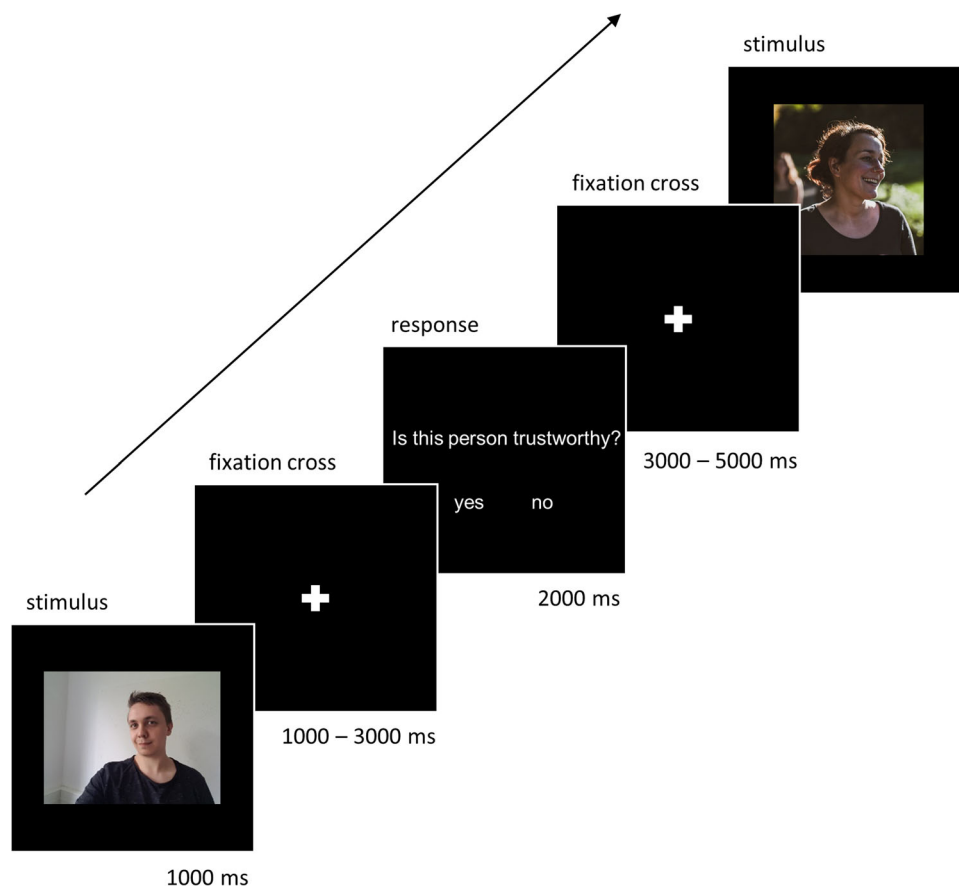


Figure 1. Schematic trial with exemplary pictures of the authors A.W. and D.M. that were not used in the study.

(slice thickness, 3.4 mm, no gap); FOV, 218 mm; number of acquisitions, 190; matrix, 64×64 ; volume size, $3.4 \times 3.4 \times 3.4$ mm. In addition, a high-resolution T1-weighted scan was performed with the following parameters: MPRAGE-Sequence; TR/TE, 2,500/4.06 ms; flip angle, 7° ; 192 axial slices (slice thickness, 1 mm, no gap); number of volumes, 1; FOV, 256; matrix, $1 \times 1 \times 1$.

Data analyses

Behavioral indices/ratings. For the FaRiP task, we determined the number of in-scanner facial stimuli categorized as trustworthy (yes/no) and calculated the sum of both sessions. A χ^2 test was performed for yes/no categorization relative to the HIV+/- precategorizations. For the postscan ratings, we analyzed the graded evaluations of trustworthiness, HIV probability, and attractiveness. An average over the two sessions was calculated for each measure. Paired sample *t* tests for comparisons between trustworthiness, HIV probability, and attractiveness ratings depending on the HIV precategorizations were calculated. For all behavioral analyses, we used IBM SPSS Statistics V27 (<https://www.ibm.com/>).

Hemodynamic activity patterns. We preprocessed and analyzed the fMRI data with SPM12 (<https://www.fil.ion.ucl.ac.uk/spm/software/spm12/>). Preprocessing consisted of slice-time correction to the middle slice, realignment to the mean image, with unwarping on the *y*-axis, normalization and resampling to a $3 \times 3 \times 3$ mm voxel size, and smoothing with a 9 mm (FWHM) Gaussian kernel.

First-level analyses were based on seven GLMs. GLM1 covered the precategorization and had one regressor for high risk (HIV+) and one for low risk (HIV-) stimuli. GLM2 addressed the individual trust ratings and accordingly had one regressor for trustworthy and one for untrustworthy rated stimuli. In addition, GLM3, 4, and 5 were set up for the parametric modulation to investigate brain-behavior associations, in which one regressor contained all stimuli and was modulated by the postratings of HIV probability, attractiveness, and trustworthiness. To examine functional connectivity, GLMs 6 and 7 that were identical to GLM1 and GLM2 were set up, with the only exception being the duration of 1 s instead of 0 s. For all GLMs, there was an additional regressor representing all button presses and six motion regressors from the realignment as regressors of no interest. Furthermore, each first-level model included both sessions with session as additional regressor. The high-pass filter was always set to 256 s. We used generalized psychophysiological interactions (gPPI; McLaren et al., 2012) with the default

settings provided in McLaren's gPPI toolbox, using bilateral insula as seed regions. Specifically, in the toolbox, the time series of the seed regions are extracted at a default threshold of 0.5 and multiplied with the condition-specific regressors.

At the second level, group analyses were calculated with t tests for the following contrasts from the first-level analyses: GLM 1 and GLM 6, HIV+, HIV-, HIV+ > HIV-, HIV- > HIV+; GLM 2 and GLM 7, untrustworthy, trustworthy, untrustworthy > trustworthy, trustworthy > untrustworthy; GLM3, 4 and 5, each stimulus event \times parametric modulator (postratings of attractiveness, trustworthiness, or HIV probability). We applied Probabilistic Threshold-Free Cluster Enhancement (TFCE; Smith and Nichols, 2009) to improve sensitivity in detecting fMRI signal differences and therefore to make our second-level results more robust and reliable. Unlike traditional cluster-based methods, TFCE enhances both small, intense activations and larger, less intense clusters, providing a better balance for detecting subtle but meaningful neural responses. This is especially important for our study, in which we examine complex social-cognitive processes, where activation patterns may be distributed across various brain regions or may be present in very small brain regions. The significance level was set to $p < 0.05$ FWE corrected. In addition, exploratory analyses with a more lenient threshold of $p < 0.001$ uncorrected were conducted. ROI analyses were performed for amygdala, insula, mPFC, mOFC, Nacc, pSTS, and ACC with small-volume correction (svc) with FWE < 0.05 peak voxel correction. Cluster size was set to a standard threshold of $k = 5$ for all analyses.

The insula mask and the mPFC mask were created with MARINA (Walter, 2003). The mOFC mask was taken from the labels atlas in WFU-PickAtlas (Maldjian et al., 2003). Nacc mask was taken from automatic anatomy labeling atlas 3 (Rolls et al., 2020a,b). The amygdala mask and the ACC mask were taken from Neuromorphometrics atlas (Neuromorphometrics, as included in SPM12). The mask for the pSTS was taken from Mier et al. (2010).

Results

Behavior

A χ^2 was used to compare HIV precategorization and categorical trustworthiness ratings (Table 1). Results show a significant interaction between HIV precategorization and trustworthiness in-scanner ratings ($p < 0.001$). Specifically, participants rated on average 59.44% (± 14.71) of HIV+ stimuli as untrustworthy and 72.96% (± 15.13) of HIV- as trustworthy.

Postratings of trustworthiness ($t_{(24)} = 12.14$, $p < 0.001$, $d = 0.63$) and attractiveness ($t_{(24)} = 3.18$, $p = 0.004$, $d = 0.67$) were significantly higher for HIV- than those for HIV+ preclassified persons (Fig. 2), and HIV probability postratings were significantly lower ($t_{(24)} = -7.01$, $p < 0.001$, $d = 0.76$; Fig. 2). Trustworthiness ratings during the fMRI measurements correlated significantly positive with trustworthiness postratings ($r = 0.73$, $p < 0.001$), on a trend level with attractiveness postratings ($r = 0.38$, $p = 0.061$), but not with HIV probability postratings ($r = -0.33$, $p = 0.107$). Trustworthiness postratings correlated significantly positive with attractiveness postratings ($r = 0.55$, $p = 0.004$), but not with HIV probability postratings ($r = -0.33$, $p = 0.105$). Attractiveness postratings did not correlate with HIV probability postratings ($r = -0.04$, $p = 0.854$).

Confirmatory fMRI analyses

Analyses according to the preclassification

Comparison of HIV+ and HIV- revealed no significant results when applying FWE correction or ROI analyses, neither for activation nor for left and right insula connectivity.

Analyses based on individual trustworthiness ratings

Whole-brain analyses of faces rated as not trustworthy (compared with trustworthy) revealed activation in left insula and left putamen, both reaching into inferior frontal cortex (Fig. 3, Table 2). For this contrast, small volume analyses showed activation in bilateral insula (left: MNI: -42, 14, -1; $T = 7.31$; $p \text{ svc} \leq 0.001$; $k = 471$; right: MNI: 42, 23, 2; $T = 4.99$; $p \text{ svc} = 0.008$; $k = 221$) and left amygdala (MNI: -27, 2, -19; $T = 4.09$; $p \text{ svc} = 0.006$; $k = 12$). Again, no significant left or right insula connectivity differences between conditions were found, neither in the whole brain nor in the ROI analyses.

Brain-behavior associations

Parametric modulation of attractiveness and graded trustworthiness ratings revealed an activation cluster in right Nacc (MNI: 18, 5, -7; $T = 3.71$; $p \text{ svc} = 0.029$; $k = 59$), but not in left Nacc or mOFC. No further results of the parametric modulation were significant, neither for the whole brain nor the ROI analyses; i.e., there were no significant associations between HIV probability, as well as trustworthiness ratings and brain activation.

Table 1. Distribution of in-scanner trustworthiness ratings in relation to the precategorization into HIV positive and HIV negative

		Trustworthiness	
		Not trustworthy	Trustworthy
HIV	+	59.8%	40.2%
	-	26.8%	73.2%

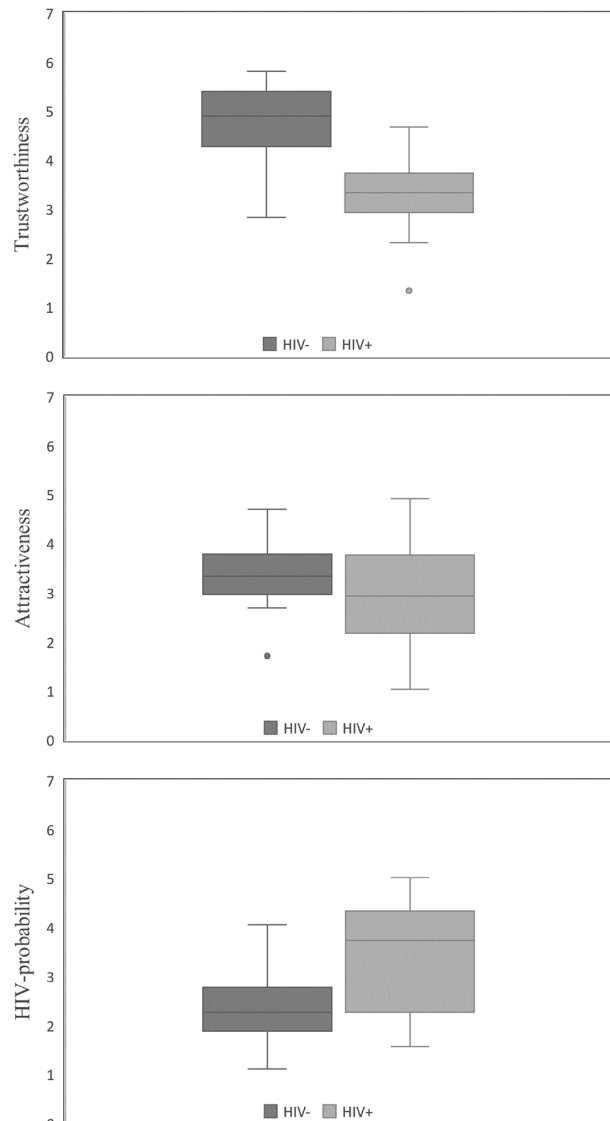


Figure 2. Trustworthiness, attractiveness, and HIV probability post ratings previously categorized in HIV– and HIV+.

Exploratory fMRI analyses

Analyses according to the preclassification

Whole-brain analyses with a liberal significance threshold ($p < 0.001$, $k = 5$) showed higher activation in mOFC, left superior frontal gyrus, and right angular gyrus for HIV– compared with HIV+ (Fig. 4, right; Table 3). Small volume analyses revealed higher activation in left pSTS (MNI: $-45, -70, 20$; $T = 4.16$; $p \text{ svc} = 0.037$; $k = 108$), mPFC (MNI: $-3, 47, -13$; $T = 5.89$; $p \text{ svc} = 0.003$; $k = 328$), and mOFC (left: MNI: $-3, 38, -22$; $T = 5.66$; $p \text{ svc} = 0.002$; $k = 18$; right: MNI: $0, 44, -19$; $T = 5.63$; $p \text{ svc} = 0.002$; $k = 14$). Further ROI analyses as well as whole-brain analyses revealed no regions with differential activation or significantly enhanced left or right insula connectivity for HIV– versus HIV+ preclassified images and vice versa.

Analyses based on individual trustworthiness ratings

Additional exploratory small volume analyses for untrustworthy (compared with trustworthy) faces showed activation in bilateral Nacc (left: MNI: $-18, 5, -7$; $T = 6.76$; $p \text{ svc} < 0.001$; $k = 81$; right: MNI: $18, 11, -4$; $T = 5.52$; $p \text{ svc} = 0.001$; $k = 67$) and left ACC (MNI: $-9, 26, 26$; $T = 4.24$; $p \text{ svc} = 0.034$; $k = 133$).

Whole-brain analyses for the reverse contrast with a more lenient significance threshold revealed activation in mOFC (Fig. 4, left; Table 4). Exploratory small volume analyses for the same contrast revealed no significant activation differences in the ROIs.

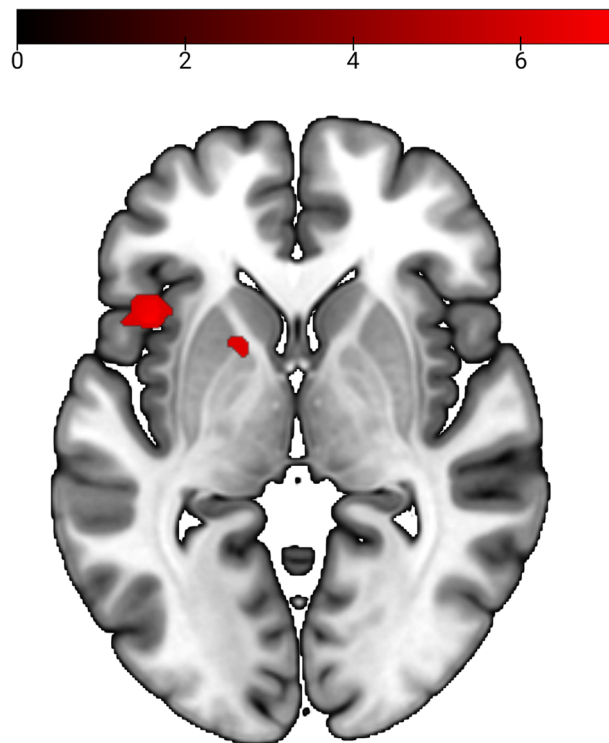


Figure 3. Contrast not trustworthy > trustworthy; $p < 0.05$ (FWE corrected), $k = 5$.

Table 2. Activation in the contrast not trustworthy > trustworthy, $k = 5$

Not trustworthy > trustworthy	Area	BA	k	MNI			t value	$P_{FWE < 0.05}$
				x	y	z		
	Left insular cortex	13	58	-42	14	-1	7.31	0.004
	Left inferior frontal gyrus	13		-48	14	-5	7.08	0.007
	Left insular cortex	13		-33	14	11	6.24	0.047
	Left putamen		42	33	26	7	6.76	0.015
	Left inferior prefrontal gyrus	47		39	11	29	6.72	0.016
	Left inferior prefrontal gyrus	47		33	26	5	6.39	0.035

Discussion

In the present fMRI study, participants judged the trustworthiness of individuals preclassified as having high or low HIV probability. We investigated how the participants' judgments of trustworthiness, HIV risk, and attractiveness are interrelated and associated to activation of regions involved in salience and trust perception, specifically dACC, insula, mPFC, Nacc, as well as amygdala and pSTS. In addition, we analyzed functional connectivity of the bilateral insulae.

Persons preclassified as HIV-, i.e., less likely to be infected with HIV, were rated more often as trustworthy than persons previously classified as HIV+. This could be interpreted as hinting toward a cross-sample perception of other persons' safety. Also, in the postratings with graded evaluations, trustworthiness ratings were higher and HIV probability was lower, for the HIV- than HIV+ stimuli. fMRI-based cortical activity both upon preclassification and upon trustworthiness evaluation (albeit only with an uncorrected significance threshold) indicated activation in mOFC. This mOFC response seems not to derive from a joint representation of attractiveness and trustworthiness, because attractiveness ratings were associated with Nacc, but not with mOFC activation. Rather, we conclude that for persons perceived as trustworthy and safe (i.e., individually evaluated as trustworthy and precategorized as HIV-), the mOFC response may signal a rewarding property (Rolls et al., 2020a). For the persons preclassified as HIV+, the agreement between in-scanner trustworthiness and prior HIV ratings was ~60%, suggesting a higher agreement of trustworthiness and HIV- classifications (>70% agreement) than of untrustworthiness and HIV+ classification. While this does not necessarily reflect the objective trustworthiness of a person, it indicates that the perception of safety going along with mOFC response has higher convergence across study samples than the perception of riskiness.

Importantly, while attractiveness ratings did not differ between HIV+ and HIV-, attractiveness and trustworthiness were associated. Previous studies have shown a (complex) relationship between ratings of attractiveness and trustworthiness

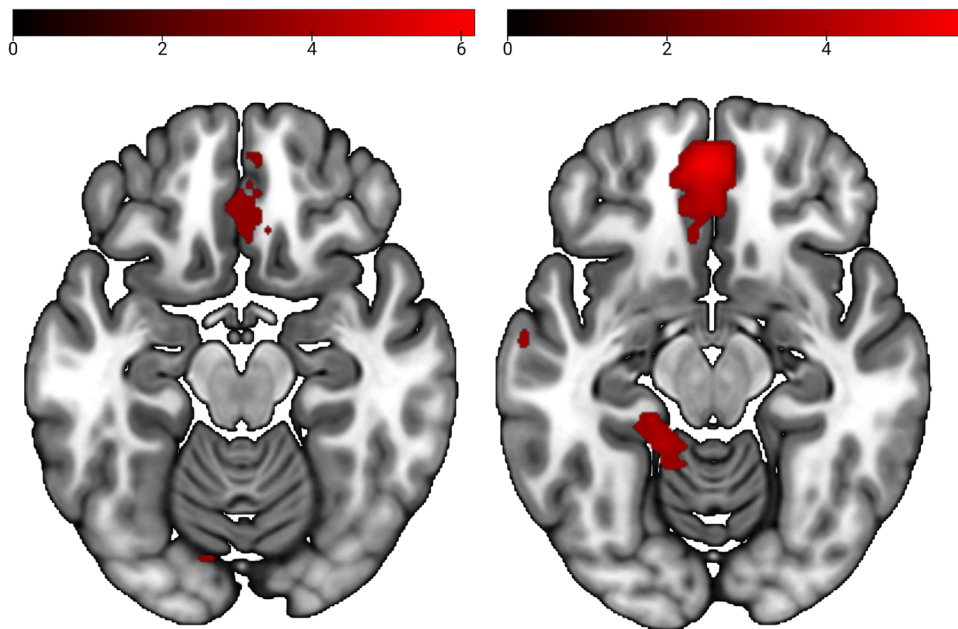


Figure 4. Left, Contrast trustworthy > not trustworthy; $p < 0.001$ (uncorrected), $k = 5$. Right, Contrast HIV- > HIV+; $k = 5$; $p = 0.001$ (uncorrected).

Table 3. Activation in the contrast HIV- > HIV+, $k = 5$

Area	BA	k	MNI			t value	$p < 0.001$
			x	y	z		
Orbitofrontal area	11	218	-3	47	-13	5.89	<0.001
Orbitofrontal area	11		-3	38	-22	5.66	<0.001
Orbitofrontal area	11		0	29	-22	5.12	<0.001
Left inferior temporal gyrus	20	36	-60	-7	-25	5.07	<0.001
Left middle temporal gyrus	21		-60	-1	-19	4.75	<0.001
Left angular gyrus	39	64	-45	-67	23	4.39	<0.001
Left angular gyrus	39		-67	-1	-19	3.66	0.001
Right retrosplenial cingulate cortex	29	643	15	-46	5	4.38	<0.001
Left primary visual cortex	17		-12	-85	-1	4.30	<0.001
Left visual association cortex	19		-18	-43	-10	4.27	<0.001
Right angular gyrus	39	52	51	-58	23	4.22	<0.001

Table 4. Activation in the contrast trustworthy > not trustworthy, $k = 5$

Area	BA	k	MNI			t value	$p < 0.001$
			x	y	z		
Left visual association cortex	18	240	-12	-82	-10	6.22	<0.001
Left lingual gyrus			-24	-73	-4	4.92	<0.001
Left lingual gyrus			-30	-73	2	3.64	0.001
Orbitofrontal area	11	13	6	53	-10	4.00	<0.001
Orbitofrontal area	11	27	0	35	-16	3.97	<0.001
Medial frontal gyrus	25		3	23	-16	3.87	<0.001

(Gutierrez-Garcia et al., 2019; Li and Liu, 2021), but not between attractiveness and HIV probability (Schmälzle et al., 2011). In contrast to earlier findings that showed a relationship between attractiveness perception and probability of careless (unprotected) sexual behavior (Eleftheriou et al., 2016, 2019), our results suggest that the attraction to another person is related to their perceived trustworthiness but seems not to lead to a faulty perception of safety with regard to STI.

Although the behavioral ratings and the activation in mOFC suggest an association between preclassified HIV probability and individual trustworthiness ratings, only individual ratings (as untrustworthy), but not preclassification (as HIV+), resulted in activation of the SN and related regions involved in salience and trust processing. Facial stimuli rated as untrustworthy in comparison with faces rated as trustworthy were associated with activity in bilateral insula, bilateral Nacc, left ACC, and left amygdala. In agreement with meta-analytical findings (Bzdok et al., 2011), this demonstrates a

prominent role of the amygdala for trustworthiness judgments. The meta-analysis of Bzdok and colleagues (Bzdok et al., 2011) also shows activation in pSTS for trustworthiness tasks. The pSTS is generally involved in social perception and interaction (Hein and Knight, 2008; Deen et al., 2015) but seems not to be differentially activated by trustworthy versus untrustworthy faces. As pSTS activation, however, was enhanced in response to faces preclassified as HIV– in contrast to HIV+, the pSTS might play a particular role in safety perception. However, this finding warrants replication, before firm conclusions can be drawn. While Häcker and colleagues reported enhanced insula and mPFC activation for persons rated as HIV+, we found higher insula, ACC, Nacc, and amygdala activation for untrustworthy versus trustworthy ratings. Insula, amygdala, ACC, and Nacc are involved in (social) salience processing (Menon and Uddin, 2010; Schmidt et al., 2019). As Häcker and colleagues propose (Häcker et al., 2015), participants have intuitive sensing of risk; i.e., a “HIV risk feeling.” As activity in insula, but not connectivity between insula and mPFC, differed between persons rated as trustworthy versus untrustworthy, the regional activity of certain structures within the SN seems modulated by such a risk feeling, but not the communication between the selected structures. The lack of connectivity results from our data stands in contrast to previous studies that reported enhanced functional connectivity dependent on HIV categorization (Renner et al., 2012; Häcker et al., 2015; Schmälzle et al., 2019b).

Across various analyses of the here presented data (both for HIV+ vs HIV– faces and trustworthy vs untrustworthy faces), there is a notable absence of significant insula connectivity differences between conditions. This could imply that the difference in processing risk-related social information within our task is more localized to specific brain regions rather than involving broader neural networks. However, a more likely explanation for our lack of connectivity findings might be our selection of the seed region rather than an absence of differential connectivity effects. Specifically, our choice of the insula as seed region was based on studies focusing on perceived HIV risk in the same stimulus set (Häcker et al., 2015). The insula is involved in interoception, emotional awareness, risk perception, and certain aspects of emotion processing, such as disgust and fear. In addition, it is often activated when processing emotionally salient or risky stimuli (Uddin et al., 2017), which makes it a reasonable candidate for examining responses to faces associated with HIV risk and low trust. While we find insula activation when persons were perceived as not trustworthy, connectivity might not change between not trustworthy and trustworthy/HIV+ and HIV– perception. Possibly, trustworthiness judgments may be related more strongly to variation in networks involved in social reward evaluation and value-based decisions. Relevant regions of interest that may serve as seed regions for future connectivity analyses in the context of trustworthiness are mPFC, amygdala, OFC, and even the Nacc.

To our knowledge, our study is the first attempt to differentiate SN activity for faces preclassified as HIV+ versus HIV–. In our study, insula, ACC, Nacc, and amygdala were only related to untrustworthiness ratings, but not to HIV+ preclassification. Since we applied an implicit risk assessment by participants rating trustworthiness and not HIV probability during fMRI, we cannot draw conclusions about an involvement of the SN in individual HIV judgments as Häcker et al. (2015). The discrepancy in SN activation in our study between individual trustworthiness ratings and precategorized HIV probability emphasizes the importance of individual trustworthiness perception. In contrast to the behavioral findings, the fMRI findings suggest a reduced transferability of HIV stereotypes to individual salience signals of trustworthiness.

Such SN activation patterns upon facial images judged as untrustworthy may have evolutionary roots. Rapid judgment of social counterpart's reduced trustworthiness and their risk of being a life-endangering disease carrier should ignite the brain's alertness response and prepare quick (risk–caution) decisions, thus bearing survival advantage. Even though average ratings indicate prototypic features that are related to trustworthiness perception (Todorov et al., 2008; Sessa and Meconi, 2015), our findings suggest that the features of presented facial images activate the alarm system seems to be highly individual in nature. Importantly, these intuitive perceptions might result in false risk and/or safety signals. We had neither objective information about the general trustworthiness of our stimulus persons nor their health status.

Limitations

Several limitations of our study have to be noted. First, the stimulus persons were shown in different naturalistic contexts. Although this makes the pictures less artificial, different contexts may have influenced the trustworthiness ratings. In addition, there was no matching for arousal and valence for HIV+ and HIV– stimuli. Matching the stimulus material with regard to these characteristics would have been detrimental to the manipulation of the independent variable and would have reduced the probability of finding differences in trustworthiness ratings between HIV+ and HIV– precategorized persons. By definition, persons perceived as not trustworthy elicit arousal and negative valence, while persons perceived as trustworthy should be rather related to positive valence and limited arousal. Second, participants were not provided with a definition or contextualization of trustworthiness prior to the experiment, which may have increased the interindividual variability of subjective definitions. To address this, we specified the context for the trustworthiness judgments in a follow-up study (Wolber et al., in preparation) and mentioned that participants should imagine a dating context.

Third, six participants had to be excluded due to a positive bias; i.e., not rating any of the stimulus persons as untrustworthy in at least one of the sessions which is an interesting finding itself. This, however, resulted in a final sample of 25 participants that is rather small for current standards in fMRI studies and which limits the reliability of the results. To

counteract, we combined the sessions that were 4–5 weeks apart in our analyses, since it was shown that aggregating data from multiple sessions can increase reliability (Noble et al., 2017).

A fourth and most important limitation is the partial stimulus overlap in the “preclassified” and “individual rating” analyses (59.8% for HIV+ and untrustworthy judgments, 73.2% for HIV– and trustworthy judgments). Both analyses share individually differing proportions of stimulus materials. Thus, a similarity in neural activation for preclassification and individual trustworthiness analyses that occurred for the persons rated as trustworthy seems logical. Since we used the same stimulus set for both sessions, independent analyses are not possible. Future studies may investigate both concepts with two stimulus sets with two sessions and analyze individual ratings in session one with stimulus set one and preclassification in session two stimulus set two or might ask in session one for trustworthiness and in session two for HIV probability.

Conclusion

Taken together we found within-sample congruence for trustworthiness and HIV perception, while across-sample congruence was higher for trustworthiness and HIV– than for untrustworthiness and HIV+ judgments. This was reflected in mOFC activation during HIV– preclassified images and individual trustworthiness ratings, and regions involved in salience and trust processing for individual untrustworthy-ratings, but not HIV+ preclassification. These findings suggest cross-sample agreement of safety perception in behavior and brain activation but highlight the importance of individuality of risk perception.

References

- Adolphs R, Tuschke A (2017) From faces to prosocial behavior: cues, tools, and mechanisms. *Curr Dir Psychol Sci* 26:282–287.
- Barth A, Schmäzle R, Hartung F-M, Renner B, Schupp HT (2015) How target and perceiver gender affect impressions of HIV risk. *Front Public Health* 3:137393.
- Beesdo-Baum K, Zaudig M, Wittchen H (2019) *Structured clinical interview for DSM-5—clinical version*. Göttingen: Hogrefe.
- Bressler SL, Menon V (2010) Large-scale brain networks in cognition: emerging methods and principles. *Trends Cogn Sci* 14:277–290.
- Bruce V, Young A (1986) Understanding face recognition. *Br J Psychol* 77:305–327.
- Bzdok D, Langner R, Caspers S, Kurth F, Habel U, Zilles K, Laird A, Eickhoff SB (2011) ALE meta-analysis on facial judgments of trustworthiness and attractiveness. *Brain Struct Funct* 215:209–223.
- Cloutier J, Heatherton TF, Whalen PJ, Kelley WM (2008) Are attractive people rewarding? Sex differences in the neural substrates of facial attractiveness. *J Cogn Neurosci* 20:941–951.
- Deen B, Koldewyn K, Kanwisher N, Saxe R (2015) Functional organization of social perception and cognition in the superior temporal sulcus. *Cereb Cortex* 25:4596–4609.
- Eleftheriou A, Bullock S, Graham CA, Skakoon-Sparling S, Ingham R (2019) Does attractiveness influence condom use intentions in women who have sex with men? *PLoS One* 14:e0217152.
- Eleftheriou A, Bullock S, Graham CA, Stone N, Ingham R (2016) Does attractiveness influence condom use intentions in heterosexual men? An experimental study. *BMJ Open* 6:e010883.
- Gutierrez-Garcia A, Beltran D, Calvo MG (2019) Facial attractiveness impressions precede trustworthiness inferences: lower detection thresholds and faster decision latencies. *Cogn Emot* 33:378–385.
- Häcker FE, Schmäzle R, Renner B, Schupp HT (2015) Neural correlates of HIV risk feelings. *Soc Cogn Affect Neurosci* 10:612–617.
- Haxby JV, Gobbini MI (2011) Distributed neural systems for face perception. In: *The Oxford handbook of face perception* (Calder AJ, ed). Oxford: Oxford University Press.
- Hein G, Knight RT (2008) Superior temporal sulcus—it’s my area: or is it? *J Cogn Neurosci* 20:2125–2136.
- Ince S, Steward T, Harrison BJ, Jamieson AJ, Davey CG, Agathos JA, Moffat BA, Glarin RK, Felmingham KL (2023) Subcortical contributions to salience network functioning during negative emotional processing. *Neuroimage* 270:119964.
- Li N, Liu N (2021) The nonlinear and gender-related relationships of face attractiveness and typicality with perceived trustworthiness. *Front Psychol* 12:656084.
- Maldjian JA, Laurienti PJ, Kraft RA, Burdette JH (2003) An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *Neuroimage* 19:1233–1239.
- McLaren DG, Ries ML, Xu G, Johnson SC (2012) A generalized form of context-dependent psychophysiological interactions (gPPI): a comparison to standard approaches. *Neuroimage* 61:1277–1286.
- Menon V (2011) Large-scale brain networks and psychopathology: a unifying triple network model. *Trends Cogn Sci* 15:483–506.
- Menon V, Uddin LQ (2010) Saliency, switching, attention and control: a network model of insula function. *Brain Struct Funct* 214:655–667.
- Mier D, Lis S, Neuthe K, Sauer C, Esslinger C, Gallhofer B, Kirsch P (2010) The involvement of emotion recognition in affective theory of mind. *Psychophysiology* 47:1028–1039.
- Noble S, Spann MN, Tokoglu F, Shen X, Constable RT, Scheinost D (2017) Influences on the test–retest reliability of functional connectivity MRI and its relationship with behavioral utility. *Cereb Cortex* 27:5415–5429.
- Renner B, Schmäzle R, Schupp HT (2012) First impressions of HIV risk: it takes only milliseconds to scan a stranger. *PLoS One* 7:e30460.
- Rolls ET, Cheng W, Feng J (2020a) The orbitofrontal cortex: reward, emotion and depression. *Brain Commun* 2:fcaa196.
- Rolls ET, Huang C-C, Lin C-P, Feng J, Joliot M (2020b) Automated anatomical labelling atlas 3. *Neuroimage* 206:116189.
- Said CP, Haxby JV, Todorov A (2011) Brain systems for assessing the affective value of faces. *Philos Trans R Soc B Biol Sci* 366:1660–1670.
- Schmäzle R, Hartung FM, Barth A, Imhof MA, Kenter A, Renner B, Schupp HT (2019a) Visual cues that predict intuitive risk perception in the case of HIV. *PLoS One* 14:e0211770.
- Schmäzle R, Imhof MA, Kenter A, Renner B, Schupp HT (2019b) Impressions of HIV risk online: brain potentials while viewing online dating profiles. *Cogn Affect Behav Neurosci* 19:1203–1217.
- Schmäzle R, Schupp HT, Barth A, Renner B (2011) Implicit and explicit processes in risk perception: neural antecedents of perceived HIV risk. *Front Hum Neurosci* 5:43.
- Schmidt SNL, Fenske SC, Kirsch P, Mier D (2019) Nucleus accumbens activation is linked to salience in social decision making. *Eur Arch Psychiatry Clin Neurosci* 269:701–712.

- Schmidt SNL, Sehrig S, Wolber A, Rockstroh B, Mier D (2024) Nothing to lose? Neural correlates of decision, anticipation, and feedback in the balloon analog risk task. *Psychophysiology* 61: e14660.
- Sessa P, Meconi F (2015) Perceived trustworthiness shapes neural empathic responses toward others' pain. *Neuropsychologia* 79:97–105.
- Sladky R, Riva F, Rosenberger LA, van Honk J, Lamm C (2021) Basolateral and central amygdala orchestrate how we learn whom to trust. *Commun Biol* 4:1–9.
- Smith SM, Nichols TE (2009) Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage* 44:83–98.
- Todorov A, Baron SG, Oosterhof NN (2008) Evaluating face trustworthiness: a model based approach. *Soc Cogn Affect Neurosci* 3:119–127.
- Uddin LQ, Nomi JS, Hébert-Seropian B, Ghaziri J, Boucher O (2017) Structure and function of the human insula. *J Clin Neurophysiol* 34:300–306.
- Van't Wout M, Sanfey AG (2008) Friend or foe: the effect of implicit trustworthiness judgments in social decision-making. *Cognition* 108:796–803.
- Walter B (2003) MARINA: an easy to use tool for the creation of MAsks for Region of Interest Analyses. The 9th International Conference on Functional Mapping of the Human Brain, June 19–22, 2003, New York, NY.
- Wilson RK, Eckel CC (2006) Judging a book by its cover: beauty and expectations in the trust game. *Polit Res Q* 59:189–202.
- Winston JS, Strange BA, O'Doherty J, Dolan RJ (2013) Automatic and intentional brain responses during evaluation of trustworthiness of faces. In: *Social neuroscience* (Berntson GB, Cacioppo JT, eds), pp 199–210. New York: Psychology Press.