Received: 18 February 2022 | Accepted: 27 September 2022

DOI: 10.1111/apps.12437

ORIGINAL ARTICLE

APPLIED PSYCHOLOGY IAAP

The neuroscience of trust violation: Differential activation of the default mode network in ability, benevolence and integrity breaches

Lisa van der Werff[1] | **Deirdre O'Shea**[2] | **Graham Healy**[3] | **Finian Buckley**[4] | **Colette Real**[1,4] | **Michael Keane**[5] | **Theo Lynn**[1]

[1]Irish Institute of Digital Business, DCU Business School, Dublin City University, Dublin, Ireland

[2]Kemmy Business School, University of Limerick, Limerick, Ireland

[3]School of Computing, Dublin City University, Dublin, Ireland

[4]DCU Business School, Dublin City University, Dublin, Ireland

[5]Actualise Neurofeedback Clinic, Dublin, Ireland

Correspondence
Lisa van der Werff, Irish Institute of Digital Business, DCU Business School, Dublin City University, Dublin, Ireland.
Email: lisa.vanderwerff@dcu.ie

Abstract

Trust is widely regarded as being foundational in workplace relationships. The violation of interpersonal trust results in a range of negative affective, cognitive and behavioural consequences for the injured party. However, research has yet to isolate the specific neural areas and processes activated when different types of interpersonal trust are breached. Using electroencephalogram with 68 participants, we identified the effects of three distinct types of trust violations—ability violation, integrity violation and benevolence violations—on electrical brain activity. Our findings indicate that trust violations are processed in social cognitive-related brain areas. Specifically, our results identify the significance of the default mode network (DMN), relevant to the processing of social information, in trust violation and further isolated distinct activity for ability, integrity and benevolence trust violation, with integrity violations demonstrating the greatest reaction in the DMN. Benevolence violations generated the next greatest reaction but were not significantly different from the ability violations. This potential distinction may be worth further investigation in future research. Our

findings highlight the potential importance of the DMN in processing cues regarding the trustworthiness of others and the distinctiveness of the processing of violation cues of the three facets of trustworthiness.

KEYWORDS

default mode network, electroencephalogram (EEG), integrity, trust violation, trustworthiness

INTRODUCTION

The ability to build, maintain and repair trust in workplace relationships has been critical to effective social exchange in organisations with outcomes including performance, cooperation and employee agility (Doeze Jager et al., 2022; Fulmer & Gelfand, 2012). The willingness to be vulnerable that defines trust is based on positive expectations of the trustworthiness of the other party (Mayer et al., 1995). Violations of trust, where the other party commits an act that lowers expectations of trustworthiness, are common and undermine trustor perceptions of both the trustee and their relationship (Lewicki & Brinsfield, 2017). Given the potential for trust violation to undermine the cooperation that is so vital to the functioning of dyadic relationships, groups and organisations, developing a thorough understanding of the mechanisms underlying trust violation is crucial.

One of the most studied aspects of trust violation is the type of trustworthiness that has been violated. Trustworthiness is an aggregate evaluation of the ability, benevolence and integrity of another party (Mayer et al., 1995); this very well-established conceptualisation is sometimes referred to as the ABI model of trustworthiness. Violations of ability include information about the lack of competence of the trustee; violations of benevolence include information that the trustee is unlikely to act in the interests of the trustor; and violations of integrity relate to the values and principles of the trustee, such as a lack of honesty. Empirical research on trust violation indicates that the type of trustworthiness violated has repercussions for the effectiveness of various methods of trust repair (Kim, 2018; Kim et al., 2004). These results have led scholars to theorise that ability and integrity violations involve different mechanisms of cognitive processing (Kim et al., 2009). Unfortunately, the predominance of experimental vignette self-report studies in the field has made it difficult to ascertain empirically if integrity and ability are processed differently or to examine the underlying processes. Furthermore, trust violation experiments have tended to focus on ability and integrity and have yet to fully consider how violations of benevolence are processed. Given the prominence of the ABI model of trustworthiness, the relative exclusion of benevolence from the empirical trust violation literature leaves an important gap in our understanding of how different types of trust violations impact trustors.

Our paper intervenes at this time to investigate the theorised differences between the categories of trust violation using a neuroscientific approach. Neuroimaging data can assist us in exploring the mechanisms underlying particular perceptions and behaviours (Haesevoets et al., 2018), in a way that avoids many of the issues with the more common survey approach (Waldman et al., 2017). These benefits are particularly timely for the trust literature where debate about the dimensionality of trust continues (Legood et al., 2022; van

Knippenberg, 2018), and preliminary distinctions observed between the different antecedents of trust violation would benefit from multi-method confirmation. We explore the neurological response to trust violation by examining changes in brain activity as indicators of responses to experienced ability, benevolence and integrity violations using electroencephalogram (EEG). Previous research has demonstrated changes in brain activity as a result of a loss versus gain frame (e.g. King-Casas et al., 2005), and for global trust violation (e.g. Dimoka, 2010; Riedl & Javor, 2012), but has not demonstrated the effects of violating different types of trust on brain activity. Building on this previous work, our study investigates the argument that different types of violation are processed differently using neuroimaging techniques.

HYPOTHESIS DEVELOPMENT

Although neuroscientific work specifically on trust has been relatively sparse, further progress has been made in the wider area of social cognition which includes the range of processes involved in understanding and interacting with others. An important aspect of social cognition is that it enables us to make inferences about people's internal states, including their emotions and motives, from their social behaviour (Adolphs, 2009). Trust, as a willingness to be vulnerable based on positive expectations (Mayer et al., 1995), is a process defined by forming perceptions and understanding of others, on which to base interaction. Whereas other dispositional, emotional and motivational drivers play a role in trust, the cognitive perceptions encompassed by trustworthiness are undoubtedly the most commonly studied antecedents (Baer & Colquitt, 2018).

Trustworthiness perceptions are formed through the consideration of the attributes of the other party to form an expectation on which a willingness to be vulnerable can be based (Mayer et al., 1995). The ability to recognise and behave with respect to socially relevant information may be uniquely human (Adolphs, 2009), and this process is central to both the formation of trust and its violation. To experience a violation of trust, an individual must perceive that their expectations of the other party have not been met (Haselhuhn et al., 2015) and that their beliefs about the other party's ability, benevolence or integrity were overly optimistic. This process requires individuals to cognitively assess the actions of their colleagues and to compare these actions to currently held judgements of trustworthiness.

Neurobiological and neuropsychological research has found a number of brain structures, networks and processes that are relevant to the processing of social information (Adolphs, 2001). For instance, the prefrontal cortices play a specific role in reasoning about the mental states of other people (Adolphs, 2001). In particular, the default mode network (DMN) has been identified repeatedly for its involvement in social cognitions including cooperation (Amodio & Frith, 2006; Bressler & Menon, 2010) and in the social understanding of others, including emotion perception, empathy, theory of mind and morality (Li et al., 2014). The DMN is an anatomically defined brain system that preferentially activates when individuals are not focused on the external environment, and core areas of the DMN include the medial posterior cortex, medial prefrontal cortex and the bilateral inferior parietal lobule (Buckner et al., 2008; Li et al., 2014). Regarding the neurocorrelates of trust and its closely related concepts, research involving economic games also highlights brain areas related to the DMN as being central to cooperation (Xiang et al., 2012). Furthermore, Dimoka (2010) demonstrates activation of centres associated with reward, uncertainty and prediction (e.g. anterior paracingulate cortex and orbitofrontal cortex) using fMRI when examining trust during an

e-commerce experiment. Conversely, individuals appear to use the amygdala in the processing of untrustworthy faces (Baron et al., 2010).

In an effort to advance understanding of the neurological basis of trust, Krueger and Meyer-Lindenberg (2019) propose a neuropsychoeconomic model that argues that evaluations of trustworthiness are likely to involve the DMN given its importance for forming impressions of others attributes and intentions. Regions of the DMN are known to activate in tasks requiring participants to understand and interact with others, including perceiving and interpreting others emotion status, inferring others' beliefs and intentions and developing moral judgements of others' behaviours (Li et al., 2014). In line with this, we hypothesise:

> **Hypothesis 1.** Trust violation will be processed in social cognitive-related brain areas reflected by activation of the default mode network.

Growing evidence suggests that responses to trust violations differ according to the type of trustworthiness violated (Kim, 2018; Kim et al., 2009). In general, empirical evidence has demonstrated that integrity violations are particularly problematic (e.g. Kim et al., 2013). Several theoretical accounts have been offered to account for this. First, integrity is seen to be a stable characteristic, and individuals who act with low integrity in one situation are deemed likely to act in a similar way in the future (Tomlinson & Mayer, 2009). Clearly, this has implications for the repair of trust in ongoing workplace relationships. Perhaps more pertinent to the new relationships included in our study is the second explanation. When assessing information about integrity, individuals tend to place more weight on negative information, in direct contrast to the processing of ability information where positive information is given more weight (Kim et al., 2006). The trust violation literature is relatively consistent in the message that integrity violations are likely to be more damaging than ability violations; however, benevolence violations have received less attention.

Benevolence perceptions reflect a judgement that another party has one's best interests at heart, that they would be motivated to help or take risks to benefit the trustor (Mayer et al., 1995). Although benevolence violations have received limited empirical attention, trust scholars have argued that benevolence is more affectively or relationally toned and includes a judgement of a personal orientation towards the trustor, not solely the characteristics of the trustee (Colquitt et al., 2012). Furthermore, Tomlinson and Mayer (2009) argue that like integrity, benevolence violations are more likely to be viewed as controllable as issues with ability, in that trustees can choose to behave in a way that is benevolent if they wish. Accordingly, we hypothesise:

> **Hypothesis 2.** (a) Integrity violation and (b) benevolence violation will result in a stronger reaction than ability violation, indicated by greater activation in the default mode network in the brain.

METHOD

Sample

To partake in the study, participants needed to be over 18 years of age, right-handed, have no personal or family history of epilepsy or seizures and have no history of head injury or stroke.

Furthermore, participants needed to be non-smokers, and not taking psychoactive substances. These criteria were used to help us compare responses across participants in line with selection criteria used in previous EEG studies (e.g. Inzlicht & Gutsell, 2007; Ray & Cole, 1985).

One hundred forty-five people started the pre-survey, and 118 surveys were completed. Of these, 91 participants agreed to partake in the EEG study. Nine people did not arrive at the laboratory at their scheduled time, and one individual who was left-handed was excluded upon arrival. Thus, the final sample comprised 81 individuals (ability violation = 25; benevolence violation = 28; integrity violation = 28). During EEG data cleaning, a further 13 respondents were excluded (see below), resulting in a sample of 68 (ability violation = 20; benevolence violation = 24; integrity violation = 24). There were no differences in terms of age, gender, level of education or whether English was the first language between those who were retained and not. This sample size is similar to that of previous EEG studies with experimental manipulations (e.g. Harmon-Jones & Peterson, 2009; Inzlicht & Gutsell, 2007). However, as past commentaries identified that EEG studies have a tendency to have low power due to recognised difficulties in obtaining large samples (Szucs & Ioannidis, 2017), we additionally conducted Bayesian analysis to test the robustness of our findings. Participants had an average age of 33.66 years (SD = 12.37 years) and ranged in age from 18 to 62 years of age. Thirty-six individuals were female, and 32 were male. The majority of participants (82.4%) had completed at least some third-level education.

Procedure

Following ethical approval from the relevant university research ethics committees, participants were recruited to partake in the study via posters posted on various university notice boards, email lists of the university and via a research recruitment firm in Ireland. Participants were offered €20 multi-store gift card for their participation. Participants were first asked to complete a pre-survey, during which they were asked screening questions for right-handedness and were assessed for disqualifying injuries, illness or medication use. Once complete, they were given a date and time to present themselves at the lab, where they were met by the EEG technician.

Prior to beginning the EEG session, participants were asked to read a plain language statement and sign a consent form. They were informed that the purpose of the study was to evaluate brain activity during a test of cognition. As the technician connected the EEG to the participant, she aimed to have as little interaction between her and the participant as possible. A scripted experimental protocol was also prepared to standardise necessary interactions in so far as practically possible. The EEG technician then left the room, explaining to the participants that this was important for the experiment, but that she would be able to communicate with them via a one-way microphone to give them instructions.

Participants were randomly assigned to one of three trust violation scenarios using a random number generator in Excel (ability violation, benevolence violation or integrity violation). During the period of time that the experimenter was out of the room, the participants were presented with one of three pre-recorded verbal protocols of the EEG technician interacting via the one-way microphone. Each recording represented one type of trust violation (ability, integrity or benevolence).

Participants were instructed so that they were engaged in an eyes open baseline (EOB) task when the trust violation event occurred. Approximately 170 s into the EOB task, participants heard one of the three possible pre-recorded trust violation events, which were presented to

participants as though it were live. After the audio segment related to the trust violation event ended, there was a further 110 s until the EOB task ended. Each trust violation event lasted approximately 20 s (conversation start to finish). In the 20 s prior to the start of the conversation, there were a number of loud audible artefacts such as a door opening. This was constant for all participants and was included to heighten the realism of the recorded audio segment.

After the session was complete, the participant was fully debriefed as to the purpose of the experiment using a pre-prepared debriefing document and were free to ask any questions they had about the research. Finally, participants were presented with their multi-store gift voucher and thanked for their time.

Measures

Pre-survey

Propensity to trust was assessed using the scale by MacDonald et al. (1972) which comprised 10 items (e.g. 'I expect other people to be open and honest'; $\alpha = 0.783$) assessed on a 5-point Likert scale ranging from 1 = *strongly disagree* to 5 = *strongly agree*.

A number of demographic questions were also asked including age, gender, level of education and level of English proficiency.

During experiment

Trust violation was experimentally manipulated using pre-recorded audio of the EEG technician. Three types of trust violation were manipulated: ability, benevolence and integrity. During the EEG sessions, the participants heard the EEG technician being interrupted by a knock on the door of the room she was in, and overheard one of the following conversations, as though the technician had forgotten to turn off the microphone.

The conversations were as follows:

Knock on door (audio)
AN Other 'Hi, oh sorry, is that mic on?'
EEG Technician 'No, you're grand, it's off'.
AN Other 'How's it going?'

Ability trust violation condition:

EEG Technician 'Grand, I think I completely messed this one up, but sure...'
AN Other 'Sure, look, you can't get them all right. Anyhow, keep at it'.

Benevolence trust violation condition:

EEG Technician 'Grand, such a waste of their time but sure look, we get our data'.
AN Other 'Well, sure look, that's the most important thing isn't it'.

Integrity trust violation condition:

EEG Technician 'Yeah, grand, I don't think they realise what we are actually doing'.
AN Other 'Sure, that's the main thing isn't it'.

The language used in our experimental manipulations was designed to reflect the conceptual definitions of ability, benevolence and integrity from Mayer et al. (1995). Given the complexity of affective and cognitive reasoning processes that are likely to accompany a trust violation in this type of experimental situation, we decided to assess the definitional correspondence of our trust violations using two methods. First, 10 researchers with expertise in the area were asked to listen to each of the audio files and identify which type of trust they were designed to violate. A satisfactory level of agreement was established for each of the three types of trust violation. Second, an independent sample of 149 participants was recruited via Prolific. Participants were asked to rate the extent to which the experimental manipulation matched definitions of ability, benevolence and integrity using a 7-point scale: 1 = *statement is an extremely bad match to the definition above*; 7 = *statement is an extremely good match to the definition above*. The mean levels of definitional correspondence were m = 5.52, 5.32 and 4.90 for ability, benevolence and integrity respectively. This compares favourably with previous uses of this procedure (e.g. Baer et al., 2018; Hinkin & Tracey, 1999).

Brain activity was measured using electroencephalography (EEG). EEG is a non-invasive method of measuring electrical activity generated by the brain by placing electrodes on the scalp. Changes in the characteristics of this electrical activity provide a way to measure changes related to ongoing cognitive processes. Past research has demonstrated that frontal theta EEG activity can be seen as an EEG index of DMN activity (Scheeringa et al., 2008), and so, we used frontal theta EEG activity in the present research as an indicator of the neurological reaction to the trust violation. Frontal theta EEG activity correlates negatively with the DMN in a resting state (Scheeringa et al., 2008) and we expect that a trust violation would activate the DMN. More specifically, electrode site Fz was chosen for analysis as this is the electrode site where frontal theta is maximal (Inanaga, 1998), and most appropriate to our study of frontal theta and the DMN (Prestel et al., 2018; Scheeringa et al., 2008).

EEG acquisition/processing

EEG was recorded at 19 scalp locations (Fp1, Fp2, F3, F4, F7, F8, C3, C4, T3, T4, T5, T6, P3, P4, O1, O2, Fz, Cz, Pz) using a fixed electrode cap, arranged as per the International 10–20 system. A Deymed TruScan EEG system with a sampling rate of 256 Hz was used for data capture with an electrode at Fpz (as a ground electrode) and Fz (as a reference electrode). EEG was band-passed filtered between 1 and 20 Hz using a zero-phase finite impulse response (FIR) filter. Following this, all EEG recordings were converted to a common average reference (CAR) scheme. A software trigger was captured via the Deymed recording software (when the audio file started playing) to enable the later co-registration of the trust violation audio event and the EEG time periods.

Two continuous segments of EEG were extracted for each participant with respect to the start and the end of the trust violation event. Power spectral density (PSD) measures were calculated using Welch's method (Gramfort et al., 2013) for theta band (4–8 Hz) activity for electrode site Fz for the two time periods preceding and following the violation event. A time period of 80 s (ending 30 s before the violation event) was used to compute a baseline measure for each

participant and is hereafter referred to as the BVE (before violation event) time period. The BVE was selected as it did not contain any loud environmental noises as part of the recording. The time period following the violation event (lasting 80 s) was expected to exhibit power changes as a result of the violation event and is hereafter referred to as the AVE (after violation event) time period. A ratio was computed on the average power (for the theta band) by dividing AVE by BVE for each participant. The ratio measure for each participant was log-transformed due to the inherent non-normality of ratio data (Miskovic et al., 2011). This can be considered a theta log ratio (TLR) as it indicates an increase or decrease in the power of theta band activity following the violation event relative to the baseline period. As we outlined above, frontal theta activity correlates negatively with the default mode network in a resting state (Scheeringa et al., 2008), and so, higher negative values of the TLR indicate higher activation of the DMN.

In order to compute an averaged PSD for each of the BVE/AVE time periods on the EEG, 3-s epochs were extracted using a 1-s overlap. A peak-to-peak thresholding filtering scheme was used to reject epochs (in the time domain) that had a peak-to-peak amplitude greater than 90 μV. Visual inspection of the filtered epochs indicated gross artefacts such as eye blinks were no longer present. The TLR values generated for each participant were used as input to our ANOVA.

EEG data cleaning

We found the number of rejected EEG trials following the trust violation event was higher than that of the baseline period [$F(2,65) = 25.902$, $\eta 2 = .285$, $p < .001$], indicating there was an increase in physical and ocular movement following the event, which we would expect. EEG is sensitive to movement-induced artefacts and other sources of noise that can often be confounded with the experimental condition under manipulation; hence, we do not use these time periods as part of our EEG analysis. In this regard, we take a highly conservative approach in order not to use any data that may potentially contain artefacts that could produce erroneous results in our analysis. This is in line with current best practice (e.g. Cohen, 2014).

In order to effectively analyse the remaining data after epoch cleaning, we deemed a participant's data as usable if more than 12.5% (threshold) of their data were available for the pre-event and post-event time periods, that is, 12.5% of the 3-s epochs. We found increasing our filtering threshold beyond this point was not optimal as a large number of participants would need to be excluded from the analysis. In Figure 2, we show the impact of varying the subject rejection threshold for the 0- to 80-s period following the breach event across three different thresholds. It can be seen that increasing the threshold results in fewer subjects as part of the analysis and at higher thresholds, detection of significance of statistical effects rapidly degrades. Notably, however, the spatial pattern of statistical effects remains highly consistent across filtering thresholds, indicating that this approach did not change the nature of significant effects found.

Statistical analysis procedure

The TLR value for participants was used in an ANOVA analysis examining between-subject effects for condition. Post hoc analyses report Bonferroni corrected *p*-values for multiple comparisons unless specifically stated. We replicated our analysis using a Bayesian ANOVA to

assess the robustness of our findings and because it has a number of advantages over null hypothesis significance testing (NHST). In particular, supplementing our analysis in this way allows us to investigate our findings using a statistical method that is not based on the central limit theorem; thus, large samples are not required and conventional approaches to power are not applicable (van de Schoot & Depaoli, 2014). Analyses were conducted using SPSS Version 24 or JASP.

RESULTS

Preliminary analysis

Prior to testing our hypotheses, we conducted a number of preliminary tests. First, we conducted a series of between subjects ANOVAs to assess the success of the randomisation. There were no differences across the three conditions with regard to participants' gender, age, level of education or propensity to trust. Thus, our randomisation was successful. Means, standard deviations and correlations between study variables can be found in Table 1.

Hypothesis testing

In order to test our hypotheses, we first tested for differences in the TLR for electrode site Fz between condition types (between-subject effect). A significant between-subject effect was found for condition [$F(2,65) = 4.665$, $\eta 2 = .126$, $p = .013$]. In the post hoc analysis, we found the mean for the ability condition ($\text{Ability}_{\text{mean}} = .0125$, $SD = .160$, $n = 20$) was greater than the mean for the integrity condition ($\text{Integrity}_{\text{mean}} = -.149$, $SD = .248$, $n = 24$) where $p = .016$. We also found that the mean for the benevolence condition ($\text{Benevolence}_{\text{mean}} = -.0056$, $SD = .161$, $n = 24$) was greater than the mean for the integrity condition where $p = .022$. No significant differences were found between the means for the benevolence and ability conditions.

Due to past criticisms of neuroscientific research with regard to power, we replicated our statistical analyses using a Bayesian ANOVA. We used the default prior specified in JASP which uses a Cauchy distribution (see Rouder et al., 2012 for an overview of the development of default Bayes factors for ANOVA designs). Wagenmakers et al. (2018) state that the use of such

TABLE 1 Pearson correlations between study variables

	1.	2.	3.	4.
1. Fz_TLR	—			
2. Dummy condition (integrity)	−0.353**	—		
3. Dummy condition (ability)	0.199	−0.477***	—	
4. Age	−0.083	−0.035	0.103	—
5. Education	−0.186	0.160	−0.174	−0.208

Note: Fz_TLR represents a theta log ratio of the before violation event time period and the after violation event time period at the electrode site Fz.
*$p < .05$.**$p < .01$.***$p < .001$.

default priors yield results that are broadly consistent with those that would be obtained with a more subjective analysis. Results indicated that the model with the three experimental conditions predicted the data almost four times ($BF_{10} = 3.973$) as well as the null model and the model average posterior R^2 equals .080, indicating that 8% of the variance in the TLR score in the Fz channel is due to the condition. The percentage error associated with the estimation of the Bayes factor is quite low (0.017), suggesting that the estimate is likely to be accurate. The model averaged posterior distribution demonstrates that it is the integrity condition that is associated with the largest differences (see Table 2 and Figure 1). Bayesian post hoc analysis indicated that the data is 3.372 times more likely under the integrity violation condition compared with the ability violation condition and 2.637 times more likely compared with the benevolence violation condition (Table 3).

TABLE 2 Model averaged posterior summary for Bayes ANOVA (1 = *ability trust violation*, 2 = *benevolence trust violation*, 3 = *integrity trust violation*)

				95% credible interval	
Variable	**Level**	**Mean**	**SD**	**Lower**	**Upper**
Intercept		−0.048	0.022	−0.093	−0.002
Cond.	1	0.049	0.033	−0.016	0.115
	2	0.035	0.031	−0.027	0.098
	3	−0.084	0.033	−0.152	−0.020

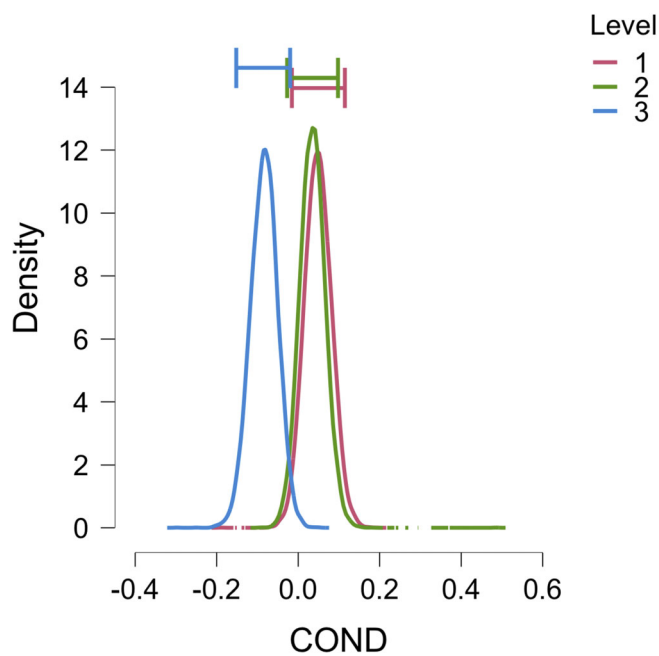

FIGURE 1 Model averaged posterior distributions Bayes ANOVA comparing the trust violation conditions (1 = *ability*, 2 = *benevolence*, 3 = *integrity*)

TABLE 3 Bayesian post hoc comparisons of experimental condition (1 = *ability trust violation*, 2 = *benevolence trust violation*, 3 = *integrity trust violation*)

		Prior odds	Posterior odds	$BF_{10, U}$	Error %
1	2	0.587	0.185	0.316	0.018
	3	0.587	1.981	3.372	0.001
2	3	0.587	1.549	2.637	8.524e-4

Notes: The posterior odds have been corrected for multiple testing by fixing to 0.5 the prior probability that the null hypothesis holds across all comparisons (Westfall et al., 1997). Individual comparisons are based on the default *t*-test with a Cauchy (0, r = 1/sqrt[K3] (2)) prior. The 'U' in the Bayes factor denotes that it is uncorrected.

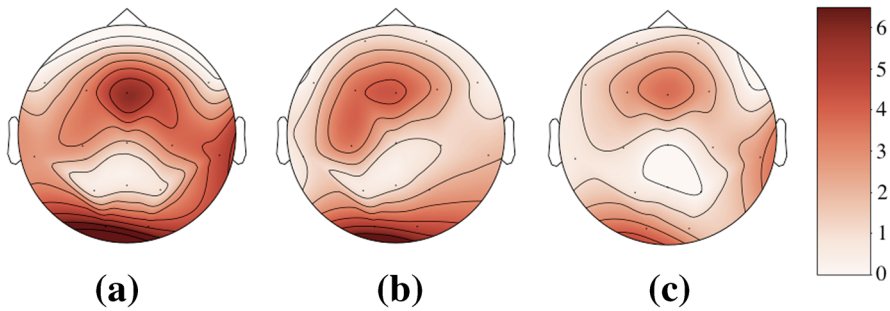

FIGURE 2 Topographical plot of *F*-values from one-way ANOVAs (applied per electrode location) showing the spatial distribution of significant effects when using baselined (AVE-BVE) log-transformed theta band power for the three trust violation conditions. In (a) (leftmost scalp plot), we use a threshold of 12.5% (where $N_{Ability} = 20$, $N_{Benevolence} = 24$ and $N_{Integrity} = 24$), in (b) (middle scalp plot), we use a threshold of 25% (where $N_{Ability} = 15$, $N_{Benevolence} = 22$ and $N_{Integrity} = 19$), and in (c) (right scalp plot), we use a threshold of 50% (where $N_{Ability} = 11$, $N_{Benevolence} = 14$ and $N_{Integrity} = 11$) for comparison.

In Figure 2, we visualise the spatial topography of statistical effects (*F* values) for condition type for theta band activity (via a one-way ANOVA). As expected we observe statistical effects related to frontal theta surrounding electrode site Fz; however, ostensible effects appear to be present at other scalp locations notably over occipital electrodes. This may have indicated that the effect was not restricted to the DMN/frontal theta exclusively, so we further investigated potential effects at other channels. Thus, we extended our analysis using a repeated-measures ANOVA to include all electrode locations (17 channels) in the EEG (as a within-subject factor). We find the between-subject effect remains similar [$F(2,65) = 4.736$, $\eta 2 = .127$, $p = .012$]; however, no significant effect is observed for electrode location. We find the same pattern of statistical effects remain (and overall means for condition type remain consistently ranked) when including all electrodes in this analysis (i.e. $Ability_{mean} = .015$, $SD = .600$, $n = 20 > Benevolence_{mean} = -.006$, $SD = .648$, $n = 24 > Integrity_{mean} = -.099$, $SD = .648$, $n = 24$). As no significant effects are found for electrode, we do not investigate these other spatially distinct patterns of theta activity any further. Notwithstanding, we observed the expected condition-related effects at electrode site Fz related to the DMN and frontal theta. To further investigate the robustness of our findings, we replicated our analysis for delta (1–4 Hz), alpha (8–12 Hz) and beta (14–20 Hz) bands and found no significant effects for condition. Thus, we can be confident in our conclusion that theta (4–7 Hz) band activity is responsible for the observed effects.

In summary, we find that the integrity condition showed the greatest decrease in theta activity, in turn reflecting a greater increase of activity in DMN compared to other conditions (see Figures 1 and 3). We confirmed this result by testing our data using a conventional one-way ANOVA and with a Bayesian ANOVA. We can confirm the observed effects are in line with our first hypothesis that *trust violation will be processed in social cognitive-related brain areas reflected by activation of the default mode network*. Our results also partially support our second hypothesis that *integrity violation and benevolence violation will result in a stronger reaction than ability violation (indicated by greater activation in the default mode network in the brain)*; however, we find this is only significant for the integrity violation and not the benevolence violation.

DISCUSSION

Our findings demonstrate the involvement of the DMN in the processing of trust violations and provide evidence that the degree of activation of the DMN differs according to the type of trustworthiness that is violated. This result lends important support for organisational trust theorists suggesting that different types of violation may be processed differently or to varying degrees. A major contribution of our research is that the involvement of the DMN suggests that this difference is related specifically to how individuals understand and interpret the emotions and intentions of others. Whereas interpretation of the intentions of others has been a feature of previous scholarly work (e.g. Tomlinson & Mayer, 2009), survey and experimental vignette studies can make it difficult to uncover the underlying processes involved. Our findings clearly underscore

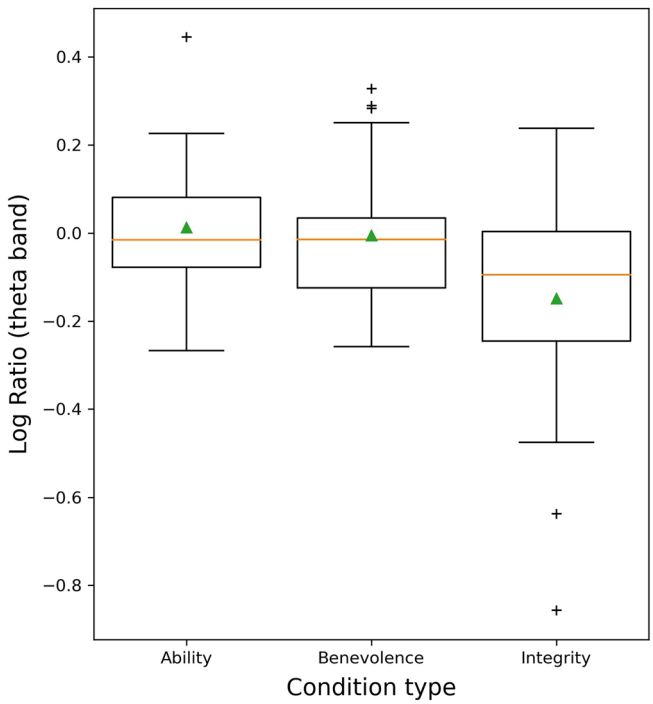

FIGURE 3 Boxplot showing theta log ratio (TLR) for the three trust violation conditions types for electrode site Fz. Means are indicated by a triangle.

the importance of social cognitions, suggesting that understanding of emotion and intentions is the crucial process influencing the experience of trust violation.

Integrity violations generated the greatest response in DMN activation, in line with the self-report reactions from previous studies. It has been proposed that the DMN plays an adaptive role in helping individuals to navigate social interactions (Buckner et al., 2008), and integrity trust violations signal that an individual is not honest in their interactions with others. Practically speaking, such integrity violations lead to the trustor being less likely to take risks with such an individual (Mayer et al., 1995) and to verify any information that may originate from this individual (Lewicki et al., 1998).

Our findings also provide some initial indication of how benevolence violations might be processed. We had predicted that benevolence violations, like integrity, would result in a stronger reaction than ability violations. Although the means for each condition in our study are in line with this, the difference between the benevolence and ability conditions was not significant, and we could not support the hypothesis. One potential explanation for this finding is the context of the experiment where the benevolence violation of the experimenter being unconcerned about wasting the participant's time might be considered to have a less serious implication for the trustor. In contrast, the perception of an integrity violation could be considered to have much more serious implications, thus triggering stronger responses. This would fit with threshold models of trust violation, where not every deviation from expectations triggers a violation perception (Jones & George, 1998). Recent theory also suggests that benevolence perceptions are likely to be more context dependent and relational than integrity (Moore et al., 2019), and so the relatively impersonal context of the experimenter–participant dyad is likely to influence their interpretation. Future research might consider obtaining self-report ratings of the arousal and valence of different types of trust violation in order to check for differences across conditions.

Practical implications

Beyond their theoretical importance, our findings also carry important implications for organisations and their employees. Trust violations are pervasive in the workplace (Lewicki & Brinsfield, 2017). Although organisations cannot always prevent these violations, particularly those that occur at an interpersonal level, they can take action to manage trust violations as they occur (Gillespie & Dietz, 2009). Our research provides insight into the types of violations that might be most damaging to trust in workplace relationships. Armed with this understanding, organisations and managers can play a proactive role in helping employees through the sensemaking process that surrounds potential trust violations (Gustafsson et al., 2020). In general terms, our study suggests that sensemaking and post-violation communication efforts that highlight benevolence and integrity in the transgressor might help frame trust violations in a less damaging manner. Further research might investigate how this process is influenced through repeated violations by the same work colleague or within the same work environment or how it is influenced by the attributes of the transgressor (Cowen & Montgomery, 2020).

Limitations and future research

Although our study demonstrates robust findings regarding the link between trust violation and the DMN, there are limitations that might be addressed by future research. First, EEG is

just one option for measuring brain activity. We chose EEG as it offers better temporal resolution and an opportunity for a more natural environment for research participants than alternative approaches (Waldman et al., 2017). However, one of the key limitations of the EEG method is spatial resolution. Future research using fMRI would offer us an increased ability to identify brain networks and regions involved in processing trust violation and may provide an opportunity to assess some of the deeper brain structures that are often associated with emotional experience.

Second, our method did not allow for a between subjects control group. We used a within-subject comparison by comparing the BVE time period the AVE time period. Thus, we were able to calculate the differences within person and compare these across conditions. As there needed to be an event in order to create this BVE–AVE ratio, it did not make sense to include a traditional control group.

Finally, our methodology offered us limited scope to ascertain differences in the length of time of activation in reaction to trust violation. It may be that certain violations involve a shorter or longer response than others and this may be related to their affective versus cognitive content. We encourage future researchers to investigate this possibility in more detail. Finally, the deception in our experimental study made it impractical to gather self-report data on participants' pre-violation perceptions of the experimenter without alerting them to the nature of the study and undermining the experimental manipulation. Scholars (e.g. Lewicki & Brinsfield, 2017) have called for studies of trust violation and repair to include more assessment of pre-existing levels of trust, and we reiterate this call.

Conclusion

The regularity of trust violations in our personal and professional lives (Lewicki & Brinsfield, 2017) means research into the processes underlying trust violation provide an important opportunity to understand this process further in order to avoid issues with cooperation or to facilitate relationship repair. Although there has been some limited prior EEG research on trust, this is the first study to examine different types of trust violation and their corresponding impacts. Taken together, our findings, combined with previous experimental vignettes, clearly indicate that different types of violation evoke different responses in relation to activation of the DMN and that integrity violations have the potential to create the greatest response.

ACKNOWLEDGEMENTS

Open access funding provided by IReL.

CONFLICT OF INTEREST

The authors have no conflict of interest to declare.

ETHICS STATEMENT

In conducting this research, we have complied with the ethical guidelines regarding research with human participants from the Psychological Society of Ireland. Our study was reviewed and approved by the DCU Research Ethics Committee.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Lisa van der Werff https://orcid.org/0000-0003-4529-4690
Deirdre O'Shea https://orcid.org/0000-0001-9107-1434
Graham Healy https://orcid.org/0000-0001-6429-6339
Finian Buckley https://orcid.org/0000-0003-2651-6868
Colette Real https://orcid.org/0000-0002-9093-5293
Theo Lynn https://orcid.org/0000-0001-9284-7580

REFERENCES

Adolphs, R. (2001). The neurobiology of social cognition. *Current Opinion in Neurobiology*, *11*, 231–239. https://doi.org/10.1016/S0959-4388(00)00202-6

Adolphs, R. (2009). The social brain: Neural basis of social knowledge. *Annual Review of Psychology.*, *60*, 693–716. https://doi.org/10.1146/annurev.psych.60.110707.163514

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience*, *7*(4), 268–277. https://doi.org/10.1038/nrn1884

Baer, M., & Colquitt, J. A. (2018). Moving toward a more comprehensive consideration of the antecedents of trust. In R. H. Searle, A. M. Neinaber, & S. B. Sitkin (Eds.), *Routledge companion to trust* (pp. 163–182). Routledge. https://doi.org/10.4324/9781315745572-12

Baer, M. D., van der Werff, L., Colquitt, J. A., Rodell, J. B., Zipay, K. P., & Buckley, F. (2018). Trusting the "look and feel": Situational normality, situational aesthetics, and the perceived trustworthiness of organizations. *Academy of Management Journal*, *61*(5), 1718–1740. https://doi.org/10.5465/amj.2016.0248

Baron, S. G., Gobbini, M. I., Engell, A. D., & Todorov, A. (2010). Amygdala and dorsomedial prefrontal cortex responses to appearance-based and behavior-based person impressions. *Social Cognitive and Affective Neuroscience*, *6*, 572–581. https://doi.org/10.1093/scan/nsq086

Bressler, S. L., & Menon, V. (2010). Large-scale brain networks in cognition: Emerging methods and principles. *Trends in Cognitive Sciences*, *14*, 277–290. https://doi.org/10.1016/j.tics.2010.04.004

Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The brain's default network: Anatomy, function, and relevance to disease. *Annals of the new York Academy of Sciences*, *1124*(1), 1–38. https://doi.org/10.1196/annals.1440.011

Cohen, M. X. (2014). *Analyzing neural time series data: Theory and practice*. MIT Press.

Colquitt, J. A., LePine, J. A., Piccolo, R. F., Zapata, C. P., & Rich, B. L. (2012). Explaining the justice–performance relationship: Trust as exchange deepener or trust as uncertainty reducer? *Journal of Applied Psychology*, *97*(1), 1–15. https://doi.org/10.1037/a0025208

Cowen, A. P., & Montgomery, N. V. (2020). To be or not to be sorry? How CEO gender impacts the effectiveness of organizational apologies. *Journal of Applied Psychology*, *105*(2), 196–208. https://doi.org/10.1037/apl0000430

Dimoka, A. (2010). What does the brain tell us about trust and distrust? Evidence from a functional neuroimaging study. *MIS Quarterly*, *34*, 373–396. https://doi.org/10.2307/20721433

Doeze Jager, S. B., Born, M. P., & van der Molen, H. T. (2022). The relationship between organizational trust, resistance to change and adaptive and proactive employees' agility in an unplanned and planned change context. *Applied Psychology*, *71*(2), 436–460. https://doi.org/10.1111/apps.12327

Fulmer, C. A., & Gelfand, M. J. (2012). At what level (and in whom) we trust: Trust across multiple organizational levels. *Journal of Management*, *38*(4), 1167–1230. https://doi.org/10.1177/0149206312439327

Gillespie, N., & Dietz, G. (2009). Trust repair after an organization-level failure. *Academy of Management Review*, *34*(1), 127–145. https://doi.org/10.5465/amr.2009.35713319

Gramfort, A., Luessi, M., Larson, E., Engemann, D., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., & Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, *7*, 267. https://doi.org/10.3389/fnins.2013.00267

Gustafsson, S., Gillespie, N. A., Searle, R., Hope Hailey, V., & Dietz, G. (2020). Preserving organizational trust during disruption. *Organization Studies*, *42*, 1409–1433. https://doi.org/10.1177/0170840620912705

Haesevoets, T., De Cremer, D., Van Hiel, A., & Van Overwalle, F. (2018). Understanding the positive effect of financial compensation on trust after norm violations: Evidence from fMRI in favor of forgiveness. *Journal of Applied Psychology*, *103*, 578–590. https://doi.org/10.1037/apl0000271

Harmon-Jones, E., & Peterson, C. (2009). Supine body position reduces neural response to anger evocation. *Psychological Science*, *20*(10), 1209–1210. https://doi.org/10.1111/j.1467-9280.2009.02416.x

Haselhuhn, M. P., Kennedy, J. A., Kray, L. J., Van Zant, A. B., & Schweitzer, M. E. (2015). Gender differences in trust dynamics: Women trust more than men following a trust violation. *Journal of Experimental Social Psychology*, *56*, 104–109. https://doi.org/10.1016/j.jesp.2014.09.007

Hinkin, T. R., & Tracey, J. B. (1999). An analysis of variance approach to content validation. *Organizational Research Methods*, *2*(2), 175–186. https://doi.org/10.1177/109442819922004

Inanaga, K. (1998). Frontal midline theta rhythm and mental activity. *Psychiatry and Clinical Neurosciences*, *52*(6), 555–566. https://doi.org/10.1111/j.1440-1819.1998.tb02700.x

Inzlicht, M., & Gutsell, J. N. (2007). Running on empty: Neural signals for self-control failure. *Psychological Science*, *18*(11), 933–937. https://doi.org/10.1111/j.1467-9280.2007.02004.x

Jones, G. R., & George, J. M. (1998). The experience and evolution of trust: Implications for cooperation and teamwork. *Academy of Management Review*, *23*(3), 531–546. https://doi.org/10.5465/amr.1998.926625

Kim, P. H. (2018). An interactive perspective on trust repair. In R. H. Searle, A. M. Neinaber, & S. B. Sitkin (Eds.), *Routledge companion to trust* (pp. 269–283). Routledge. https://doi.org/10.4324/9781315745572-19

Kim, P. H., Cooper, C. D., Dirks, K. T., & Ferrin, D. L. (2013). Repairing trust with individuals vs. groups. *Organizational Behavior and Human Decision Processes*, *120*(1), 1–14. https://doi.org/10.1016/j.obhdp.2012.08.004

Kim, P. H., Dirks, K. T., & Cooper, C. D. (2009). The repair of trust: A dynamic bilateral perspective and multilevel conceptualization. *The Academy of Management Review*, *34*(3), 401–422. https://doi.org/10.5465/AMR.2009.40631887

Kim, P. H., Dirks, K. T., Cooper, C. D., & Ferrin, D. L. (2006). When more blame is better than less: The implications of internal vs. external attributions for the repair of trust after a competence-vs. integrity-based trust violation. *Organizational Behavior and Human Decision Processes*, *99*(1), 49–65. https://doi.org/10.1016/j.obhdp.2005.07.002

Kim, P. H., Ferrin, D. L., Cooper, C. D., & Dirks, K. T. (2004). Removing the shadow of suspicion: The effects of apology versus denial for repairing competence- versus integrity-based trust violations. *Journal of Applied Psychology*, *89*(1), 104–118. https://doi.org/10.1037/0021-9010.89.1.104

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to know you: Reputation and trust in a two-person economic exchange. *Science*, *308*(5718), 78–83. https://doi.org/10.1126/science.1108062

Krueger, F., & Meyer-Lindenberg, A. (2019). Toward a model of interpersonal trust drawn from neuroscience, psychology, and economics. *Trends in Neurosciences*, *42*(2), 92–101. https://doi.org/10.1016/j.tins.2018.10.004

Legood, A., van der Werff, L., Lee, A., den Hartog, D., & van Knippenberg, D. (2022). A critical review of the conceptualization, operationalization, and empirical literature on cognition-based and affect-based trust. *Journal of Management Studies*. Online first publication. https://doi.org/10.1111/joms.12811

Lewicki, R. J., & Brinsfield, C. (2017). Trust repair. *Annual Review of Organizational Psychology and Organizational Behavior*, *4*, 287–313. https://doi.org/10.1146/annurev-orgpsych-032516-113147

Lewicki, R. J., McAllister, D. J., & Bies, R. J. (1998). Trust and distrust: New relationships and realities. *Academy of Management Review*, *23*(3), 438–458. https://doi.org/10.2307/259288

Li, W., Mai, X., & Liu, C. (2014). The default mode network and social understanding of others: What do brain connectivity studies tell us. *Frontiers in Human Neuroscience*, *8*, 74. https://doi.org/10.3389/fnhum.2014.00074

MacDonald, A. P. Jr., Kessel, V. S., & Fuller, J. B. (1972). Self-disclosure and two kinds of trust. *Psychological Reports*, *30*(1), 143–148. https://doi.org/10.2466/pr0.1972.30.1.143

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*(3), 709–734. https://doi.org/10.2307/258792

Miskovic, V., Moscovitch, D. A., Santesso, D. L., McCabe, R. E., Antony, M. M., & Schmidt, L. A. (2011). Changes in EEG cross-frequency coupling during cognitive behavioral therapy for social anxiety disorder. *Psychological Science*, *22*(4), 507–516. https://doi.org/10.1177/0956797611400914

Moore, A. K., Munguia Gomez, D. M., & Levine, E. E. (2019). Everyday dilemmas: New directions on the judgment and resolution of benevolence–integrity dilemmas. *Social and Personality Psychology Compass*, *13*, e12472. https://doi.org/10.1111/spc3.12472

Prestel, M., Steinfath, T. P., Tremmel, M., Stark, R., & Ott, U. (2018). fMRI BOLD correlates of EEG independent components: Spatial correspondence with the default mode network. *Frontiers in Human Neuroscience*, *12*, 478. https://doi.org/10.3389/fnhum.2018.00478

Ray, W. J., & Cole, H. W. (1985). EEG alpha activity reflects attentional demands, and beta activity reflects emotional and cognitive processes. *Science*, *228*(4700), 750–752. https://doi.org/10.1126/science.3992243

Riedl, R., & Javor, A. (2012). The biology of trust. *Journal of Neuroscience, Psychology, and Economics*, *5*, 63–91. https://doi.org/10.1037/a0026318

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356–374. https://doi.org/10.1016/j.jmp.2012.08.001

Scheeringa, R., Bastiaansen, M. C., Petersson, K. M., Oostenveld, R., Norris, D. G., & Hagoort, P. (2008). Frontal theta EEG activity correlates negatively with the default mode network in resting state. *International Journal of Psychophysiology*, *67*(3), 242–251. https://doi.org/10.1016/j.ijpsycho.2007.05.017

Szucs, D., & Ioannidis, J. P. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology*, *15*, e2000797. https://doi.org/10.1371/journal.pbio.2000797

Tomlinson, E. C., & Mayer, R. C. (2009). The role of causal attribution dimensions in trust repair. *Academy of Management Review*, *34*(1), 85–104. https://doi.org/10.5465/amr.2009.35713291

van de Schoot, R., & Depaoli, S. (2014). Bayesian analyses: Where to start and what to report. *European Health Psychologist*, *2*(1), 75–84.

van Knippenberg, D. (2018). Reconsidering affect-based trust: A new research agenda. In R. H. Searle, A.-M. I. Nienaber, & S. B. Sitkin (Eds.), *The Routledge companion to trust* (pp. 3–13). Routledge. https://doi.org/10.4324/9781315745572-2

Wagenmakers, E. J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Selker, R., Gronau, Q. F., Dropmann, D., Boutin, B., Meerhoff, F., Knight, P., Raj, A., van Kesteren, E. J., van Doorn, J., Šmíra, M., Epskamp, S., Etz, A., Matzke, D., ... Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, *25*, 58–76. https://doi.org/10.3758/s13423-017-1323-7

Waldman, D. A., Ward, M. K., & Becker, W. J. (2017). Neuroscience in organizational behavior. *Annual Review of Organizational Psychology and Organizational Behavior*, *4*(1), 425–444. https://doi.org/10.1146/annurev-orgpsych-032516-113316

Westfall, P. H., Johnson, W. O., & Utts, J. M. (1997). A Bayesian perspective on the Bonferroni adjustment. *Biometrika*, *84*(2), 419-427. https://doi.org/10.1093/biomet/84.2.419

Xiang, T., Ray, D., Lohrenz, T., Dayan, P., & Montague, P. R. (2012). Computational phenotyping of two-person interactions reveals differential neural response to depth-of-thought. *PLoS Computational Biology*, *8*, e1002841. https://doi.org/10.1371/journal.pcbi.1002841

How to cite this article: van der Werff, L., O'Shea, D., Healy, G., Buckley, F., Real, C., Keane, M., & Lynn, T. (2022). The neuroscience of trust violation: Differential activation of the default mode network in ability, benevolence and integrity breaches. *Applied Psychology*, 1–17. https://doi.org/10.1111/apps.12437