COGNITION AND EMOTION, 2018
VOL. 32, NO. 3, 480–493
https://doi.org/10.1080/02699931.2017.1316706

Routledge
Taylor & Francis Group

Rule-violations sensitise towards negative and authority-related stimuli

Robert Wirth, Anna Foerster, Hannah Rendel, Wilfried Kunde and Roland Pfister

Department of Psychology, Julius-Maximilians-University of Würzburg, Röntgenring 11, 97070 Würzburg, Germany

ABSTRACT

Rule violations have usually been studied from a third-person perspective, identifying situational factors that render violations more or less likely. A first-person perspective of the agent that actively violates the rules, on the other hand, is only just beginning to emerge. Here we show that committing a rule violation sensitises towards subsequent negative stimuli as well as subsequent authority-related stimuli. In a Prime-Probe design, we used an instructed rule-violation task as the Prime and a word categorisation task as the Probe. Also, we employed a control condition that used a rule inversion task as the Prime (instead of rule violations). Probe targets were categorised faster after a violation relative to after a rule-based response if they related to either, negative valence or authority. Inversions, however, primed only negative stimuli and did not accelerate the categorisation of authority-related stimuli. A heightened sensitivity towards authority-related targets thus seems to be specific to rule violations. A control experiment showed that these effects cannot be explained in terms of semantic priming. Therefore, we propose that rule violations necessarily activate authority-related representations that make rule violations qualitatively different from simple rule inversions.

ARTICLE HISTORY
Received 29 September 2016
Revised 29 March 2017
Accepted 30 March 2017

KEYWORDS
Rule violation; non-conformity; conflict; aversive signal; authority

How to approach rule violations

A pedestrian faces a red traffic light. What does he do? The normative response would be to stop and wait for the signal to turn green, and this course of action will be chosen by many people. A considerable share will respond differently, however, by crossing without further ado. Behaviour such as crossing at a red light resembles a violation of commonly accepted rules, and previous research has taken pains to develop models that allow for predicting whether such violations will occur in a given situation. This is not only true for pedestrians and red lights (Dommes, Granié, Cloutier, Coquelet, & Huguenin-Richard, 2015; King, Soole, & Ghafourian, 2009; Rosenbloom, 2009), but also for numerous other situations for which rule-violation behaviour has been identified as a potential risk factor (Berry, Ones, & Sackett, 2007; Parker, Reason, Manstead, & Stradling, 1995; Phipps et al., 2008).

Studies in this framework typically asked the question whether it is possible to predict rule violations in order to prevent them from happening. Answering this question requires a thorough analysis of observational data in which the occurrence of a rule violation is the critical measure (Phipps et al., 2008; Reason, 1990; Yap, Wazlawek, Lucas, Cuddy, & Carney, 2013). We label this approach as the "third-person approach" in the following. The third-person approach, however, does not allow for a precise understanding of the cognitive and affective processes involved for the agent who violates a rule. Still, first studies have begun to uncover the peculiarities of rule-breaking for the violating agent, and we label this emerging perspective as the "first-person approach" (Pfister, Wirth, Schwarz, Steinhauser, & Kunde, 2016; Wirth, Pfister, Foerster, Huestegge, & Kunde, 2016).

Studying the cognitive processes involved in rule-violation behaviour from a first-person perspective requires experimental paradigms in which a behaviour can be clearly identified as a deliberate rule violation rather than an unintended action slip or mistake. In a set of experiments, we therefore instructed

CONTACT Robert Wirth robert.wirth@uni-wuerzburg.de Department of Psychology, Julius-Maximilians-University of Würzburg, Röntgenring 11, 97070 Würzburg, Germany

Supplemental data for this article can be accessed at https://doi.org/10.1080/02699931.2017.1316706.

participants to respond to stimuli on the screen according to a fixed mapping rule: For one stimulus, they had to reach a left target area with the computer mouse, and for the other stimulus, they had to reach a right target area. Crucially, participants were given the opportunity to choose freely whether or not they intended to follow the mapping rule and indicate it via button press before the start of each trial (Pfister, Wirth, Schwarz, Steinhauser et al., 2016). Results showed that committing such a violation incurred cognitive conflict, as indexed by prolonged initiation times for violation responses and a reliable attraction of the corresponding mouse movement trajectories to the rule-based response option. Further, rule violations have been shown to be accompanied by a specific electrophysiological signature (Pfister, Wirth, Schwarz, Foerster et al., 2016).

However, violations can be defined on a wide spectrum (Reason, 1990), ranging from complex moral violations on the one end to simple instructions about what (not) to do on the other end (Pfister, 2013). A common taxonomy therefore distinguishes at least three types of rule violations: optimizing violations, routine violations, and necessary violations (Reason, 1990, 1995). Optimizing violations comprise behaviours such as the described pedestrian who crosses the road against a red traffic light. Routine violations comprise, for instance, sustained and systematic violations of safety procedures at the workplace. Necessary violations finally are defined as violations in which the agent does not choose to violate by him- or herself; rather, situational factors prompt the violation and the agent deliberately performs this action. These "necessary violations" are more common than one would think: Imagine, for example, a taxi driver whose passenger happens to be a policeman. The policeman shouts at the taxi driver to get him to a crime scene as fast as possible. The taxi driver, probably, will drive his car faster than allowed, violating traffic rules by an external prompt. But even when violations are prompted externally, violations are more difficult to execute and show a signed influence of the original rule (Wirth, Pfister, Foerster et al., 2016). Obviously, committing a rule violation comes with cognitive burdens, and it does so irrespective of whether violations are prompted internally or externally. The taxi driver, even though he now has an allowance to drive as fast as he wants, would still be literally thwarted by his automatic attempt to follow the traffic rules.

In experiments aiming for the cognitive mechanisms underlying rule violations, instructed violations provide a better experimental control than having participants choose freely whether to conform or violate. In the following experiments, participants were therefore confronted with an arbitrary mapping rule, and violating these rules did not have any consequences for the participants, and they did so in a non-social setting. This deliberate design choice allowed us to isolate the cognitive processes that arise purely due to the fact that an agent deliberately behaves counter to a rule – without potential confounding factors such as social influences, prior experience with specific types of non-conformity, morals, and expectations of punishment. Further, participants were instructed whether to follow or break a rule (akin to necessary violations as discussed above) to control for the ratio of both response options. A systematic comparison of participants who could choose freely whether to follow or break a rule to participants who were instructed what to do showed that the behavioural signature of non-conformity was independent of whether participants chose freely or whether they were instructed (Pfister, Wirth, Schwarz, Steinhauser et al., 2016).

What sets rule violations apart?

Even though the findings outlined above demonstrate that rule violations pose a considerable effort for the agent who intends to commit a violation, they do not allow for assessing the source of cognitive conflict that rule violations incur. That is: Why is there stronger cognitive conflict for rule violations compared to simple rule-based responses?

One candidate mechanism seems to be the difficulty to process negations. Whereas rule-based responses can be readily encoded (as "rule-based"), the human cognitive system tends to represent negated concepts in their non-negated form, accompanied by a negation tag that has to be resolved when negations are accessed (as "not" + "rule-based"; Fillenbaum, 1966; Gilbert, 1991; Strack & Deutsch, 2004; Wegner, Coulton, & Wenzlaff, 1985). Assuming that rule-violation behaviour is indeed represented as the corresponding rule-based behaviour accompanied by a negation tag, such a tag would have to be resolved during each instance of a rule violation. Indeed, negation processing has been shown to partially account for the effects of rule violation (Wirth, Pfister, Foerster et al., 2016). The mechanics of planning and executing a violation seem to follow a two-step activation model, where first the original rule is activated automatically and then has to be manipulated (in this case, negated) to

overcome the default, rule-based response tendency. However, this negation process alone did not explain the effects of rule violations entirely. When participants were asked to either violate or invert a two-choice mapping rule (which results in identical responses for both framings), the costs for rule violations were larger than the costs for rule inversions (Wirth, Pfister, Foerster et al., 2016). Even though both framings required the negation of the original rule, there seems to be something more difficult with violations than with more neutrally labelled, but ultimately identical actions.

So, what is it then that renders rule violations special? Here we propose that rule violations may differ from normal, rule-based responding in terms of evaluation and appraisal processes that occur automatically during or after response execution. Such evaluative processes have recently been documented for committing unintended errors (Aarts, De Houwer, & Pourtois, 2012, 2013; Lindström, Mattsson-Mårn, Golkar, & Olsson, 2013). For instance, when participants were asked to classify positive or negative target words after either correct responses or errors, negative words were classified more quickly after errors than after correct responses (Aarts et al., 2012). This bias suggests an automatic emotional reaction driven by the appraisal of own actions (see also Rabbitt & Rodgers, 1977).

Automatic emotional responses to rule violations seem likely in the light of experiments that showed cognitive conflict to cause emotional responses, though two opposing predictions can be derived from the literature. For one, conflicting situations in general seem to be linked to a negative emotional component, that is, conflicts appear to be aversive signals (Botvinick, 2007; Fritz & Dreisbach, 2013, 2015; Wirth, Pfister, & Kunde, 2016). For example, participants were first confronted with a congruent or an incongruent Stroop target, afterwards, word or picture targets had to be categorised as positive or negative. Positive targets were categorised faster when preceded by a congruent Stroop word as compared to incongruent Stroop words, and negative targets were categorised faster when preceded by an incongruent Stroop word as compared to congruent Stroop words (Dreisbach & Fischer, 2012). Based on this finding, the authors argued that the conflict in Stroop words is coded as an aversive signal, therefore incongruent Stroop words can sensitise toward negative targets in the subsequent task. Assuming that rule violations come with cognitive conflict (as do incongruent Stroop words), we would predict violations to sensitise to negative events. At the same time, however, the successful resolution of cognitive conflict has been demonstrated to represent a reward signal (Schouppe et al., 2015). Successfully overcoming a rule-based response tendency, that is, successfully committing a rule violation, would thus predict violations to sensitise to positive events instead.

At the same time, we set out to explore a further possible evaluative process triggered by rule violations. Rules and violations are mostly associated with the concept of authority. Authoritarian figures are the ones that make and enforce the rules in our daily lives (e.g. parents during childhood and adolescence, children and adolescents during parenthood, superordinates at the workplace, or officials such as police officers). Especially for young children, rules of authoritarian figures are mainly obeyed to avoid punishment (Kohlberg, 1963; Piaget, 1932). And even the behaviour of many adults is described as "orientation towards authority, fixed rules, […], [and] showing respect for authority" (description of the conventional, adult level in Kohlberg & Hersh, 1977, p. 55). The concept of authority seems to be strongly linked to rules and rule violations. Negations (as for example studied with rule inversions), arguably, are less closely associated with authority, especially those that are usually studied in experimental settings. Therefore, we additionally tested whether committing a rule violation modulates the categorisation of not only valent stimuli, but also of authority-related targets. Faster responses to authority-related words after a preceding rule violation as compared to preceding rule-based responses would speak for the idea that violations include authority-related processes. For inversions, however, such a modulation seems less likely.

In Experiment 1, we therefore tested how committing rule violations and rule inversions, respectively, modulates the categorisation of subsequent valent and authority-related stimuli. Finally, Experiment 2 tested whether the effects of Experiment 1 are simply driven by semantic priming due to the framing of the instructions, or whether they are tied to having executed the violation response.

Experiment 1

Introduction

In Experiment 1, we probed for the hypothesised affective and authority-related components of rule violations and compared them to those possibly elicited

by simple rule inversions (see Wirth, Pfister, Foerster, et al., 2016). To do so, we introduced a simple and arbitrary two-choice mapping rule with two stimuli to two response keys. This mapping had to be followed in most of the trials, and had to be violated in a fraction of trials by half of the participants, while the other half of participants received instructions to invert the rule. With this semantic variation, all participants had to apply the same responses in the same frequency, just the labelling of the deviant response differed, with inversions being the more neutral (and, importantly, rule-conform) option compared to violations, although both would result in the same action. Both violations and inversions only required a negation of the instructed mapping rule. We deliberately designed these responses to not entail negative feedback or punishment for not following the rule.

The experiment consisted of two tasks (see Figure 1): A violation/inversion task as the Prime, where an instructed mapping rule had to be violated or inverted in a fraction of trials, and a valence task as the Probe, where target words had to be categorised as either positive or negative. To make this complex design more accessible, the following hypotheses relate to the violation framing. The inversion framing, by contrast, served as a baseline to assess if the effects found for rule violations are specific to the semantic framing of the response or not.

If we assume rule violations to have an affective component, committing a violation should alter the categorisation of subsequent valent words (Aarts et al., 2012; Dreisbach & Fischer, 2012). More specifically, the following hypotheses can be derived: If committing a violation represents a negative event, subsequent negative targets should be categorised faster as compared to after rule-based responses. Alternatively, successfully committing a violation might even be a positive event, just like the successful resolution of difficult and conflicting tasks has been argued to represent a reward signal (Schouppe et al., 2015). Consequently, positive targets might be categorised faster after a violation than after a rule-based response, because the successful completion of the more demanding and difficult response might be considered a positive event.

Further, if we assume rule violations to entail an authority-related component, then committing a violation should also alter the categorisation of subsequent authority-related words relative to after rule-based responding, with faster responses to words that are strongly linked to the concept after violations relative to after rule-based responses.

Methods

Participants

Fifty-six participants were recruited (mean age = 27.9 years, SD = 8.6, 18 male, 3 left-handed) and received either course credit or €5 monetary compensation. All participants gave informed consent, were naïve to the purpose of the experiment and were debriefed after the session. This sample size was based on the effect size that was obtained in a pilot study (cf. Supplementary Material). Seven participants were removed from the sample due to high error rates (>30%) or less than 10 trials per design cell and were replaced.

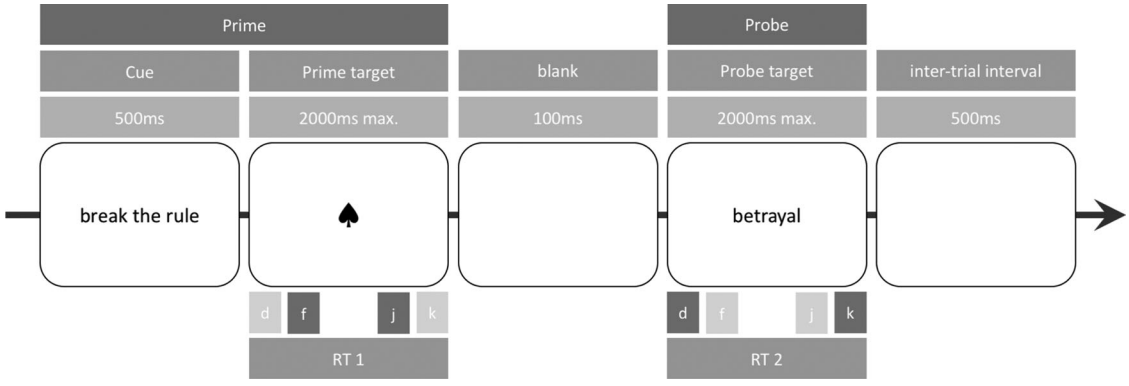

Figure 1. Setup of Experiment 1. The Prime task consisted of a Cue that informed whether the instructed mapping rule had to be used or violated/inverted in the following trial. After 500 ms, the Prime target appeared and called for responses with the f- or j-key (RT1). The Probe target appeared after a blank of 100 ms. It had to be categorised as positive or negative with the d- or k-key (RT2), and we analysed the impact of the Prime response on the following valence categorisation.

Apparatus and stimuli

The experiment was run on a PC with a 22-inch monitor and participants placed their index and middle fingers on the d-, f-, j-, and k-key of the keyboard. Each trial consisted of two tasks, the Prime and the Probe that followed each other in close temporal succession. The stimuli of the Prime were two card symbols (spade: ♠, and diamond: ♦) that prompted left or right presses of the f- and j-key on a standard QWERTZ-keyboard. In the Prime, one of two response options was prompted via written instructions, with the standard response as "follow the rule", and the deviant response option as either "break the rule" or "invert the rule" in the center of the screen before each trial started. Stimuli for the Probe task consisted of 24 nouns – 12 positive and 12 negative – with 6 of each with a strong authority-relation, and the remaining 6 with a weak authority-relation (Figure 2). All Probe words were pre-rated in a pilot study (cf. Supplementary Material) and had to be categorised as positive or negative with the d- and k-key. The Probe words' authority-relation was neither mentioned nor relevant for the completion of the task. All cues and targets were presented centrally in black against a white screen.

Probe target words

For Experiment 1, we selected 24 words from a pre-rated item pool. The Probe words were chosen so that they were either clearly positive or clearly negative, and that they either strongly or weakly related to the concept of authority. This resulted in four clusters (positive strong relation, positive weak relation, negative strong relation, and negative weak relation) that contained six words each. For the words with a weak authority-relation, the Probe words were the German equivalents of present, luck, sun, peace, gain, and benefit ($M_{Valence} = 7.70$, $SD_{Valence} = 1.01$, $M_{Authority} = 2.06$, $SD_{Authority} = 0.50$) for the positive target words and corpse, accident, lie, bankruptcy, betrayal, and disloyalty ($M_{Valence} = 1.78$, $SD_{Valence} = 0.30$, $M_{Authority} = 2.18$, $SD_{Authority} = 0.29$) for the negative words. All these target words were rated lower than 2.5 on the authority scale and were therefore considered weakly related at best. The marked difference in the valence ratings between both types of items, $t(10) = 13.77$, $p < .001$, $d = 9.04$, should, however, allow for easy discrimination between positive and negative words.

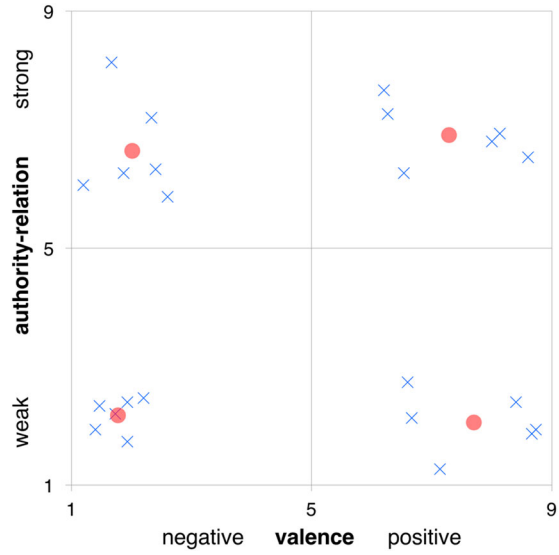

Figure 2. Ratings of the Probe target words. Probe target words were taken from an item pool of 168 words that were pre-rated concerning word valence and their authority-relation, both on a nine-point scale. Mean ratings for valence are depicted on the abscissa (1 = negative, 9 = positive), mean ratings for authority-relation are depicted on the ordinate (1 = not related, 9 = strongly related). Crosses represent mean ratings for individual target words; dots represent the mean ratings for each cluster (descriptive values for individual words can be found in the Appendix).

For the strong authority-relation, we chose mentor, mother, father, parents, doctor, and professor ($M_{Valence} = 7.29$, $SD_{Valence} = 1.07$, $M_{Authority} = 6.91$, $SD_{Authority} = 0.50$) for the positive words, and violence, weapon, punishment, prison, dictatorship, and admonition ($M_{Valence} = 2.01$, $SD_{Valence} = 0.53$, $M_{Authority} = 6.64$, $SD_{Authority} = 0.86$) for the negative words. These words were again chosen because they provided a strong discrimination between positive and negative words, $t(10) = 10.83$, $p < .001$, $d = 6.60$, while the ratings of both, the positive and the negative words, were similar to the ratings of the weakly authority-related target words, $|t|s < 1$, $ps > .364$.

Not only were the valence-ratings matched between the strongly and weakly authority-related target words, but also the authority ratings were similar within the strong and the weak authority-relation, $|t|s < 1$, $ps > .527$. Still, authority ratings clearly differentiated between strong and weak relation within all positive target words, $t(10) = 16.76$, $p < .001$, $d = 9.67$, as well as within all negative target words, $t(10) = 12.08$, $p < .001$, $d = 7.82$.

These four clusters, each containing six words, that were either positive or negative and had a strong or weak relation to authority, allowed for valence and

authority-relation to be manipulated orthogonally (see Figure 2). Note, however, that only the valence dimension was relevant for the participants, as in the Probe, the target words only had to be categorised as positive or negative. While the authority-relation of the Probe target words was manipulated in this experiment, it was neither explicitly instructed nor relevant for the completion of the task.

Procedure

Each trial started with a cue that instructed participants to either follow or break/invert the instructed mapping rule of the Prime task. The instructed rule held that half of the participants were to press the left key when a spade appeared, and the right key if a diamond appeared. The other half was instructed with the opposite mapping for counterbalancing. In 75% of all cases, the cue required participants to employ the instructed mapping rule ("follow the rule") and in 25% of all cases, the instruction called for a deviant response: Half of the participants were instructed with a violation framing, in which the displayed mapping rule had to be violated, and the other half was instructed with an inversion framing, in which the displayed mapping rule had to be inverted. Crucially, both the violation framing and the inversion framing resulted in the same task requirements. This cue was displayed for 500 ms, immediately followed by the Prime target. The Prime target was either a spade or a diamond and required a left or right keypress. It was presented for a maximum of 2000 ms and disappeared as soon as a response was given. A blank screen of 100 ms separated the Prime from the Probe.

For the Probe task, a randomly chosen target word appeared for a maximum of 2000 ms and had to be categorised as positive or negative via keypress. Half of the participants were instructed to press the left key if the target word was positive, and the right key if it was negative. The other half was instructed with the opposite mapping for counterbalancing. The Probe word disappeared as soon as a response was given, the next trial started after an inter-trial interval of 500 ms (Figure 1). Feedback in case of errors was only provided in the training blocks, and not during the experimental blocks. This was done so that the expectation of negative feedback would not overshadow possible negative signals elicited by violation/inversion responses.

Participants completed two short training blocks where the two tasks were presented separately (one block with 24 Prime trials, one block with 24 Probe trials). After that, participants completed 3 experimental blocks of 192 trials each.

Results

Data selection and analyses

For the following analyses, we only used trials from the experimental blocks. We omitted trials in which participants failed to act according to the instruction (Prime: 8.6%, with more commission errors for violations/inversions than for rule-based responses, $t(55) = 9.63$, $p < .001$, $d = 1.29$; Probe: 7.5%, irrespective of Probe valence and Probe authority-relation, $|t|s < 1.19$, $ps > .240$, $ds < 0.16$) and the immediately following trials. Trials were discarded as outliers if any of the measures (RT1, RT2) deviated more than 2.5 standard deviations from the participant's respective cell mean (4.5%). RT1 was then analysed in a 2×2 analysis of variance (ANOVA) with Prime response (rule-based vs. deviant) as within-subjects factor and framing (violation vs. inversion) as between-subjects factor, whereas RT2 was analysed in a $2 \times 2 \times 2 \times 2$ ANOVA with Prime response (rule-based vs. deviant), Probe valence (positive vs. negative) and Probe authority-relation (weak vs. strong) as within-subjects factors and framing (violation vs. inversion) as a between-subjects factor.

Prime responses

A significant effect of Prime response emerged, $F(1,54) = 152.35$, $p < .001$, $\eta_p^2 = .74$, driven by slower responses for the deviant response option (787 ms) than for rule-based responses (693 ms). The violation framing produced descriptively larger effects ($\Delta =$ 105 ms, Figure 3(A)) relative to the inversion instruction ($\Delta = 84$ ms, Figure 3(B)), $F(1,54) = 1.92$, $p = .170$, $\eta_p^2 = .03$, but overall, there were no differences between framings, $F < 1$.

Probe responses

There was an interaction of Prime response and Probe valence, $F(1,54) = 9.29$, $p = .004$, $\eta_p^2 = .15$, as negative words were evaluated faster after violations relative to rule-based responses ($\Delta = 12$ ms, $t(55) = 1.97$, $p = .053$, $d = 0.26$), and positive words were evaluated descriptively faster after rule-based responses relative to after violations ($\Delta = -4$ ms, $t(55) = 0.69$, $p = .495$, $d = 0.09$). This effect was present for both framings, as indicated by a non-significant three-way interaction of Prime response type, Probe valence and framing, $F < 1$ (Figure 3(C,D)). A main effect of Probe authority-relation described responses to Probe words

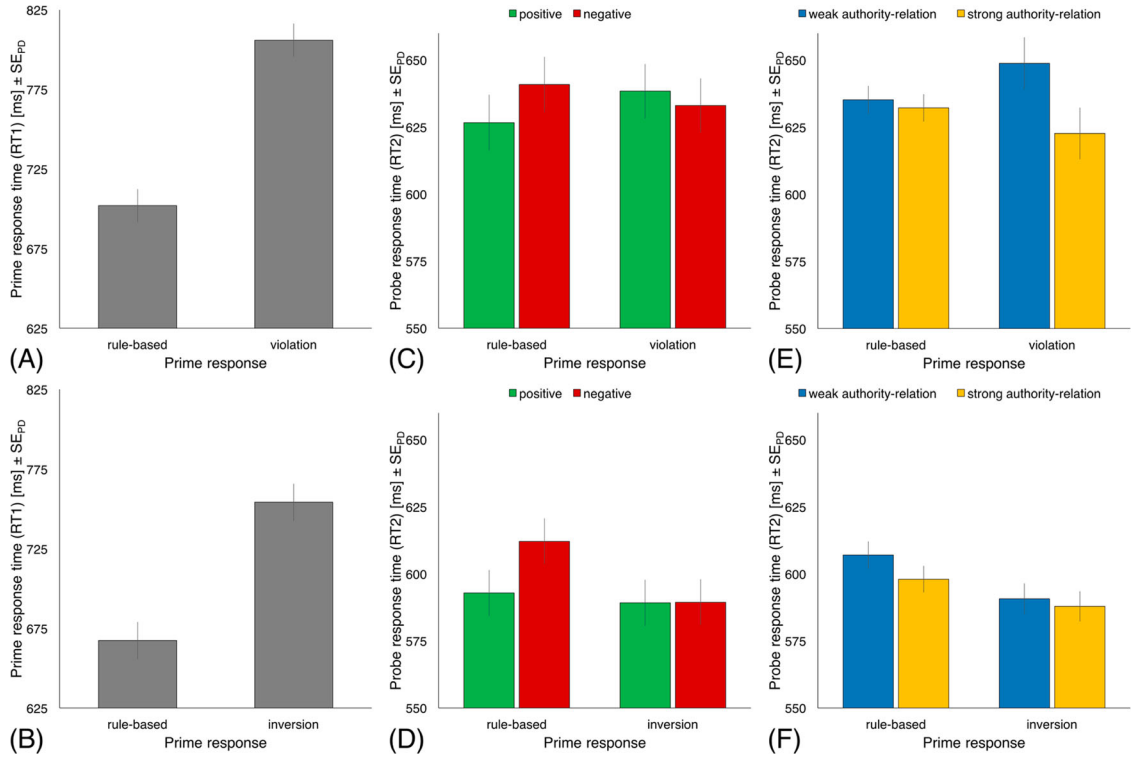

Figure 3. Results for Experiment 1 for participants who received violation instructions (upper row) and participants who received inversion instructions (lower row). Prime response times (RT1; panels (A) and (B)) and Probe response times (RT2, panels (C)–(F)) as a function of Prime response(abscissa), Probe valence (panels (C) and (D): left, green bars for positive targets; right, red bars for negative targets), and Probe authority-relation (panels (E) and (F): left, blue bars for weakly authority-related targets; right, yellow bars for strongly authority-related targets). Error bars represent standard errors of paired differences, for the interactions calculated separately for each instance of Prime response (Pfister & Janczyk, 2013).

that were strongly related to authority as faster (600 ms) than to weakly related Probe words (610 ms), $F(1,54) = 7.66$, $p = .008$, $\eta_p^2 = .12$. Overall, there was a three-way interaction between Prime response type, Probe authority-relation and framing, $F(1,54) = 5.79$, $p = .020$, $\eta_p^2 = .10$, driven by a significant interaction for the violation framing, $F(1,27) = 6.12$, $p = .020$, $\eta_p^2 = .30$ (Figure 3(E), with response benefits for Probe words with a strong authority-relation after violations relative to rule-based responses, $\Delta = 6$ ms, $t(27) = 0.66$, $p = .516$, $d = 0.12$, and response costs for Probe words with a weak authority-relation after violations relative to rule-based responses, $\Delta = -17$ ms, $t(27) = 1.33$, $p = .195$, $d = 0.25$), but not for the inversion framing $F(1,27) = 0.28$, $p = .602$, $\eta_p^2 = .01$ (Figure 3(F), with response benefits after inversions irrespective of Probe authority-relation, $\Delta = 5$ ms, $t(27) = 1.70$, $p = .101$, $d = 0.32$). Finally, there was an interaction between Prime target valence and Prime target authority-relation, $F(1,54) = 13.20$, $p = .001$, $\eta_p^2 = .20$, with no benefit for target words that have a strong authority-relation over target words with a weak relation for negative words ($\Delta = -2$ ms, $t(27) = 0.38$, $p = .704$, $d = 0.05$), but a strong benefit for positive words ($\Delta = 23$ ms, $t(27) = 3.87$, $p < .001$, $d = 0.52$). No other effects or higher-order interactions were significant, Fs < 1.75, ps > .191.

Discussion

In Experiment 1, we probed for affective and authority-related components of rule violations and rule inversions. By employing a Prime-Probe design with a violation/inversion task as the Prime and a word categorisation task as the Probe, we tested whether having committed a violation/inversion modulates subsequent categorisation of positive and negative target words, as well as of words that strongly or weakly relate to the concept of authority.

Results of the Prime task replicated previous findings by showing that it is harder to commit a violation response compared to an inversion response (Wirth, Pfister, Foerster et al., 2016). These results showed a quantitative differences between rule violations and the similar inversion condition, demonstrating that violations are more difficult even compared to instructions that seemingly require the same mental operation and produce the same motor response. This can merely represent a first step at understanding the cognitive architecture of rule violations. What is important here is to show that there are fundamentally different cognitive processes at work when actively committing a violation, even though a mere observer could not differentiate between a violation and an inversion.

In the Probe trials, we found both, violations and inversions, to speed up the categorisation of negative target words. Violations and inversions seem to be considered a negative event. This result helps to distinguish between the two competing hypotheses that were raised in the Introduction. Despite these responses being more difficult and demanding, the successful resolution of these tasks did not seem to trigger a reward signal (Schouppe et al., 2015), but instead promoted the detection of negative stimuli. However, the affective component did not pose as a unique feature of rule violations, as rule inversions triggered the same subsequent sensitivity towards negative stimuli. We therefore conclude that the affective component is likely driven by the negation that is included in both, violations and inversions. The processing of violations and inversions via negation of an instructed mapping rule requires that the original rule and the modulated rule are concurrently active (see Wirth, Pfister, Foerster et al., 2016). This dual activation triggers conflict, which has been linked to negative valence (Dreisbach & Fischer, 2012).

The qualitative difference between violations and inversions can be found by analysing the authority-relation of the Probe trials. Here, we saw that only violations sensitised towards subsequent authority-related stimuli, while inversions did not. This process is unique to violations, and it might reflect latent expectations of punishment (Pfister, Wirth, Schwarz, Steinhauser et al., 2016). Punishment, after all, has strong ties to the concept of authority, because usually, authoritarian figures (parents, teachers, and superordinates) are the ones entitled to punish. Even though we designed the experiment to not include any feedback, it might be that violations and punishment are inseparably interlinked, so that this expectation cannot be shut down easily.

Finally, the lack of a three-way interaction between Prime response type, Probe valence and Probe authority-relation suggests that the affective component and the authority-related component are separate entities. After a violation, authority-related words were not evaluated faster *because* they are negative or vice versa, so both components are not the symptom of the same cognitive mechanism, but seem to work independently. Still, the interaction between the Probe valence and the Probe authority-relation was surprising. Even though the Probe target words were chosen via their explicit ratings so that valence and authority-relation could be manipulated orthogonally, their implicit evaluation via response times seems to differ, with positive authority-related words categorised fastest. However, since this overall pattern of results is present for all conditions (rule-based responses, violations and inversions), and differs from the specific pattern found after rule violations, we believe that this interaction can hardly drive the main results. Before drawing any further conclusions from this data, Experiment 2 will serve to address a possible confound of Experiment 1.

Experiment 2

Introduction

In Experiment 1, we found that violating a rule triggers affective and authority-related processes that modulate subsequent information processing. The affective aftereffects of rule violations seem to reflect the cognitive demands of resolving cognitive conflict and are therefore not specific to rule violations. A heightened sensitivity towards authority-related stimuli, by contrast, seems to be specific to rule violations and does not occur for behaviour that is in accordance with a given rule. However, an alternative explanation might be that it is not the breaking of a rule itself that causes these effects; rather, the results of Experiment 1 could stem from semantic priming (Meyer & Schvaneveldt, 1971): Participants who were instructed to violate rules were obviously confronted with the concept of rule violation as part of each rule violation cue, whereas participants who were instructed to invert a mapping rule were not confronted with any semantics that would relate to rule-breaking. The

observed aftereffects might therefore not reflect a property of rule violations, but may alternatively be due to a pre-activation of the corresponding semantic networks.[1] For Experiment 2, we adjusted the experimental procedure so that for the Prime, the instructional cue was still displayed (*follow* vs. *break the rule*), but the corresponding action did not have to be executed. If the effects found in Experiment 1 were simply due to semantic priming, the same effects should emerge again. However, if the effects are tied to the execution of the response, they should diminish or even vanish.

Methods

Participants

Twenty-eight participants were recruited (mean age = 24.7 years, SD = 6.3, 5 male, no left-handed) and received either course credit or €5 monetary compensation. All participants gave informed consent, were naïve to the purpose of the experiment and were debriefed after the session. This sample size was again based on the effect size that was obtained in the pilot study (cf. Supplementary Material). Four participants were removed from the sample due to high error rates (>30%) or less than 10 trials per design cell and were replaced.

Procedure

The experiment was similar to the violation framing of Experiment 1 with the following changes: Instead of executing the Prime response, participants were now confronted with the cue ("follow the rule" or "break the rule") without having to act on it. After the cue was presented, there was a blank screen of 475 ms (mean RT1 in Experiment 1) instead of the Prime target, which was then followed by the 100 ms blank. This setup ensured that the temporal structure between Experiments 1 and 2 was comparable. The Probe task was unchanged. To further ensure that the cue was still read and processed, participants were tasked with counting how often the instruction to break a rule appeared. At the end of each block, they were then asked to specify their result, and in case of an error, feedback was provided together with the correct answer. To not have the exact same number of required rule violations per block (as in Experiment 1), in each trial the cue was chosen randomly, with a 25% chance of a violation cue. This way, the overall probability of encountering a violation cue was still similar for both experiments.

To account for the counting task, the experiment was further divided into a larger number of blocks while decreasing the number of trials per block. That is, participants completed two short training blocks where the two tasks were presented separately (one block with 24 counting Prime trials, one block with 24 Probe trials). After that, participants completed 8 experimental blocks of 72 trials each.

Results

Data selection and analyses

For the following analyses, we only used trials from the experimental blocks. We omitted trials in which participants failed to act according to the instruction (Probe: 7.2%, irrespective of Probe valence and Probe authority-relation, $|t|$s < 1.53, ps > .136, ds < 0.29) and the immediately following trials. Further, the data of an entire block were discarded if participants' estimate of the number of "break the rule" cues was off by more than 3 to ensure that participants properly processed the cues (12.1%). Trials were discarded as outliers if RT2 deviated more than 2.5 standard deviations from the participant's respective cell mean (2.7%).[2] RT2 was then analysed in a 2 × 2 × 2 ANOVA with Prime cue (rule-based vs. violation), Probe valence (positive vs. negative) and Probe authority-relation (weak vs. strong) as within-subjects factors.

Probe responses

As in Experiment 1, there was an interaction between Probe valence and Probe authority-relation, $F(1,27) = 5.11$, $p = .032$, $\eta_p^2 = .16$, with descriptive costs for target words with a strong authority-relation over target words with a weak relation for negative words ($\Delta = -9$ ms, $t(27) = 1.20$, $p = .240$, $d = 0.23$), but a benefit for positive words ($\Delta = 16$ ms, $t(27) = 2.30$, $p = .029$, $d = 0.44$). However, neither Prime cue, $F < 1$, nor any interaction involving Prime cue, Fs < 1.38, ps > .250 (Figure 4), was significant. No other effects or higher-order interactions turned significant, Fs < 2.71, ps > .111.

Discussion

In Experiment 2, we tested whether the effects obtained in Experiment 1 can be explained by

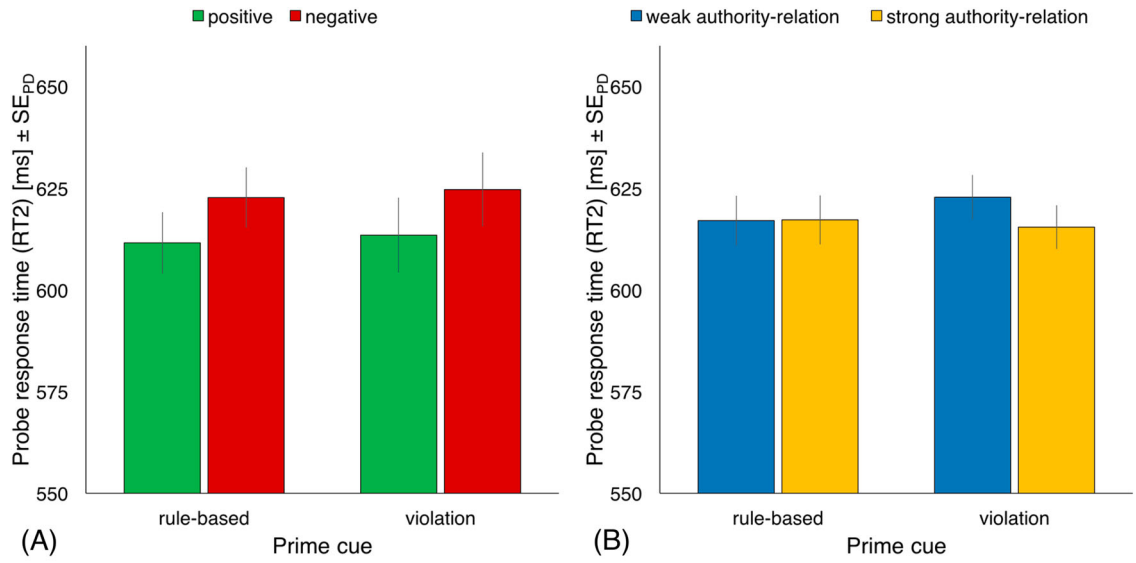

Figure 4. Results for Experiment 2. Probe response times (RT2) as a function of Prime cue (abscissa), Probe valence (panel (A): left, green bars for positive targets; right, red bars for negative targets), and Probe authority-relation (panel (B): left, blue bars for weakly authority-related targets; right, yellow bars for strongly authority-related targets). Error bars represent standard errors of paired differences, calculated separately for each instance of Prime cue (Pfister & Janczyk, 2013).

semantic priming (Meyer & Schvaneveldt, 1971). We adapted the procedure of Experiment 1 to no longer include the Prime response, but only retained the Prime cue including its semantics (*follow* vs. *break the rule*). If the subsequent sensitivity towards both, negative and authority-related stimuli, was driven by the semantic content of the cue rather than the following response, then the omission of the Prime response should yield the same results as in Experiment 1. If, however, the affective and authority-related components of rule violations were triggered by their execution, then we should find no effect with this new setup. Data showed that omitting the Prime response annulled both effects found in Experiment 1, and thus semantic priming is unlikely to account for the affective or the authority-related after-effects of rule violations.

While a semantic priming explanation of the sensitivity for negative targets after a violation is not supported by the present data, it might still be that the prompt to violate a rule is inherently negative, but the current setup is unable to identify such an effect, as Prime and Probe task are not presented in sufficiently close temporal succession. We can, however, conclude that the affective component found in the Probe trials of Experiment 1 does not rely on semantic priming by the violation prompt alone, but is due to having executed the corresponding response.

Interestingly, the interaction between Probe valence and Probe authority-relation was replicated in Experiment 2, again with positive authority-related target words categorised faster than the remaining combinations. This shows that participants respond consistently to the Probe words across both experiments. Without the Prime response, any systematic influence of the Prime response on the Probe response times vanished, but the regularities within the Probe response times remained.

General discussion

In the present experiments, we investigated affective and authority-related components of rule violations and compared them to rule inversions. By employing a Prime-Probe design with a violation task as the Prime and a word categorisation task as the Probe, we could identify how having committed a violation response modulates the sensitivity towards valent and authority-related stimuli. As a baseline, the after-effects of rule inversions were tested the same way by having participants respond according to an inverted rule (Experiment 1). Additionally, we ruled out a possible alternative explanation in terms of semantic priming effects (Experiment 2).

The data of the Prime responses showed that it is indeed slightly harder to commit a violation compared

to a rule inversion, replicating previous results (Wirth, Pfister, Foerster et al., 2016). Even though a violation and an inversion seem to require the same mental operation in our scenario (inhibiting the automatic conformity tendency, inverting the instructed mapping rule and then applying the newly derived rule), the labelling of the response influenced the difficulty of these responses: violations were harder and more effortful than rule inversions.

This distinction is an important first step, but it is only a quantitative one. To show that violations are not simply an especially difficult instance of a conflicting task, to show that they are not something more, but something *else*, something that is qualitatively different, we tested how violations and inversion modulated a subsequent categorisation of valent and authority-related words. The data of these Probe responses showed that both violations and inversions sensitise towards negative stimuli: while after rule-based responses, positive target words were categorised faster, a violation and an inversion seem to promote the processing of negative target words, which were consequently categorised faster afterwards relative to after rule-based responses. This result stresses the conflicting nature and aversive quality of violations (Aarts et al., 2012; Dreisbach & Fischer, 2012). However, this was not unique to violations, but is also true for inversions. Mere priming of violations, however, did not produce this effect, highlighting that the sensitivity towards subsequent negative events is tied to having carried out the action. So while the execution of rule violations seems to entail an affective component, it can be attributed to the simultaneous activation of two responses, the default, rule-based response and the deviant, negated response, as this is also the case for inversion responses. This double activation makes these responses more difficult and triggers an aversive signal, which in return promotes the processing of subsequent negative stimuli.

The analysis of the authority-related dimension of the Probe target words, however, tells a different story. Here, we observed a clear dissociation between violations and inversions. While rule violations seem to specifically promote the processing of authority-related stimuli, this is not the case with inversions. This shows that violations additionally trigger heightened attention towards authorities, as authority-related figures might be especially relevant in these situations. Next to sensitizing towards negative stimuli, violations can also act as a prime for subsequent authority-related stimuli.

Heightened attention towards authority-related stimuli that is specific to violations might reflect latent expectations of sanctions and punishment (Pfister, Wirth, Schwarz, Steinhauser et al., 2016): Even though we explicitly omitted this in our experimental design, participants might automatically expect negative feedback after committing a violation response. After all, punishment after breaking a rule is at the core of the development of moral and social behaviour (Kohlberg, 1963; Kohlberg & Hersh, 1977; Piaget, 1932), and is also essential to strengthen cooperation within groups (Fehr & Gächter, 2002; Yamagishi, 1986). These expectations of punishment after breaking a rule might therefore represent an automatic process that cannot be invalidated by instruction, at least in the timeframe of our experiments. On the other hand, an analysis of situational factors in enterprises has identified "perceived lack of management care", "poor supervision", and "belief that bad outcomes will not happen" as key factors to promote the likelihood of violations at the workplace (Reason, 1995, p. 86). So in the long run, this latent expectancy of punishment for breaking the rules could possibly be suspended by local, situational factors.

The examples above describe situations in which a rule-breaker is confronted with an authority that is entitled to make and enforce the rules. What if we changed the perspective here? How about people with authority, people with power, breaking the rules? Assuming an expansive power posture (consciously or inadvertently) could be shown to increase the probability of stealing money, cheating on a test, and committing a traffic violation compared to participants who assumed a contractive body posture (Yap et al., 2013). An authoritarian feeling (or posture) seems to annul any latent expectation of punishment after rule violations, thereby making them more likely (see also Trautmann, van de Kuilen, & Zeckhauser, 2013; van Kleef, Homan, Finkenauer, Gündemir, & Stamkou, 2011; van Kleef, Wanders, Stamkou, & Homan, 2015). And again, a "perceived license to bend the rules" by an individual has been identified as a factor to increase the likelihood of violations at the workplace (Reason, 1995, p. 86).

In the current set of experiments, we show that rule violations sensitise towards negative and authority-related stimuli. But maybe this is not a one-way street: Based on these results, the question arises whether being confronted with a negative or authority-related stimulus makes violations (a) more or less likely and (b) more or less difficult. If in fact this is a two-way street and the confrontation with a

valent- or authority-related setting can reliably modulate the likelihood of violations, this will have strong implication for cognitive psychology, social psychology, and even forensics (Jusyte et al., in press).

Further, future research should investigate more closely how these results relate to rule-breaking behaviour in real life. The current experiments were deliberately designed to investigate the cognitive mechanisms underlying rule violations, and therefore factors such as prior experience with violations, consequences or punishment for breaking the rules, and social factors were intentionally omitted. Introducing these factors systematically to the experimental setup might reveal commonalities as well differences to the behaviour observed in the present study. This could be done by measuring walk onset time and walking speed when jaywalking or crossing red traffic lights in real or simulated environments, with or without social bystanders, with an incentive (saving time) or punishment (possible traffic fine).

Whatever the outcome of these experiments might be, in a highly controlled setup, we show that violation responses trigger processes that sensitise not only toward negative stimuli, which likely reflects an automatic evaluation of the agent's own response (Aarts et al., 2012), but also toward authority-related stimuli, which is suggestive of even latent expectations of punishment after breaking a rule (Pfister, Wirth, Schwarz, Steinhauser et al., 2016). This authority-related sensitivity after breaking a rule is specific to violation responses and cannot be explained by negation processing, showing that violations are not just quantitatively different from simple conflict tasks, but also differ in a qualitative manner.

Notes

1. We thank two anonymous reviewers for raising this alternative explanation.
2. Even though there was no Prime response in Experiment 2, we still abbreviate Probe response times as RT2 to remain consistent with the terminology of Experiment 1.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Aarts, K., De Houwer, J., & Pourtois, G. (2012). Evidence for the automatic evaluation of self-generated actions. *Cognition*, *124*(2), 117–127.

Aarts, K., De Houwer, J., & Pourtois, G. (2013). Erroneous and correct actions have a different affective valence: Evidence from ERPs. *Emotion*, *13*(5), 960–973.

Berry, C. M., Ones, D. S., & Sackett, P. R. (2007). Interpersonal deviance, organizational deviance, and their common correlates: A review and meta-analysis. *Journal of Applied Psychology*, *92*(2), 410–424.

Botvinick, M. M. (2007). Conflict monitoring and decision making: Reconciling two perspectives on anterior cingulate function. *Cognitive, Affective, & Behavioral Neuroscience*, *7*(4), 356–366.

Dommes, A., Granié, M. A., Cloutier, M. S., Coquelet, C., & Huguenin-Richard, F. (2015). Red light violations by adult pedestrians and other safety-related behaviors at signalized crosswalks. *Accident Analysis & Prevention*, *80*, 67–75.

Dreisbach, G., & Fischer, R. (2012). Conflicts as aversive signals. *Brain and Cognition*, *78*(2), 94–98.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 137–140.

Fillenbaum, S. (1966). Memory for gist: Some relevant variables. *Language and Speech*, *9*(4), 217–227.

Fritz, J., & Dreisbach, G. (2013). Conflicts as aversive signals: Conflict priming increases negative judgments for neutral stimuli. *Cognitive, Affective, & Behavioral Neuroscience*, *13*(2), 311–317.

Fritz, J., & Dreisbach, G. (2015). The time course of the aversive conflict signal. *Experimental Psychology*, *62*, 30–39.

Gilbert, D. T. (1991). How mental systems believe. *American Psychologist*, *46*(2), 107–119.

Jusyte, A., Pfister, R., Mayer, S. V., Schwarz, K. A., Wirth, R., Kunde, W., & Schönenberg, M. (in press). Smooth criminal: The profound cognitive flexibility of convicted rule-breakers. *Psychological Research*, (ahead of print), 1–8. doi:10.1007/s00426-016-0798-6

King, M. J., Soole, D., & Ghafourian, A. (2009). Illegal pedestrian crossing at signalised intersections: Incidence and relative risk. *Accident Analysis & Prevention*, *41*(3), 485–490.

Kohlberg, L., & Hersh, R. H. (1977). Moral development: A review of the theory. *Theory into Practice*, *16*(2), 53–59.

Kohlberg, L. (1963). The development of children's orientations toward a moral order. *Human Development*, *6*(1–2), 11–33.

Lindström, B. R., Mattsson-Mårn, I. B., Golkar, A., & Olsson, A. (2013). In your face: Risk of punishment enhances cognitive control and error-related activity in the Corrugator supercilii muscle. *PLoS ONE*, *8*(6), e65692.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*(2), 227–234.

Parker, D., Reason, J. T., Manstead, A. S. R., & Stradling, S. G. (1995). Driving errors, driving violations and accident involvement. *Ergonomics*, *38*(5), 1036–1048.

Pfister, R., & Janczyk, M. (2013). Confidence intervals for two sample means: Calculation, interpretation, and a few simple rules. *Advances in Cognitive Psychology*, *9*(2), 74–80.

Pfister, R. (2013). *Breaking the rules: Cognitive conflict during deliberate rule violations*. Berlin: Logos.

Pfister, R., Wirth, R., Schwarz, K. A., Foerster, A., Steinhauser, M., & Kunde, W. (2016). The electrophysiological signature of deliberate rule violations. *Psychophysiology*, *53*, 1870–1877.

Pfister, R., Wirth, R., Schwarz, K., Steinhauser, M., & Kunde, W. (2016). Burdens of non-conformity: Motor execution reveals

cognitive conflict during deliberate rule violations. *Cognition*, *147*, 93–99.

Phipps, D. L., Parker, D., Pals, E. J. M., Meakin, G. H., Nsoedo, C., & Beatty, P. C. W. (2008). Identifying violation-provoking conditions in a healthcare setting. *Ergonomics*, *51*(11), 1625–1642.

Piaget, J. (1932). *The moral judgement of the child*. Glencoe, IL: The Free Press.

Rabbitt, P., & Rodgers, B. (1977). What does a man do after he makes an error? An analysis of response programming. *Quarterly Journal of Experimental Psychology*, *29*(4), 727–743.

Reason, J. (1990). *Human error*. New York, NY: Cambridge University Press.

Reason, J. (1995). Understanding adverse events: Human factors. *Quality and Safety in Health Care*, *4*(2), 80–89.

Rosenbloom, T. (2009). Crossing at a red light: Behaviour of individuals and groups. *Transportation Research Part F: Traffic Psychology and Behaviour*, *12*(5), 389–394.

Schouppe, N., Braem, S., De Houwer, J., Silvetti, M., Verguts, T., Ridderinkhof, K. R., & Notebaert, W. (2015). No pain, no gain: The affective valence of congruency conditions changes following a successful response. *Cognitive, Affective, & Behavioral Neuroscience*, *15*(1), 251–261.

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review*, *8*(3), 220–247.

Trautmann, S. T., van de Kuilen, G., & Zeckhauser, R. J. (2013). Social class and (un)ethical behavior: A framework, with evidence from a large population sample. *Perspectives on Psychological Science*, *8*(5), 487–497.

van Kleef, G. A., Homan, A. C., Finkenauer, C., Gündemir, S., & Stamkou, E. (2011). Breaking the rules to rise to power: How norm violators gain power in the eyes of others. *Social Psychological and Personality Science*, *2*(5), 500–507.

van Kleef, G. A., Wanders, F., Stamkou, E., & Homan, A. C. (2015). The social dynamics of breaking the rules: Antecedents and consequences of norm-violating behavior. *Current Opinion in Psychology*, *6*, 25–31.

Wegner, D. M., Coulton, G. F., & Wenzlaff, R. (1985). The transparency of denial: Briefing in the debriefing paradigm. *Journal of Personality and Social Psychology*, *49*(2), 338–346.

Wirth, R., Pfister, R., Foerster, A., Huestegge, L., & Kunde, W. (2016). Pushing the rules: Effects and aftereffects of deliberate rule violations. *Psychological Research*, *80*(5), 838–852.

Wirth, R., Pfister, R., & Kunde, W. (2016). Asymmetric transfer effects between cognitive and affective task disturbances. *Cognition and Emotion*, *30*(3), 399–416.

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology*, *51*(1), 110–116.

Yap, A. J., Wazlawek, A. S., Lucas, B. J., Cuddy, A. J., & Carney, D. R. (2013). The ergonomics of dishonesty the effect of incidental posture on stealing, cheating, and traffic violations. *Psychological Science*, *24*(11), 2281–2289.

Appendix

Table A1. Descriptive values for the individual ratings of the Probe target words, taken from the pre-rated pool of 168 words.

German word	Authority-relation	strong	Gewalt	Diktatur	Waffe	Gefängnis	Strafe	Mahnung	Professor	Arzt	Mentor	Vater	Eltern	Mama
English equivalent			violence	dictatorship	Weapon	prison	punishment	admonition	professor	doctor	mentor	father	parents	mother
Valence rating			1.20	1.67	1.87	2.33	2.40	2.60	6.20	6.27	6.53	8.00	8.13	8.60
Authority rating			6.07	8.13	6.27	7.20	6.33	5.87	7.67	7.27	6.27	6.80	6.93	6.53
German word		weak	Unfall	Leiche	Untreue	Bankrott	Verrat	Lüge	Vorteil	Gewinn	Geschenk	Frieden	Sonne	Glück
English equivalent			accident	corpse	disloyalty	bancruptcy	betrayal	lie	benefit	gain	present	peace	sun	luck
Valence rating			1.40	1.47	1.73	1.93	1.93	2.20	6.60	6.67	7.13	8.40	8.67	8.73
Authority rating			1.93	2.33	2.20	1.73	2.40	2.47	2.73	2.13	1.27	2.40	1.87	1.93
			negative								Positive			
			Valence											

Note: All 24 German target words with their English equivalent, their valence rating (1 = negative, 9 = positive) and their authority rating (1 = not related, 9 = strongly related), arranged in clusters for strong and weak authority-relation (rows) and negative and positive valence (columns).