Consequences of bending and breaking the rules

Inaugural-Dissertation
zur Erlangung der Doktorwürde der
Fakultät für Humanwissenschaften
der
Julius-Maximilians-Universität Würzburg

vorgelegt von
Robert Wirth
aus Limburg an der Lahn
Würzburg, 2017

Erstgutachter:	Prof. Dr. Wilfried Kunde
Zweitgutachter:	Prof. Dr. Andrea Kiesel
Tag des Kolloqiums:	18.09.2017

Zusammenfassung.

Soziales Miteinander ist durch Regeln und Normen organisiert. Die hier beschriebenen Experimente untersuchen die kognitive Architektur von absichtsvollen Regelverstößen. Dazu wurde ein Setting entwickelt, in dem Regeln befolgt oder gebrochen werden mussten, und das Brechen dieser Regeln keine negativen Konsequenzen nach sich zog. Selbst ohne soziale Unterstützung, die das Brechen von Regeln leichter oder schwerer machen könnte, fanden wir, dass allein das Bezeichnen eines Verhaltens als Regelverletzung spezifische Kosten erzeugte: Die Planung dieses Verhaltens ist deutlich erschwert, und die Ausführung zeigt spezifische Verhaltensmuster. Regelverletzungen ähneln hierbei im weitesten Sinne Negationen, aber beinhalten zusätzliche Komponenten.

Die Frage wie genau sich die kognitive Kontrolle regelwidriger Verhaltensweisen von der Verarbeitung von Negationen unterscheidet, steht im Zentrum der vorliegenden Arbeit. Die folgenden Experimente zeigen darüber hinaus neben negativen affektiven Konsequenzen, die sowohl Regelbrüche als auch Negationen vorweisen, insbesondere eine direkte Bahnung autoritätsbezogener Konzepte, die eine spezifische Begleiterscheinung absichtsvoller Regelverstöße darstellt.

Als nächstes wurde getestet, wie die kognitiven Kosten von Regelverletzungen durch kürzliche oder häufige Ausführung gemindert werden können. Hier zeigte sich, dass die Kombination aus beiden Faktoren die größte Reduktion

kognitiver Kosten des Regelverstoßes erbrachte. Ein Transfer von kognitiver Kontrolle von einer anderen Aufgabe konnte jedoch nicht beobachtet werden.

Ein Modell, das die hier dargestellten empirischen Ergebnisse vereint, wird abschließend diskutiert. Als Variation eines Modells zum Aufgabenwechsel erklärt es die kognitiven Prozesse, die einer Regelverletzung unterliegen und zeigt Verarbeitungsschritte auf, die für Regelverletzungen spezifisch sind.

Summary.

Social life is organized around rules and norms. The present experiments investigate the cognitive architecture of rule violations. To do so, a setting with arbitrary rules that had to be followed or broken was developed, and breaking these rules did not have any negative consequences. Removed from any social influences that might further encourage or hinder the rule breaker, results suggest that simply labeling a behavior as a rule violation comes with specific costs: They are more difficult to plan and come with specific behavioral markers during execution. In essence, rule violations resemble rule negations, but they also trigger additional processes.

The question of what makes rule violations more difficult than rule inversions is the major focus of the remaining experiments. These experiments revealed negative affective consequences of rule violation and rule inversions alike, while rule violations additionally prime authority-related concepts, thus sensitizing towards authority related stimuli.

Next, the question how these burdens of non-conformity can be mitigated was investigated, and the influence of having executed the behavior in question frequently and recently was tested in both negations and rule violations. The burdens of non-conformity can best be reduced by a combination of having violated/negated a rule very frequently and very recently. Transfer from another task, however, could not be identified.

To conclude, a model that accounts for the data that is currently presented is proposed. As a variant of a task switching model, it describes the cognitive processes that were investigated and highlights unique processing steps that rule violations seem to require.

1. Introduction.

You are a brave one. Despite the clear instruction on the sleeve not to open this book and read this dissertation, you did so anyway. You broke the rules. You did what was clearly interdicted. And what is worse, you knew better. But what took you so long? You clearly hesitated, just for a bit. Barely noticeable to a bystander, you held off just for an instant before you opened these pages. That is what gave you away.

The rule that was communicated on the title page of this book is a rather simple and straightforward example of what a rule can be. However, rules can be defined on a wide spectrum, from such universal and explicit instructions ("Keep off the grass"), over more complex rules that incorporate situational demands ("If the traffic light is red, then stop"), up to social and moral norms ("Do no harm").

Interestingly, most humans automatically comply to rules, even when they are not explicit. Take for example the Line Judgement Task (Asch, 1956). In this setup, a participant matches the length of target lines to a standard reference line after a couple of confederates have given their judgement publicly within the group, thereby implicitly providing a group norm. Crucially, in some of the cases, the confederates unanimously give an obviously wrong answer, and the participant, in the end, adjusts their judgement according to this wrong group norm. The participants implicitly conformed to the group norm and overcame their initial estimate even though these norms were at no point explicit. However, with explicit norms, this can result in even more extreme cases. Consider the Milgram experiment (Milgram, 1963). Here, participants were required to administer painful electrical shocks with an increasing voltage to a seemingly suffering confederate in the next room, and they were instructed to administer these shocks even though the confederate begged them not to. Still, the simple instruction by an authority figure sufficed to lead a considerable percentage of participants to administer electrical shocks that would potentially be lethal. The simple command of an authority seems to create considerable pressure to obey the rules and conform.

Taken together, we see that rules have the power to steer the behavior of an individual, both in groups and in isolated settings. Plus, adherence to rules also comes with positive side-effects, as rule-compliant individuals are viewed as trustworthy and good social partners (Everett, Pizarro, & Crockett, 2016). Such positive side-effects even seem to be sufficiently adaptive to have rendered rule-based behavior an evolutionary default (Hoffman, 1981). People like people who stick to the rules. And an evolutionary advantage for those who adhere to (social) rules might be a driving force that allowed humans to create and maintain complex social structures in the first place.

However, by reading this, you did the exact opposite of what you were supposed to do. You knew what (not) to do, but you did it anyway. Rule violations, such as the one you are just committing, are far less understood. Mostly, rule violations have been studied from a third-person perspective. Studies in this framework 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 prime measure (Phipps et al., 2008;

Reason, 1990; Yap, Wazlawek, Lucas, Cuddy, & Carney, 2013). 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.

For the present work, we want to adopt a first-person perspective, where we analyze the cognitive and affective processes taking place in the agent when he or she actively violates a rule. To do so, we will assume the minimal criteria for a rule violation: a person knows which behavior to apply to which situation, but deliberately does not do so. Put differently, that means that an agent needs to be aware of a rule and the behavior that it prescribes, but deliberately performs a different action. Other factors, such as the locus of the decision (whether the person decides for themselves whether they want to violate a rule, of whether they do that by an external prompt; Reason, 1995), negative consequences that might ensue violation behavior (e.g., punishment), the actual content of the rule (does it make sense to follow the rule or is it arbitrary), and inter-individual differences might also play an important role, but might differ from case to case. This minimal defining feature, knowing what to do and deliberately not doing it, is what unites all types of rule violations, and can therefore serve as our basic premise.

Studying the cognitive processes involved in rule-violation behavior from a first-person perspective requires experimental paradigms in which a behavior can be clearly identified as a deliberate rule violation rather than an unintended action slip or mistake. There are first studies that tackled the idea of deliberate rule violations (Pfister, Wirth, Schwarz, Steinhauser, & Kunde, 2016). In a setup where participants were confronted with arbitrary mapping rules (If target = X, then response 1, if target = Y, then response 2), they had to respond by moving the mouse cursor on a computer screen and they could either choose freely or were instructed to violate these mapping rules at times. The data suggests that it is indeed hard to overcome even simple, arbitrary rules: When violating a rule, the original rule remains activated and therefore shapes our behavior. In these experiments, an impact of the rule representation was visible in terms of movement trajectories that were attracted toward the rule-conform option during rule violation. It is almost ironic that when we try hard not to follow a rule, this is exactly when we cannot suppress its influence (Wegner, 2009). Even though these findings are suggestive of a continued rule representation during violation behavior, they can merely represent a first step towards understanding the cognitive architecture of deliberate rule violations.

For the current work, I aimed at broadening our understanding of the cognitive mechanisms that process how our automatic tendency to adhere to the rules can be overcome, and whether this violation behavior then leaves any traces on the violating agent.

The first line of research (Chapter 2. What. Experiments 1-3) explores **what** effects rule violations pose on the acting agent and what consequences follow the execution of rule violation behavior. After that (Chapter 3. Why. Experiments 4-7), I will investigate **why** rule violations might be special, and finally (Chapter 4. How. Experiments 8-10), I will test **how** these burdens of committing rule violations can be mitigated via experimental manipulations. The General Discussion (Chapter 5.) will then integrate these results and provide a working model that explains how rule violations are processed.

2. What.

In the first line of research, I refined the experimental approach that was used in the initial experiments described in the Introduction (Pfister et al., 2016) and replicate the main findings.[1] Further, one critical limitation of the described work is its focus on rule violations as single, isolated instances. With Experiments 1-3, I aimed at setting rule violations in context by investigating the impact of previous instances of rule-based or violation behavior and the impact of different instructional framings on an agent's performance. Based on the assumption that rule violations entail a conflict between the rule-based and the violation response, I assume to find conflict adaptation processes (Botvinick, Barch, Carter, & Cohen, 2001; Gratton, Coles, & Donchin, 1992). I therefore adapted our previous methods to investigate which cognitive processes go along with rule violations, and, importantly, how these violations influence subsequent behavior. These experiments were designed to capture and compare parameters of not only the decision process between rule-based and violation behavior, but also of the execution of the response. Namely, I analyzed the movement trajectories of participants' sweeping responses on the touchscreen of an iPad while they followed or violated an instructed rule, which required the movement of the finger to a certain location, according to a certain stimulus. Based on these trajectories, I was able to compute specific parameters that mirror specific cognitive processes, i.e., the speed of

[1] The data that this work is based on is published in Wirth, Pfister, Foerster, Huestegge, & Kunde, 2016, in *Psychological Research*.

response planning, or spatial and temporal aspects of the response execution (Pfister, Janczyk, Wirth, Dignath, & Kunde, 2014; Wirth, Pfister, & Kunde, 2016). Experiment 1 employed one simple rule that had to be violated at times, whereas Experiment 2 added a second rule that specifically called for the response that was previously labeled “violation”, to control for effects of responding in a reversed rule mapping (Schroder, Moran, Moser, & Altmann, 2012). Finally, Experiment 3 provided an additional control group by addressing inversions of an instructed rule as compared to two reversed rules in Experiment 2. A direct comparison of the experiments therefore allowed me to pinpoint specific effects and aftereffects of violating a simple S-R rule.

2.1 Experiment 1.

Experiment 1 was designed (a) to quantify the difficulty that violations pose on the acting agent and (b) to investigate the impact of such violations on subsequent behavior. But isolating the burdens of non-conformity comes with unique challenges, because when comparing rule-based behavior to rule violations, in the latter case there are a lot of additional factors that might influence the results. Therefore, I conceptualized rule violations as responses that are counter-indicated by an instructed, but totally arbitrary mapping rule. 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 me to isolate the cognitive architecture that processes (non-)conformity removed from any social influences, prior experience with non-conformity, morals, and expectations of punishment, and left me with a highly controlled experimental setup: Neither, following or breaking a rule, comes with a prior training benefit, or an avoidance due to (social) punishment. Further, participants were instructed whether to follow or break a rule. Again, this was done 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 signature of non-conformity was uninfluenced by the type of choice (Pfister et al., 2016). To gauge the cognitive mechanism that processes non-conformity (rather than measuring real-life costs of violating real-life rules), I opted for the highly-controlled approach in the following studies.

So for Experiment 1, I used a simple S-R rule that mapped two target stimuli to a left and a right sweeping response on an iPad. This rule had to be followed most of the time, but had to be violated in a fraction of trials (i.e., akin to the definition of “necessary violations”; Reason, 1990, 1995). I applied a two-dimensional finger tracking design to not only depict the impact of violations in terms of an extra amount of processing time, but also in terms of distinct spatial signatures. Participants had to sweep their finger from a starting area in the bottom center of the display to an upper-left or an upper-right target area on the iPad’s touchscreen. The critical question was whether the initiation and execution of movements would vary as a function of current response type (rule-based vs. violation behavior) and, crucially, also as a function of preceding response type.

Methods.

Participants.

Twenty participants were recruited (mean age = 21.0 years, *SD* = 2.3, 5 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.

Apparatus and stimuli.

The experiment was run on an iPad in portrait mode, which sampled the participants' finger movements at 100 Hz. Viewing distance was about 50cm. I used two chess symbols (king, ♔, and pawn, ♙) as target stimuli to prompt movements to the left or to the right target area (two circles of 2cm in diameter in the upper left and right corners of the display). The target areas were separated by 11cm (center-to-center). In between trials, the two chess symbols in the center of the screen reminded participants which symbol called for a movement to the left (the one displayed on the left side) and which symbol called for a movement to the right (the one displayed on the right side). A written instruction between the two chess figures instructed the rule-compliance for the following trial. The starting position for the movement (a circle of 1cm in diameter) was located at the bottom center of the screen, 17cm from the middle of the two target positions at an angle of 31° to each side. Stimuli were presented against a light gray background (see Figure 1).

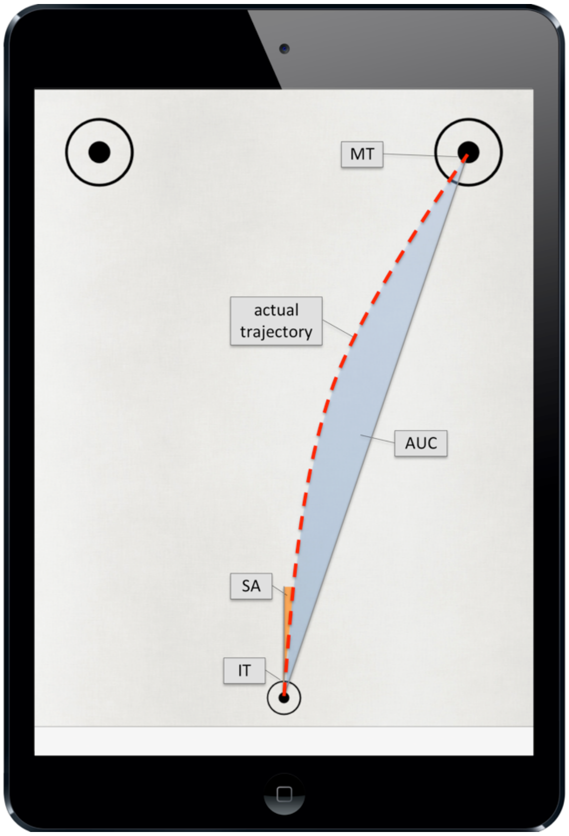

Figure 1. Setting of Experiments 1-3 and relevant measures.
Participants dragged their finger in a continuous movement from the starting area on the bottom of the screen to one of the two areas in the upper corners of the screen. In Experiment 1, they followed an instructed mapping rule in 75% of the trials whereas they violated the rule in 25% of the trials. In Experiment 2, participants performed an instructed primary task in 75% of the trials and performed a task with reversed mapping in 25% of the trials. In Experiment 3, participants followed an instructed mapping rule in 75% of the trials and had to invert the rule in 25% of the trials. IT (initiation time) was defined as the time from target onset to movement initiation, MT (movement time) as the time of movement execution. SA (starting angle) mirrors the angle of the movement trajectory against the vertical midline (orange) upon leaving the home area, AUC (area under the curve) measures the area between the actual trajectory and a straight line from start- to endpoint (blue).

Procedure.

Participants started each trial by touching the starting area with the index finger of the dominant hand. Immediately, a target symbol appeared in the center of the screen to indicate whether a movement to the left or a movement to the right had to be executed. Simultaneously, the reminder of the S-R mapping and the written instruction disappeared. Half of the participants were instructed to make a smooth finger movement to the left target area if the center showed a pawn symbol and to the right target area if it showed a king symbol. The other half of the participants was instructed with the opposite S-R mapping for counterbalancing. The target symbol disappeared as soon as the finger left the starting area. One out of four trials included a written instruction to rule violation instead of rule-based behavior before trial start (for example "👑 break the rule ♙", displayed in between trials). In these trials, the displayed mapping rule had to be violated; the response that a target required originally was now contraindicated. A trial ended when the finger was lifted from the touchscreen. Error feedback was displayed only if participants failed to hit one of the designated target areas. Participants were instructed to respond quickly and accurately; still the experiment was self-paced, so participants chose on their own when to start a trial and how long they took breaks in between blocks. Participants completed 12 blocks of 48 trials, with each of the target symbols presented equally often.

Results.

Preprocessing.

I analyzed four variables of each movement: The time from stimulus onset to movement initiation (initiation time; IT), the duration of the movement (movement time; MT), the angle between the trajectory and the vertical midline at response initiation (starting angle; SA) and the area between the actual trajectory and a straight line from start- to endpoint (area under the curve; AUC). IT therefore mirrors the speed of response selection and motor planning; MT, SA and AUC depict specific temporal and spatial parameters of the executed response. Positive values for AUC and smaller (or negative) values of SA indicate that a movement is attracted to the competing response alternative indicating a persisting influence of the original mapping rule.

IT was defined as the time that it takes for the finger to leave the starting area. From this point, x- and y-coordinates were recorded; MT was determined when the finger left the touchscreen. AUC and SA were computed from the time-normalized coordinate data of each trial by using custom MATLAB scripts (The Mathworks, Inc.). Movements to the left were mirrored at the vertical midline for all analyses. AUC was computed as the signed area relative to a straight line from start- to endpoint of the movement (positive values indicating attraction toward the opposite side, negative values indicating attraction toward the nearest edge of the display). SA was defined as the angle between the actual trajectory and the vertical midline (see Fig.1, negative values indicating attraction toward the opposite side, positive values indicating attraction toward the rule-based target area in case of rule violations).

Data selection and analyses.

For the following analyses, I omitted trials in which participants failed to act according to the instruction (3.5%) and the immediately following trials (3.0%). I also excluded trials in which participants failed to hit any of the two target-areas at all (2.5%). Trials were discarded as outliers if any of the measures (IT, MT, SA, AUC) deviated more than 2.5 standard deviations from the respective cell mean (6.3%). Each measure was then analyzed in a separate 2x2 ANOVA with current response type (rule-based vs. violation) and preceding response type as within-subject factors (see Figure 2). Additionally, repetition benefits for each measure and each response type were computed as the difference between switch and repetition trials. That is, repetition benefit for rule-based responses in IT were computed as (IT of rule-based responses after violation trial) minus (IT of rule-based responses after rule-based trial); all other repetition benefits were computed accordingly. Repetition benefits are only mentioned if significant.

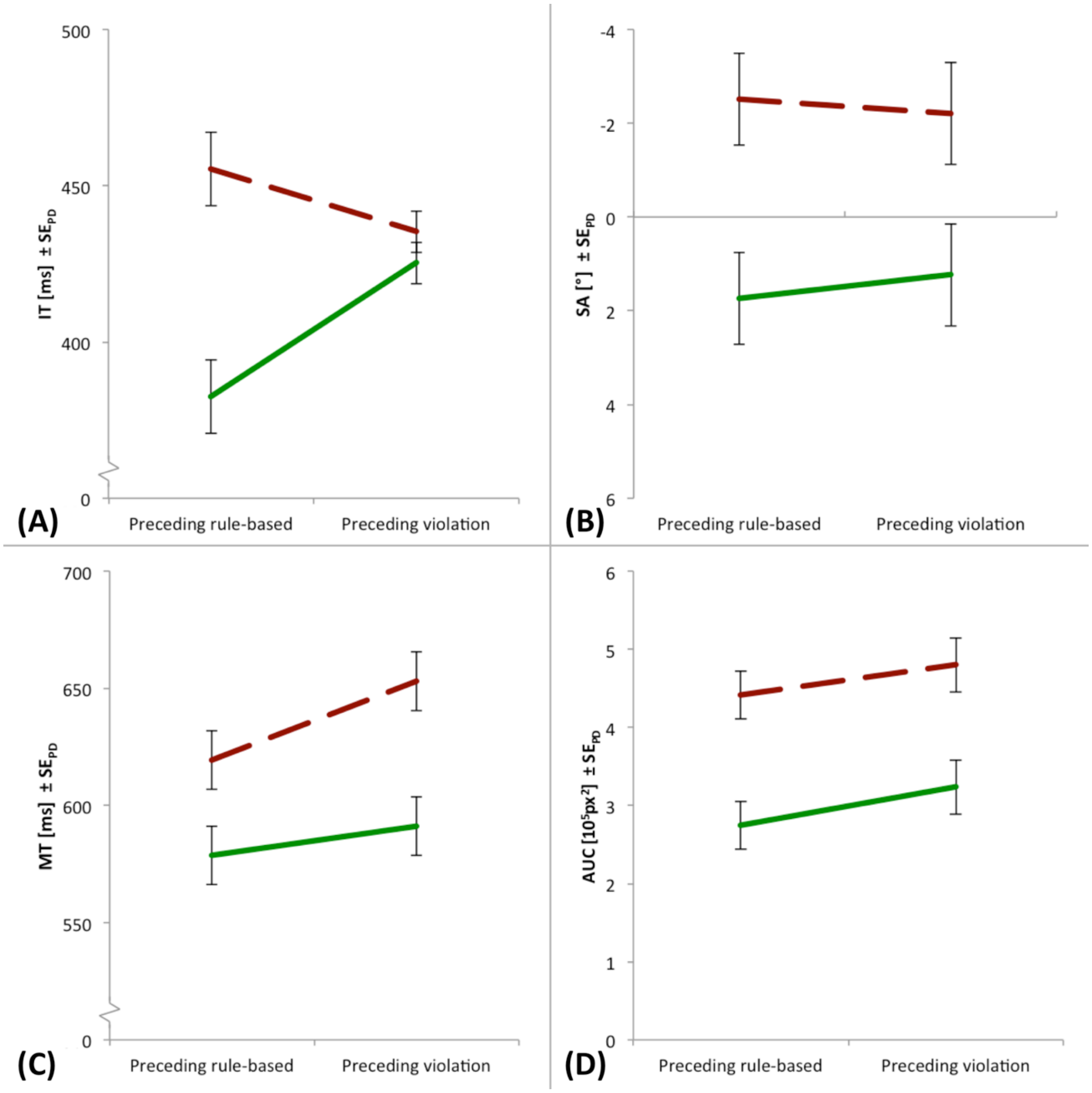

Figure 2. Results of Experiment 1.
Initiation times (ITs; panel A), starting angles (SA; panel B), movement times (MT; panel C) and areas under the curve (AUC; panel D) are plotted as a function of preceding response type (abscissa) and current response type (continuous line for rule-based responses; dashed line for violation responses). Error bars represent standard errors of paired differences (SE_{PD}), calculated separately for each instance of preceding response type (Pfister & Janczyk, 2013).

Initiation times.

A significant effect of current response type, $F(1,19) = 26.14$, $p < .001$, $\eta_p^2 = .58$, was driven by slower response initiation for violations (450ms) than for rule-based behavior (392ms). A similar effect emerged for preceding response type, *F*(1,19) = 10.43, *p* = .004, η_p^2 = .35, with slower responses following violations (428ms) compared to rule-based behavior (399ms). The interaction between preceding and current response type was also significant, *F*(1,19) = 32.34, *p* < .001, η_p^2 = .63, with a profound effect of current response type only after rule-based responses (Δ = 73ms), *t*(19) = 6.04, *p* < .001, *d* = 1.35, but not after violation responses (Δ = 10ms), *t*(19) = 1.47, *p* = .157, *d* = 0.33. Repetition benefits were smaller for violation responses (Δ = 20ms), *t*(19) = 3.54, *p* = .002, *d* = 0.79, compared to rule-based responses (Δ = 43ms), *t*(19) = 5.82, *p* = .002, *d* = 1.30 (Figure 2A).

Starting angles.

A significant effect of current response type, *F*(1,19) = 27.18, *p* < .001, η_p^2 = .59, indicated shallower initial trajectories for rule-based behavior (1.6°) compared to violations, which were steeper and initially directed to the opposite side (-2.2°). Neither preceding response type nor the interaction approached significance, *F*s < 1 (Figure 2B).

Movement times.

Response execution was slower for violations (628ms) than for rule-based behavior (581ms), *F*(1,19) = 29.52, *p* < .001, η_p^2 = .61. A significant effect of preceding response type, *F*(1,19) = 10.84, *p* = .004, η_p^2 = .36, further indicated slower movements following violations (608ms) compared to rule-based behavior (588ms). The interaction between preceding and current response type was not significant, *F*(1,19) = 1.48, *p* = .239, η_p^2 = .07. Rule-based responses produces repetition benefits (Δ = 12ms), *t*(19) =

2.19, p = .041, d = 0.49, while violation responses produced a negative repetition benefit (repetition costs), with repeated violations leading to slower movements compared to single instances (Δ = -34ms), t(19) = 2.28, p = .034, d = 0.51 (Figure 2C).

Areas under the curve.

A significant effect for current response type, F(1,19) = 46.56, p < .001, η_p^2 = .71, again indicated more curved trajectories for violations (45073px^2) than for rule-based behavior (28464px^2). The effect of preceding response type was marginally significant, F(1,19) = 4.24, p = .053, η_p^2 = .18, with descriptively more curved trajectories following violations (36359px^2) compared to rule-based behavior (31183px^2). The interaction between the two factors did not approach significance, F < 1. Repetition benefits were significant only for rule-based responses (Δ = 4891px^2), t(19) = 3.86, p = .001, d = 0.86 (Figure 2D).

Discussion.

In Experiment 1, I investigated the difficulties that rule violations pose on the acting agent. Replicating previous findings (Pfister et al., 2016) I found violation responses to be more effortful than rule-based responses. They took longer to initiate and execute, and their movement trajectory is heavily deflected towards the alternative target, suggestive of a continued influence of the original mapping rule.

The resulting pattern of ITs further suggests that repeatedly violating a rule facilitates the initiation of rule violations. This finding reminds of sequential patterns that are typically reported by studies on cognitive conflict and conflict adaptation (Botvinick et al., 2001; Gratton et al., 1992). This could ultimately suggest that the planning of a violation response is associated with cognitive conflict between the automatic rule-based and the violation response, and that this conflict lessens with previous violation responses.

Surprisingly, however, there were no sequential effects of rule violations on the actual movement trajectories. That is, the signature of rule violations on SA, MT, and AUC remained visible even after having committed a rule violation only a few seconds before. Participants thus appear not to adjust their response execution according to recent events after a rule violation (which is unusual even for response trajectories, cf. Scherbaum, Dshemuchadse, Fischer, & Goschke, 2010). Before drawing further conclusions from these findings, two experiments provide important control conditions to clarify the interpretation of these data.

2.2 Experiment 2.

Experiment 2 investigated whether the pattern of results observed in the preceding experiment is specific to rule violations or whether they represent just an instance of task switching (Monsell, 2003). I did this by employing the same task as in Experiment 1, but slightly varied the instructions. Instead of prompting participants to follow or break a given rule, I introduced two response mappings that were labeled “Task 1” and “Task 2”, with Task 2 being the reversed mapping of Task 1 (Schroder et al., 2012). As Task 2 was presented equally often as the violation prompt in Experiment 1, participants virtually had to employ the exact same responses in both experiments. In Experiment 2, however, participants were presented with two equally neutral and separate task sets, whereas the corresponding actions were labeled as deviant behavior in Experiment 1.

Methods.

Participants.

A new set of twenty participants was recruited (mean age = 21.8 years, *SD* = 4.2, 4 male, 2 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.

Apparatus, stimuli and procedure.

The experiment was mostly identical to the first experiment. But instead of instructing participants to break a given rule in one out of four trials, participants were asked to complete either “Task 1” (frequent task) or “Task 2” (infrequent task), with Task 2 consisting of the inverted S-R mapping of Task 1 and occurring in one out of four trials. This way, Experiment 2 required the exact same movements as Experiment 1, but instead of introducing rule-based and violation behavior, participants were presented with two separate and equally neutral response mappings.

Results.

Data treatment and analyses.

The data was treated exactly as in Experiment 1. I omitted trials in which participants failed to act according to the instructions (3.5%), the immediately following trials (3.0%) and trials in which participants failed to hit any of the two target-areas

(2.9%). Trials were discarded as outliers if any of the measures (IT, MT, SA, AUC) deviated more than 2.5 standard deviations from the respective cell mean (6.8%). The four measures were then analyzed in separate 2x2 ANOVAs with current response type (frequent task vs. infrequent task) and preceding response type as within-subject factors (Figure 3).

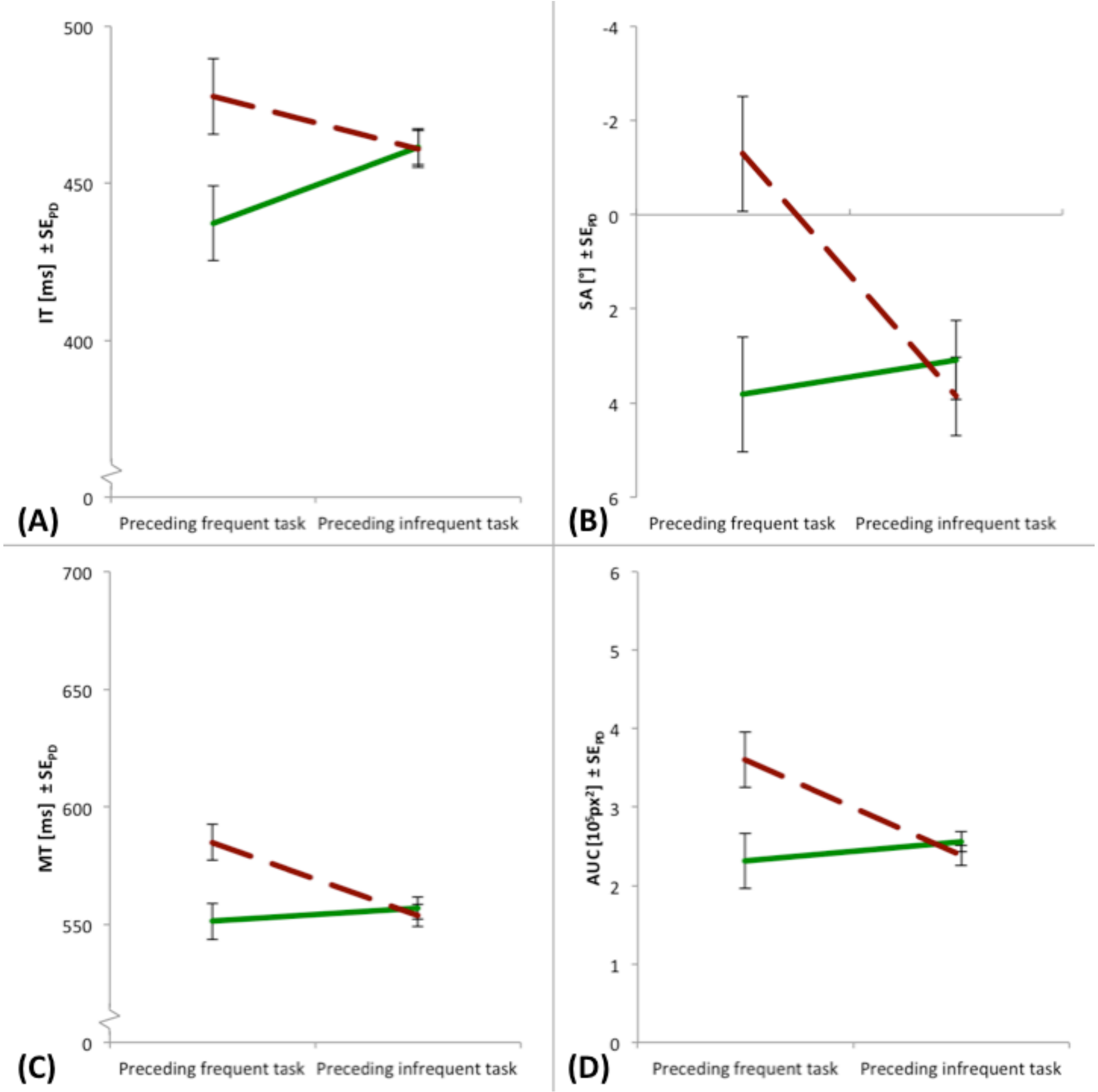

Figure 3. Results of Experiment 2.
Initiation times (ITs; panel A), starting angles (SA; panel B), movement times (MT; panel C) and areas under the curve (AUC; panel D) are plotted as a function of preceding response type (abscissa) and current response type (continuous line for responses to the frequent task; dashed line for

responses to the infrequent task). Error bars represent standard errors of paired differences (SE_{PD}), calculated separately for each instance of preceding response type (Pfister & Janczyk, 2013).

Initiation times.

A significant effect of current response type, $F(1,19) = 6.82$, $p = .017$, $\eta_p^2 = .26$, indicated slower response initiation for the infrequent task (472ms) than for the frequent task (442ms). The interaction between preceding response type and current response type was also significant, $F(1,19) = 11.84$, $p = .003$, $\eta_p^2 = .38$, with a pronounced effect of response type after frequent tasks (Δ = 40ms), $t(19) = 3.29$, $p = .004$, $d = 0.73$, and no response costs after infrequent tasks (Δ = 0ms), $t(19) = 0.12$, $p = .903$, $d = 0.03$. Repetition benefits were significant for the frequent task (Δ = 24ms), $t(19) = 2.74$, $p = .013$, $d = 0.61$, as well as for the infrequent task (Δ = 16ms), $t(19) = 1.86$, $p = .079$, $d = 0.42$ (Figure 3A).

Starting angles.

A significant effect of current response type, $F(1,19) = 7.06$, $p = .016$, $\eta_p^2 = .27$, was driven by shallower response initiation for the frequent task (3.7°) compared to the infrequent task (0.2°). A similar effect of preceding response type emerged, $F(1,19) = 12.67$, $p = .002$, $\eta_p^2 = .40$, with shallower responses following infrequent tasks (3.4°) compared to frequent tasks (2.8°). The interaction between preceding response type and current response type was also significant, $F(1,19) = 17.58$, $p < .001$, $\eta_p^2 = .48$, with effects of response type after frequent tasks (Δ = -5.1°), $t(19) = -4.07$, $p = .001$, $d = 0.91$, and no significant differences after infrequent tasks (Δ = 0.7°), $t(19) = 0.90$, $p =$

.328, *d* = 0.18. Repetition benefits were only significant for the infrequent task (Δ = 5.2°), *t*(19) = 5.01, *p* = .002, *d* = 1.12 (Figure 3B).

Movement times.

A significant effect of current response type emerged, *F*(1,19) = 9.20, *p* = .007, η_p^2 = .33, with slower movements on infrequent tasks (576ms) than on frequent tasks (552ms), as well as a significant effect of preceding response type, *F*(1,19) = 10.17, *p* = .005, η_p^2 = .35, with slightly faster movements following infrequent tasks (556ms) compared to frequent tasks (568ms). The interaction between preceding response type and current response type was also significant, *F*(1,19) = 18.85, *p* < .001, η_p^2 = .50, indicating a pronounced effect of response type after frequent tasks (Δ = 34ms), *t*(19) = 4.23, *p* < .001, *d* = 0.95, and no response costs after infrequent tasks (Δ = -3ms), *t*(19) = -0.65, *p* = .524, *d* = 0.14. Repetition benefits were significant for infrequent tasks (Δ = 31ms), *t*(19) = 4.08, *p* = .001, *d* = 0.91, and marginally significant for frequent tasks (Δ = 5ms), *t*(19) = 1.79, *p* = .090, *d* = 0.40 (Figure 3C).

Areas under the curve.

A significant effect for current response type, *F*(1,19) = 8.10, *p* = .010, η_p^2 = .30, was driven by more curved response execution on infrequent tasks (32464px^2) than on frequent tasks (23571px^2). A similar effect of preceding response type emerged, *F*(1,19) = 9.45, *p* = .006, η_p^2 = .33, with less curved response execution following infrequent tasks (24989px^2) compared to frequent tasks (25705px^2). The interaction between preceding response type and current response type was also significant, *F*(1,19) = 15.22, *p* = .001, η_p^2 = .45, with bigger effects of response type after frequent tasks (Δ = 12873px^2), t(19) = 3.59, $p < .001$, d = 0.80, and descriptively reversed response costs after infrequent tasks (Δ = -1757px^2), t(19) = -1.31, p = .204, d = 0.29. Repetition benefits were significant for infrequent tasks (Δ = 12154px^2), t(19) = 3.78, p = .001, d = 0.85, and marginally significant for frequent tasks (Δ = 2477px^2), t(19) = 1.93, p = .069, d = 0.43 (Figure 3D).

Discussion.

In Experiment 2, I slightly changed the task instructions, as compared to Experiment 1: instead of instructing one task set that had to be violated, I provided participants with two separate task sets that called for the exact same behavior as in Experiment 1.

I found that infrequent, rule-based behavior still differs from frequent, rule-based behavior with infrequent behavior being more effortful in both, planning and execution. Responses based on the Task 2 rule are also deflected to the opposite side, which could indicate an influence of the dominant rule of Task 1, or reflect task-switching effects between the two instructed task sets (Monsell, 2003, see the Preliminary Discussion of this chapter for a more thorough discussion). As of now, I can conclude that task-switching effects and the presentation frequencies that I used here could potentially account for the signature that I found to be associated with rule violations. However, the effects observed in Experiment 2 were substantially smaller than those observed in Experiment 1 and came with a different pattern of adaptations according to recent events (for a corresponding between-experiment analysis, see Paragraph 2.4). These diverging results might be driven by two procedural differences:

The labeling of the infrequent response as rule violation versus an alternative but rule-conform option for one, and the instruction in terms of one versus two task sets for another. Both differences might partly account for these diverging results and Experiment 3 therefore aimed at clarifying the role of both contributions.

2.3 Experiment 3.

In Experiment 2, participants were instructed with two separate task sets for both conditions (frequent vs. infrequent task), whereas Experiment 1 only employed one task set (rule-based), while rule violations had to be derived from the instructed one. To test whether this difference in available task sets can account for the specific behavioral signatures of rule violations compared to task-switches, Experiment 3 provided an additional control condition that only provided one task set, and participants had to derive the alternative responses from this task set by inversion (Wason, 1959; Wegner, 2009). Compared to Experiment 1, I now employed an instruction that put less emphasis on the deviating nature of the infrequent task, but offered a more neutral response alternative.

Methods.

Participants.

A new set of twenty participants was recruited (mean age = 23.4 years, *SD* = 2.6, 4 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.

Apparatus, stimuli and procedure.

The experiment was mostly identical to the first experiment. But instead of instructing participants to break a given rule in one out of four trials, participants were asked to either “follow the standard rule” or “invert the rule”. This way, Experiment 3 required the exact same movements as Experiments 1 and 2, but, as in Experiment 1, now only instructed one task set. At the same time, I took care to instruct the inversion as part of the mapping rule rather than labeling the behavior as violation as I had done in Experiment 1.

Results.

Data treatment and analyses.

The data was treated exactly as in Experiments 1 and 2. I omitted trials in which participants failed to act according to the instructions (3.4%), the immediately following trials (3.3%) and trials in which participants failed to hit any of the two target-areas (2.3%). Trials were discarded as outliers if any of the measures (IT, MT, SA, AUC) deviated more than 2.5 standard deviations from the respective cell mean (6.3%). The four measures were then analyzed in separate 2x2 ANOVAs with current response type (standard vs. inverted) and preceding response type as within-subject factors (see Figure 4).

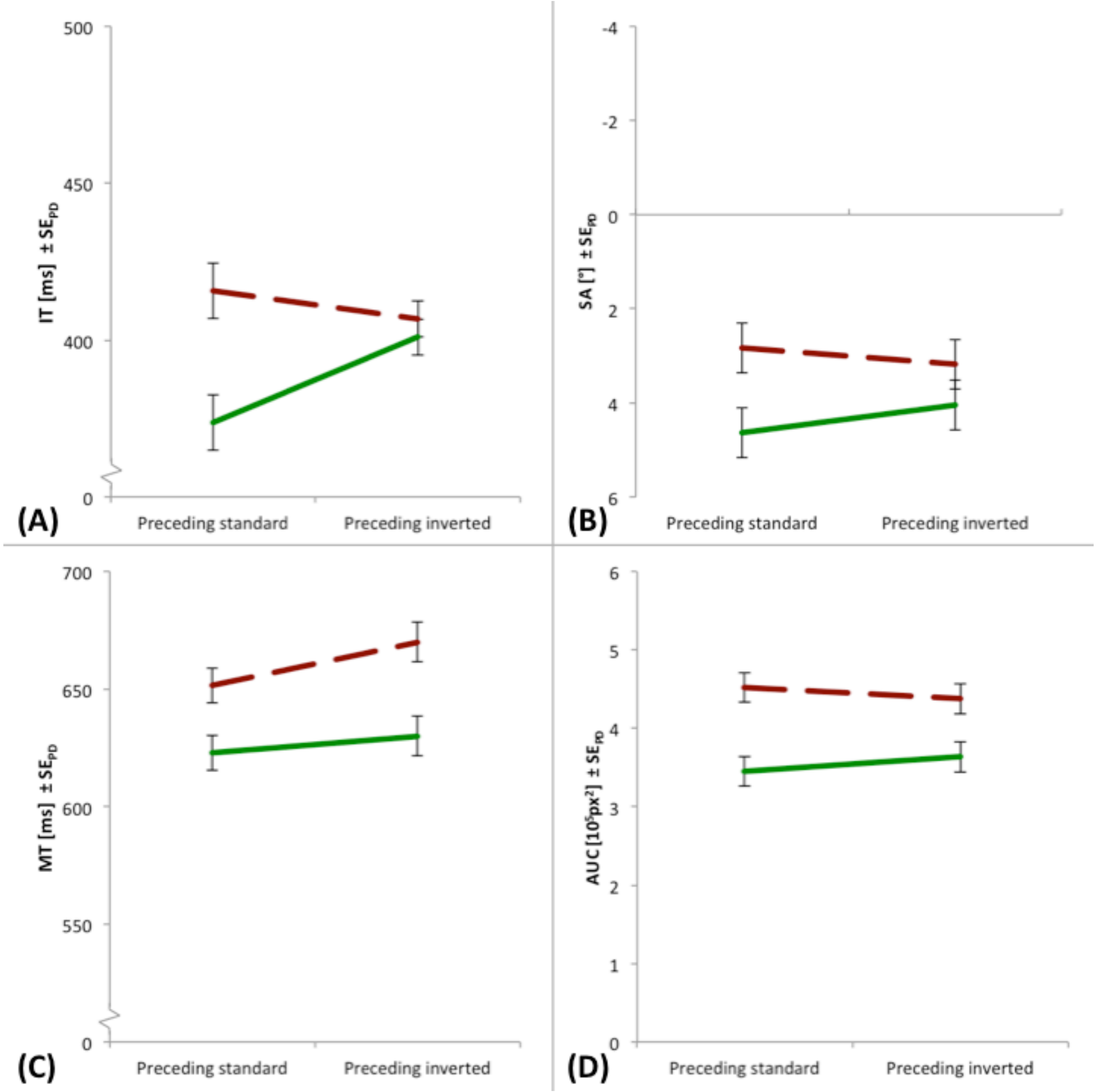

Figure 4. Results of Experiment 3.

Initiation times (ITs; panel A), starting angles (SA; panel B), movement times (MT; panel C) and areas under the curve (AUC; panel D) are plotted as a function of preceding response type (abscissa) and current response type (continuous line for standard responses according to the original rule; dashed line for inverted responses). Error bars represent standard errors of paired differences (SE_{PD}), calculated separately for each instance of preceding response type (Pfister & Janczyk, 2013).

Initiation times.

A significant effect of current response type, *F*(1,19) = 14.68, *p* = .001, η_p^2 = .43, was driven by slower response initiation for inversions (411ms) than for standard responses (387ms). A similar effect emerged for preceding response type, *F*(1,19) = 5.08, *p* = .036, η_p^2 = .21, with slower responses following inversions (404ms) compared to standard responses (395ms). The interaction between preceding and current response type was also significant, *F*(1,19) = 17.04, *p* = .001, η_p^2 = .47, with a profound effect of current response type only after standard responses (Δ = 42ms), *t*(19) = 4.65, *p* < .001, *d* = 1.04, but not after inverted responses (Δ = 6ms), *t*(19) = 0.98, *p* = .339, *d* = 0.22. Repetition benefits were smaller for inverted responses (Δ = 8ms), *t*(19) = 1.41, *p* = .173, *d* = 0.32, compared to standard responses (Δ = 27ms), *t*(19) = 4.77, *p* < .001, *d* = 1.07 (Figure 4A).

Starting angles.

A significant effect of current response type, *F*(1,19) = 13.07, *p* = .002, η_p^2 = .41, indicated shallower initial trajectories for standard responses (4.4°) compared to inversions (3.0°). Neither preceding response type nor the interaction approached significance, *F*s < 1.39, *p*s > .254 (Figure 4B).

Movement times.

Response execution was slower for inversions (661ms) than for standard responses (627ms), *F*(1,19) = 23.85, *p* < .001, η_p^2 = .56. A significant effect of preceding response type, *F*(1,19) = 7.96, *p* = .011, η_p^2 = .30, further indicated slower movements following inversions (650ms) compared to standard responses (637ms). The interaction between preceding and current response type was not significant, *F*(1,19) = 1.91, *p* = .183, η_p^2 = .09. Unexpectedly, inverted responses produced repetition costs rather than benefits (Δ = -18ms), *t*(19) = 2.74, *p* = .034, *d* = 0.61 (Figure 4C).

Areas under the curve.

A significant effect for current response type, $F(1,19) = 33.69$, $p < .001$, $\eta_p^2 = .64$, again indicated more curved trajectories for inverted responses (44494px^2) compared to standard responses (35423px^2). No other effects reached significance, *F*s < 2.02, *p*s $> .172$. Standard responses produced marginally significant repetition benefits ($\Delta = 1850$px^2), $t(19) = 2.88$, $p = .075$, $d = 0.42$ (Figure 4D).

Discussion.

In Experiment 3, I tested whether the instruction of a single task set that had to be inverted could account for the pattern of data that I found for violation responses in Experiment 1. And indeed, I again found that responses based on the inverted task set were slower and more attracted to the competing response alternative. Moreover, I was able to replicate the sequential adaptation effect of ITs and the additive effect of SAs, MTs and AUCs that indicate that the selection and planning of an inverted response becomes more efficient with previous experience, while the execution of these responses does not.

To compare the size of the response costs and adaptation effects that come with rule-violations compared to task-switches and inversions, I conducted between-experiments analyses.

2.4 Between-experiment analyses.

Results.

For all between-experiment analyses, I conducted ANOVAs on the immediate effects of the experimental manipulation and on the corresponding sequential effects. Immediate effects were computed as the mean differences between the two current response types (violation/infrequent/inverted minus rule-based/frequent/standard) and they were analyzed via planned contrasts that pitted Experiment 1 against Experiment 2, and Experiment 1 against Experiment 3, and I finally tested the contrast between Experiments 2 and 3.

Sequential effects were computed as the differences between the effects after the two response types (violation-/infrequency-/inversion-effect after rule-based/frequent/standard responses minus effect after deviant responses) and they were analyzed via planned contrasts that compared the adaptation effect between Experiment 1 against Experiment 2, and those that compared the adaptation effect between Experiment 1 and Experiment 3. I finally compared the adaptation between Experiments 2 and 3. Because I expected the effects of rule violations (Exp. 1) to exceed the effects of inversion (Exp. 3) and, likewise the effects of inversion to exceed the effects of task frequency (Exp. 2), I report the following contrast estimates as one-tailed.

Initiation times.

Regarding the immediate effects, the comparison of the effect size of Experiment 1 (Δ = 40ms) against Experiment 2 (Δ = 19ms) was significant, $t_{1/2}(57)$ =

2.10, p = .020, d = 0.63, while the contrast between Experiment 1 and Experiment 3 (Δ = 24ms) only produced a marginally significant effect, $t_{1/3}(57) = 1.58$, p = .060, d = 0.51. The comparison between Experiments 2 and 3 was not significant, $t_{2/3} < 1$.

For sequential effects, only the comparison of Experiment 1 (Δ = 60ms) against Experiment 3 (Δ = 36ms) was marginally significant, $t_{1/3}(57) = 1.42$, p = .085, d = 0.52, the adaptation effect in Experiment 2 (Δ = 40ms) differed from neither experiment, $|t|$s < 1.17, ps > .123.

Starting angles.

Regarding the immediate effects, the comparison of the effect size of Experiment 1 (Δ = -4.27°) against Experiment 2 (Δ = -1.34°) was significant, $t_{1/2}(57)$ = 1.78, p = .041, d = 0.86, as was the contrast between Experiment 1 and Experiment 3 (Δ = -2.43°), $t_{1/3}(57) = 2.83$, p = .003, d = 0.71. The comparison between Experiments 2 and 3 was not significant, $t_{2/3}(57) = 1.05$, p = .144, d = 0.38.

For sequential effects, the comparison of Experiment 1 (Δ = -0.88°) against Experiment 2 (Δ = -5.09°) was significant, $t_{1/2}(57) = 2.57$, p = .007, d = 0.74, as was the contrast between Experiment 2 and Experiment 3 (Δ = -0.93°), $t_{2/3}(57) = 2.54$, p = .007, d = 0.98. The contrast between Experiment 1 and Experiment 3 was not significant, $t_{1/3}$ < 1.

Movement times.

Regarding the immediate effects, the comparison of the effect size of Experiment 1 (Δ = 43ms) against Experiment 2 (Δ = 10ms) was significant, $t_{1/2}(57)$ =

3.22, p = .001, d = 1.08, as was the contrast between Experiment 2 and Experiment 3 (Δ = 34ms), $t_{2/3}(57)$ = 2.32, p = .012, d = 0.92. The comparison between Experiments 1 and 3 was not significant, $t_{1/3}(57) < 1$.

For sequential effects, the comparison of Experiment 1 (Δ = -22ms) against Experiment 2 (Δ = 20ms) was significant, $t_{1/2}(57)$ = 2.90, p = .003, d = 0.91, as was the contrast between Experiment 2 and Experiment 3 (Δ = -11ms), $t_{2/3}(57)$ = 2.17, p = .017, d = 0.64. The contrast between Experiment 1 and Experiment 3 was not significant, $t_{1/3} < 1$.

Areas under the curve.

Regarding the immediate effects, the comparison of the effect size of Experiment 1 (Δ = 14582px^2) against Experiment 2 (Δ = 5200px^2) was significant, $t_{1/2}(57)$ = 3.24, p = .001, d = 0.94, as was the contrast between Experiment 1 and Experiment 3 (Δ = 9071px^2), $t_{1/3}(57)$ = 2.90, p = .003, d = 1.09. The comparison between Experiments 2 and 3 was marginally significant, $t_{2/3}(57)$ = 1.33, p = .094, d = 0.55.

For sequential effects, the comparison of Experiment 1 (Δ = 342px^2) against Experiment 2 (Δ = 10789px^2) was significant, $t_{1/2}(57)$ = 2.45, p = .009, d = 0.70, as was the contrast between Experiment 2 and Experiment 3 (Δ = 3273px^2), $t_{2/3}(57)$ = 1.78, p = .042, d = 0.60. The contrast between Experiment 1 and Experiment 3 was not significant, $t_{1/3} < 1$.

Discussion.

The direct comparison of the experiments allowed us to scrutinize the impact of the rule violation instructions as compared to both control conditions. Results showed that the impact of violations (Experiment 1) was more detrimental than the impact of task-switches or inversions (Experiments 2 and 3). These differences, especially those between Experiment 1 and Experiment 3 can be solely attributed to the labeling of the actions as rule violations in the former experiment but not in the latter. It therefore seems as if simply relabeling the deviant response as an inversion instead of a violation might be an effective way to minimize the impact of deviant responses.

There was also an apparent difference in the adaptation based on the previous response type between the experimental groups. While participants in the task-switching group were able to take parameters of the previous trial into account to adjust their performance on the current trial, this was only partly the case for both the violation group and the inversion group. Participants of those groups could adapt their response selection according to recent events, but failed to do so when it came to planning and executing the corresponding response. Here, the second violation or inversion in a sequence was just as slow and contorted, if not more, than the first one.

2.5 Preliminary Discussion.

In Experiments 1-3, I investigated the impact that rule violations pose on the acting agent even when the rule in question is a simple S-R rule that was instantiated by instruction. I employed a two-dimensional finger-tracking task in which participants had to drag their finger from a starting area to one of two target areas on a touchscreen according to a pre-specified rule.

In Experiment 1, I probed for the behavioral signature of rule violations regarding temporal and spatial parameters of the executed responses. In addition to analyzing how current rule violations influence participants’ behavior, I also took previous experience with rule violations into account. I found a profound impact of current rule violations in both, temporal and spatial measures. Rule violations took longer to be initiated and executed, and their movement trajectories were heavily bent towards the opposite side, which could indicate an ongoing influence of the original rule. And even though repeated rule violations were initiated with greater ease, I did not find any modulating influence of preceding rule-compliance for measures capitalizing on response execution: repeated rule violations were as strongly affected by the original mapping rule as singular events of a rule violation.

In Experiment 2, I isolated the effect of rule violations by means of a first control condition by creating a task-switching experiment that called for the exact same behavior as Experiment 1. To this end, I instructed participants to respond in a frequent “Task 1” in most trials but prompted them to respond in an infrequent “Task 2” that was the reverse of the frequent task. Again, I found a strong temporal and spatial impact of the infrequent task set, but this time I also observed a profound sequential modulation: for repeated reversed responses, movement trajectories were as efficient as for responses based on the frequent task set.

Finally, in Experiment 3, I tested for inversions of an instructed task set and found similar sequential effects for the inverted responses compared to standard responses as I did for violations, whereas the overall impact of inversions was less pronounced than the impact of violations.

Before drawing conclusions about the possible mechanisms underlying rule violation, I would like to give a structured comparison of the instructions used. For one, the instructions differed as to whether one or two task sets were instructed, with Experiment 1 and 3 featuring only one task set and Experiment 2 featuring two distinct task sets. The main difference between the instructions involving one and the instructions involving two task sets is that, while two separate task sets allow for adaptation to the infrequent task, instructing only one task set seems to hinder participants from adjusting their performance based on recent events. The task sets for violations (Exp. 1) and inversions (Exp. 3) do not seem to be represented independently, but dependent on the frequent instruction. This representation *with strings attached* might cause the sequential modulation that I obtained here, which will be explained in more detail in the following sections.

For another, the instructions of Experiment 1 and Experiment 3 differed as to whether I emphasized that the infrequent task was not in accordance with the rule of the frequent task. For the violation instructions (Exp. 1), I specifically highlighted that the violation behavior ran counter to the original rule, whereas I did not use such an emphasis for the inversion instructions (Exp. 3). I will come back to this distinction in the following discussion.

Rule violations and cognitive conflict.

The pronounced effects for rule violations as compared to rule-based responses accord with the idea that participants experience ongoing cognitive conflict during rule violations. Assuming that rules trigger automatic compliance (Asch, 1965; Milgram, 1963), rule violations inherently provoke the activation of two response alternatives: the rule-based, automatic response and the planned violation response (for a recent perspective, see Kim & Hommel, 2015). The solution of this response conflict takes time, which explains the prolonged response initiation of violation responses. The conflict is not resolved completely, however, because the automatic, rule-based response still shapes violation behavior. This accounts for the ironic effect that rule violations are heavily influenced by the rule that participants try to violate (Wegner, 2009): They are confronted with a rule, however arbitrary, and they activate the corresponding response. Violations inherently include the recollection of the rule that has to be violated, so the activation of the rule-based response is strong enough that it cannot be suppressed entirely (Pfister et al., 2016). Thereby, I was able to isolate cognitive mechanisms that process (non-) conformity even in non-social settings.

Rule violations as a derived task set.

What further differentiates violation behavior is how parameters of previous responses are taken into account. Alternatively, or in addition to the notion of cognitive conflict, the two response alternatives might be seen not as instances of the same task, but rather as distinct task sets. In this view, the observed adaptation after responses based on the infrequent task set might be taken to indicate task-switching effects (Allport, Styles, & Hsieh, 1994; Monsell, 2003; Rogers & Monsell, 1995). When simply switching to the infrequent task set, further responses based on this task set are easy, fast and efficient; parameters of previous responses are used to speed up the current response. But when violating rules, these parameters are not taken into account; a series of repeated rule violations poses repeated difficulty on the agent. This striking pattern of results was also found when participants were asked to invert a rule. If the current task set has to be derived from an instructed one and is not based on a separate, instructed task set (which is true for both, violations and inversions), this derived task set seems to be either (1) short lived and decays immediately, or (2) is not instituted as strongly as if it were an instructed task set or, moreover, (3) it could be used and attenuated immediately after finishing response execution, indicating repeated effort for repeated derivations of the currently relevant task set. Consequently, committing a violation response would always entail an immediate, endogenous switch back to rule-based responses (Arrington, & Logan, 2004; Arrington, Weaver, & Pauker, 2010; Kessler, Shencar, & Meiran, 2009; Liefooghe, Demanet, & Vandierendonck, 2010; Vandierendonck, Demanet, Liefooghe, & Verbruggen, 2012), as the derived task set for violations (and inversions) might not be as easily accessible or maintainable as an instructed task set.

A two-step activation model.

This notion seems to be supported by research on how negations (and inversions) are represented in the cognitive system. Indeed, negations are assumed to be represented and retrieved in two separate steps: The non-negated concept is retrieved at first, followed by applying the negation for each individual retrieval process (Clark & Chase, 1972, 1974; Gilbert, 1991; Strack & Deutsch, 2004; Wegner, Coulton, &

Wenzlaff, 1985). This holds true especially for negations that do not have a graspable meaning on their own, whereas negations seems to have only limited impact if participants can form an alternative representation (Hasson, Simmons, & Todorov, 2005; Fillenbaum, 1966; Mayo, Schul, & Burnstein, 2004). Even though such an alternative representation could have been formed in the present experiments (akin to the two task sets that I instructed in Experiment 2), the violation label might have worked against this tendency. This line of thought will be addressed in the next chapter.

In any case, as violations and inversions produce the same sequential modulations, I propose that it is safe to say that the inversion of a rule (or in a broader picture: derivation, manipulation, negation, reformulation or modification of an existing task set) is one of the cognitive mechanisms that drive the behavioral parameters of rule violations. While this process partly explains the effects of rule violations, it does not drive them exclusively. Even though the sequential adaptation does not differ between these conditions, the burdens that violations pose on the agent at the moment of response execution exceed those of inversions.

Therefore, I conclude that in addition to this “cold cognition” explanation, it could be that “violate the rule” instructions have an emotional component (“hot cognition”), and participants might exhibit an active tendency to steer away from mental representations reflecting (socially) unwanted behavior. In this view, rule violations might be best described as an *inversion of an existing rule with an add-on*. Which components this add-on might include will be addressed in the next chapter.

3. Why.

Based on the results of Experiments 1-3, we can assume that rule violations pose a special instance of a rule transformation (negation, inversion) with an add-on that makes them more difficult to plan and execute than their neutrally labeled counterparts. In this chapter, we will investigate why that might be the case.

To do so, I 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[2]. Such evaluative processes have recently been documented for the commission of 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

[2] The data of Experiments 5-7 is published in Wirth, Foerster, Rendel, Kunde, & Pfister, 2017, in *Cognition & Emotion*.

in general seem to be linked to a negative emotional component, i.e. 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 categorized as positive or negative (Dreisbach & Fischer, 2012). Positive targets were categorized faster when preceded by a congruent Stroop word as compared to incongruent Stroop words, and negative targets were categorized faster when preceded by an incongruent Stroop word as compared to congruent Stroop words. Based on this finding, the authors argued that the conflict in the Stroop words is coded as an aversive signal, therefore incongruent Stroop words can sensitize toward negative targets in the subsequent task. Assuming that rule violations come with cognitive conflict (as do incongruent Stroop words), I would predict violations to sensitize toward negative events. At the same time, however, successful resolution of cognitive conflict has been demonstrated to represent a reward signal (Schouppe et al., 2015). Successfully overcoming a rule-based action tendency, i.e., successfully committing a rule violation, would thus predict violations to sensitize toward positive events instead. Experiment 4 attempted to decide between both hypotheses.

Based on the findings of Experiment 4, I set out to explore a further possible evaluative process triggered by rule violations. In addition to cognitive conflict during actually committing violations, rule violations are further inherently related to authorities that are often able to punish potential violators to ensure rule-conformity (e.g., parents during childhood and adolescence, superordinates at the workplace, or officials such as police officers). I therefore tested whether rule-violations would not only sensitize toward emotional stimuli but also toward stimuli that are related to authorities (Experiment 5), and how potential effects would compare against rule inversions that are not labeled as violations (Experiment 6). Finally, I address a possible confound that might be introduced by the labeling (Experiment 7).

3.1 Experiment 4.

In Experiment 4, I probed for the hypothesized affective component of rule violations. Following the logic that I applied in Chapter 2, I introduced a simple and arbitrary 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. Violations, here, only required a negation of the instructed mapping rule. I deliberately designed these violation responses to not entail negative feedback or punishment for breaking the rule. So the only differences between rule-based responses and violation responses were labelling of the responses as either rule-based or violation, presentation frequency, and the additional negation of the instructed mapping rule in case of violations. Consequently, any differences between rule-based and violation responses can be attributed to either of these factors, which allows for a subsequent breakdown of these components in the following experiments.

The experiment consisted of two tasks: A violation task, where an instructed mapping rule had to be violated in a fraction of trials, and a valence task, where target words had to be categorized as either positive or negative. If we assume rule violations to have an affective component, the commission of a violation should alter the performance in the second task (Aarts et al., 2012; Dreisbach & Fischer, 2012). More specifically, the following hypotheses can be derived: If the commission of a violation represents a negative event, subsequent negative targets should be categorized faster as compared to preceding rule-based responses. Alternatively, the successful commission of a violation might even be a positive event, just like the successful resolution of difficult and conflicting trials has been argued to represent a reward signal (Schouppe et al., 2015). Consequently, positive targets might be categorized faster after a violation than after a rule-based response, because the successful completion of the more demanding and difficult response is considered a positive event.

Methods.

Participants.

Twenty-four participants were recruited (mean age = 25.4 years, *SD* = 5.8, 11 male, 2 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. Based on previous studies that used a similar design, I estimated the expected effect size as medium (Dreisbach & Fischer, 2012, report *d* = 0.52 for the effect of negative targets after incongruent trials compared to after congruent trials). Consequently, I chose a sample of 24 participants, as this should provide a power of 0.80 for a medium-sized effect. The observed effect size of Experiment 4 then served as an estimate to calculate the sample size for the following experiments.

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. The Prime task was cued by two written instructions, “follow the rule” or “break the rule” in the center of the screen before each trial started. Stimuli for the Probe task consisted of 12 nouns – 6 positive and 6 negative – that were pre-rated in a pilot study. These probe words had to be categorized as positive or negative with the d- and k-key. All cues and targets were presented centrally in black against a white screen.

Pilot study.

To arrive at a standardized stimulus set, I asked an independent sample of 15 participants to rate a total of 168 words regarding their valence and their relation to authority. Ratings were given on a nine-point scale with verbal labels at the end points of the scale: (1) extremely negative to (9) extremely positive for valence, and (1) no relation to authority to (9) very strong relation to authority. See the Appendix for the mean ratings of all the Probe target words that were used in the present experiments.

For Experiment 4, I selected 6 negative and 6 positive words with low authority-relation from this item pool. The probe words were the German equivalents of present, luck, sun, peace, gain and benefit ($M_{Valence}$ = 7.70, $SD_{Valence}$ = 1.01) for the positive target words and corpse, accident, lie, bankruptcy, betrayal and disloyalty ($M_{Valence}$ = 1.78, $SD_{Valence =}$ 0.30) 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.

Procedure.

Each trial started with a cue that instructed participants to either follow or break the instructed mapping rule of the Prime task. Per instruction, 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, this instructed mapping rule had to be violated (“break the rule”). This cue was displayed for 500ms, 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 2000ms and disappeared as soon as a response was given. A blank screen of 100ms separated the Prime from the Probe.

For the Probe task, a randomly chosen target word appeared for a maximum of 2000ms and had to be categorized as positive or negative by the press of a key. 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 500ms.

Participants completed two short training blocks where the two tasks were presented separately to reinforce the instructed mapping rules (one block with 24 Prime trials, one block with 12 Probe trials). After that, participants completed 4 experimental blocks of 96 trials where Prime and Probe were interleaved as described above (see Figure 5).

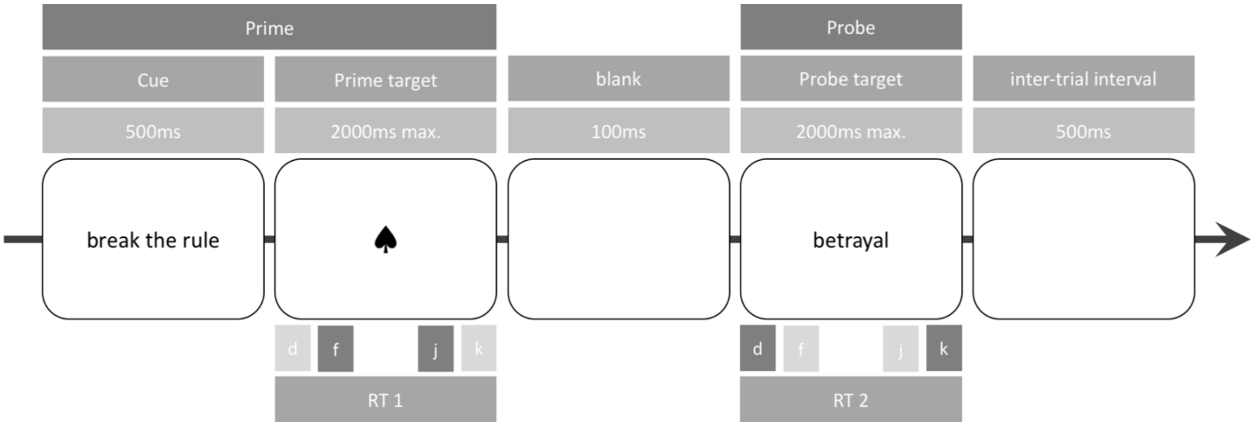

Figure 5. Setup of Experiments 4-7.
The Prime task consisted of a Cue that informed whether the instructed mapping rule had to be used or violated in the following trial. After 500ms, the Prime target appeared and called for responses with the f- or j-key (RT1). The Probe target appeared after a blank of 100ms. It had to be categorized as positive or negative with the d- or k-key (RT2), and I analyzed the impact of the Prime response type on the following valence categorization.

Results.

Data selection and analyses.

For the following analyses, I only used trials from the experimental blocks. I omitted trials in which participants failed to act according to the instruction (Prime:

8.5%, with more commission errors for violations than for rule-based responses, *t*(23) = 6.12, *p* < .001, *d* = 0.97; Probe: 5.8%, irrespective of Probe target valence, *t*(23) = 0.97, *p* = .341, *d* = 0.05) and the immediately following trials (Prime: 7.2%, Probe: 5.3%). Trials were discarded as outliers if any of the measures (RT1, RT2) deviated more than 2.5 standard deviations from the respective cell mean (4.7%). RT1 was then analyzed in an analysis of variance (ANOVA) with Prime response type (rule-based vs. violation) as within-subjects factor, whereas RT2 was analyzed in a 2×2 ANOVA with Prime response type (rule-based vs. violation) and Probe target valence (positive vs. negative) as within-subjects factors (see Figure 6 for the corresponding mean RTs).

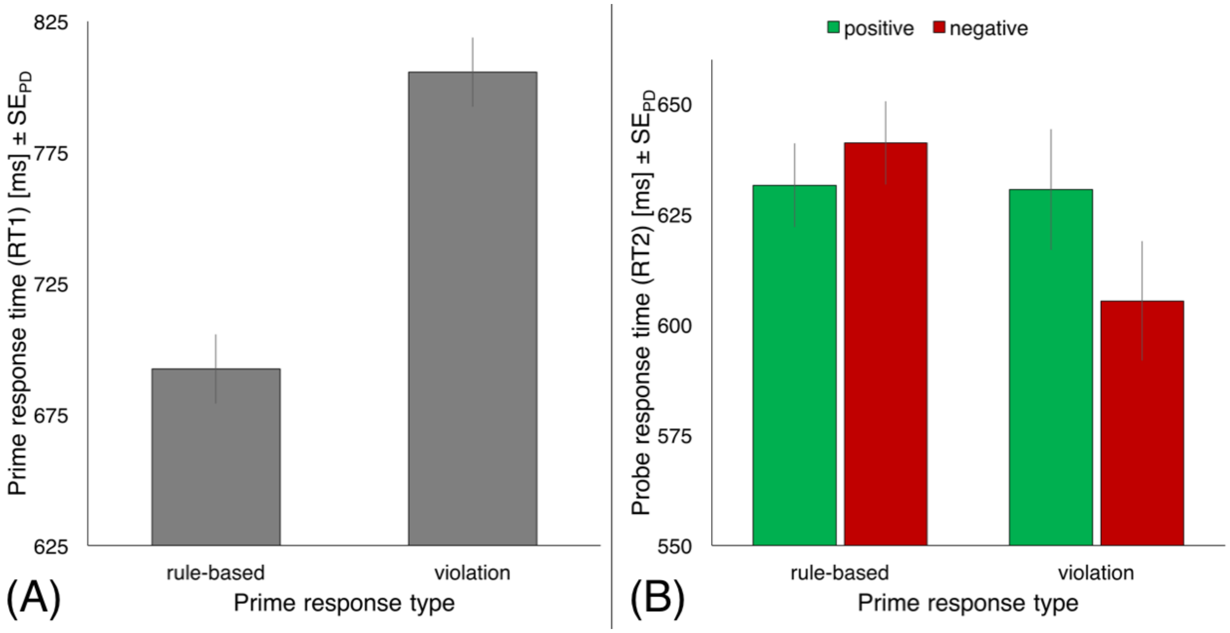

Figure 6. Results of Experiment 4.
Prime response times (RT1; Panel A) and Probe response times (RT2; Panel B) as a function of Prime response type (abscissa) and Probe target valence (left, green bars for positive targets; right, red bars for negative targets). Error bars represent standard errors of paired differences (SE_{PD}) that were calculated separately for each instance of Prime response type in Panel B (Pfister & Janczyk, 2013).

Prime responses.

A significant effect of Prime response type emerged, $F(1,23) = 75.74$, $p < .001$, $\eta_p^2 = .77$, driven by slower responses for violations (806ms) than for rule-based responses (692ms, Figure 6A).

Probe responses.

A significant effect of Prime response type, $F(1,23) = 6.86$, $p = .015$, $\eta_p^2 = .23$, indicated slower responses after rule-based behavior (636ms) compared to violations (617ms). There was an interaction between Prime response type and Probe target valence, $F(1,23) = 13.91$, $p = .001$, $\eta_p^2 = .37$, with response costs for negative targets after rule-based responses ($\Delta = 10$ms), and a response benefit for negative targets after rule violations ($\Delta = -25$ms, Figure 6B).

Power analysis.

Experiment 1 did not only serve as a confirmation that rule-violations do indeed entail an affective component, but it also served as an estimate for the effect size that could be obtained in this design. I achieved an effect size of $d = 0.76$ for the critical interaction in the Probe task between the Prime response type and the Probe target valence. This returns a statistical power of .94 when using a corresponding sample size of $n = 24$. To elevate the power of the next experiments above .95, I therefore increased the sample size to 28 participants for Experiments 5-7.

Discussion.

In Experiment 4, I probed for an affective component of rule violations. By employing a Prime-Probe-design with a violation task as the Prime and a valence task as the Probe, I could test whether having committed a violation modulated subsequent categorization of positive and negative target words. And indeed I found that after a violation response, negative target words are categorized faster than after a rule-based response. Rule violations seem to be considered a negative event. However, this result is not specific to rule violations, the same pattern of results is observed for trials with preceding errors (Aarts et al., 2012) and cognitive conflict (Dreisbach & Fischer, 2012). Rule violations have been argued to represent a special instance of a conflict task, with conflict between the rule-based, default response and the currently required response (see Chapter 2). Based on this result, we still cannot disentangle the cognitive processes of rule violations from those of cognitive conflicts.

At any rate, this result helps to distinguish between the two competing hypotheses that were raised in the Introduction. Despite violation responses being more difficult and demanding, the successful resolution of these trials does not seem to trigger a reward signal (Schouppe et al., 2015), but instead promotes the detection of negative stimuli. Although I did not include any feedback in Experiment 1, this heightened attention towards negative stimuli could reflect even latent expectations of punishment after violations (Pfister et al., 2016). Such latent expectations might derive from the fact that rules are usually instituted by authorities (parents, teachers, police), and violations are usually punished by those authority figures. In addition to an affective component, rule violations might consequently also sensitize toward authority-related stimuli that announce some sort of punishment. Therefore, in Experiment 5, I additionally tested for an authority-related component of rule violations.

3.2 Experiment 5.

In Experiment 4, I found that rule violations do indeed modulate the categorization of valent target words, giving rise to the assumption that violations are considered a negative event. This pattern of result is, however, not specific to violations and might simply be due to the conflicting nature of these responses: They are more difficult than simple rule-based responses because they require the inhibition of the default response to allow for the deviant response. To understand what sets rule violations apart from simple conflict tasks, I further aimed at identifying the cognitive processes that make the commission of a violation response more difficult than the response to a conflicting target.

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. Especially for young children, rules of authoritarian figures are mainly obeyed to avoid punishment (Kohlberg, 1963; Piaget, 1932). And even the behavior of many adults is described as "*orientation towards authority, fixed rules,* […], [and] *showing respect for authority*" (description of the conventional, adult level in Kohlberg, 1977, p. 55).

The concept of authority seems to be strongly linked to rules and rule violations. Cognitive conflicts, however, seem to be less associated with authority, especially those that are usually studied in experimental settings. Therefore, in

Experiment 5, I tested whether the commission of a violation modulates the categorization not only of valent stimuli, but also of authority-related targets. To do so, I adapted the design of Experiment 4, but added Probe target words that were rated as being related to the concept of authority. A modulation of the categorization of authority-related words by a preceding rule violation would speak for the idea that violations include authority-related processes.

Methods.

Participants.

A new set of twenty-eight participants was recruited (mean age = 26.3 years, *SD* = 4.3, 11 male, none 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 for the interaction of Prime response type and Probe target valence in Experiment 4. Four participants were removed from the sample due to high error rates and were replaced.

Apparatus, stimuli and procedure.

The experiment was similar to the first experiment, but with an additional 12 authority-related Probe target words that I extracted from the pool of pre-rated target words. Six words were authority-related positive: mentor, mother, father, parents, doctor and professor ($M_{Valence}$ = 7.29, $SD_{Valence}$ = 1.07), and six words were authority-related negative: the German equivalents of violence, weapon, punishment, prison, dictatorship, and admonition ($M_{Valence}$ = 2.01, $SD_{Valence}$ = 0.53). 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 original Probe target words, $|t|s < 1$, $ps > .364$. Further, these new words were all strongly related to authority (positive words: $M_{Authority} = 6.91$, $SD_{Authority} = 0.50$; negative words: $M_{Authority} = 6.64$, $SD_{Authority} = 0.86$), while the original Probe target words were not (positive words: $M_{Authority} = 2.06$, $SD_{Authority} = 0.50$; negative words: $M_{Authority} = 2.18$, $SD_{Authority} = 0.29$). Not only were the valence-ratings matched between the new and the original target words, but also the authority-ratings were similar for strong and weak authority-relation within the original and the new set of target words, $|t|s < 1$, $ps > .527$. 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, $t(10) = 12.08$, $p < .001$, $d = 7.82$. This resulted in four clusters, each containing 6 words, that were either positive or negative and had a strong or weak relation to authority, and thereby, valence and authority-relation could be manipulated orthogonally (see Figure 7). Note, however, that only the valence dimension was relevant for the participants, as in the Probe, the target words still had to be categorized as positive or negative. While the authority-relation of the Probe target words was manipulated in this experiment, it was neither explicitly instructed nor was it relevant for the completion of the task.

Participants again 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.

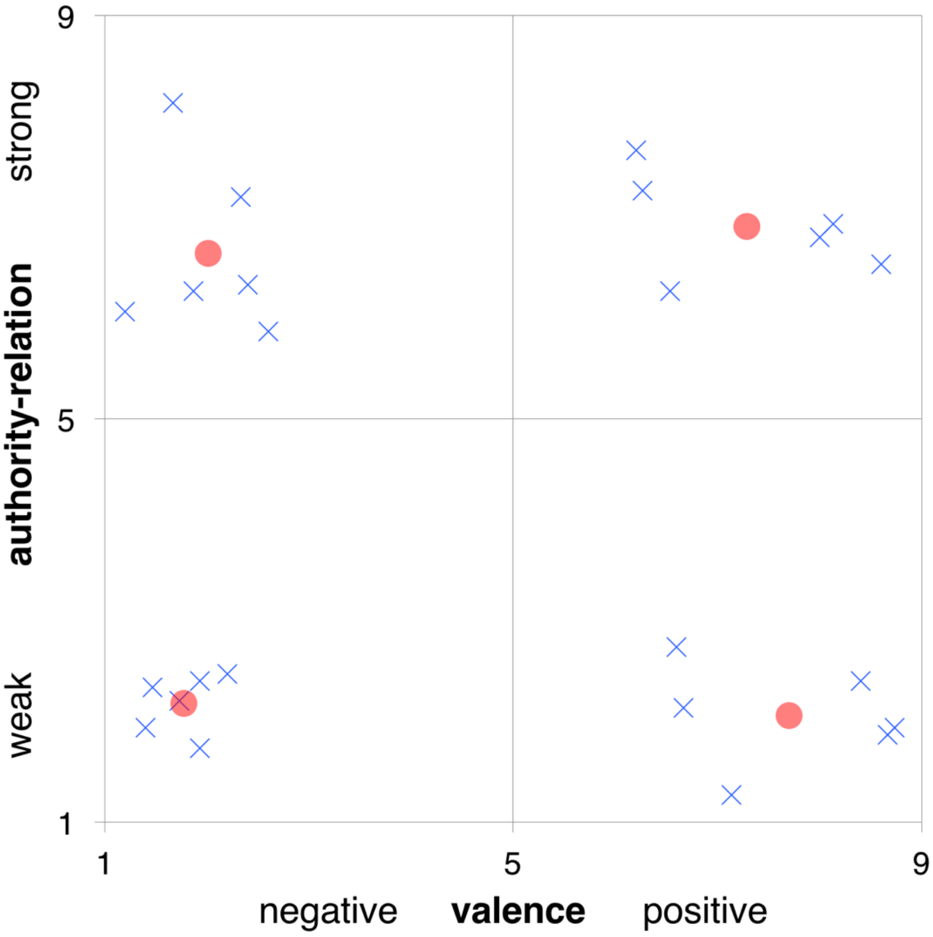

Figure 7. Ratings of the Probe target words.
Probe target words were taken form an item pool of 168 words that were pre-rated concerning word valence and their authority-relation, both on a nine-point scale. Mean ratings for valence are depicted on the abscissa (1 = negative, 9 = positive), mean ratings 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.

Results.

Data selection and analyses.

The data was treated exactly as in Experiment 5. I omitted trials in which participants failed to act according to the instruction (Prime: 8.8%, with more errors for violations than for rule-based responses, *t*(27) = 7.02, *p* < .001, *d* = 0.86; Probe: 7.7%, irrespective of Probe target valence and Probe target authority-relation, |t|s < 1.65, ps > .110) and the immediately following trials (Prime: 6.9%, Probe: 6.2%). Trials were discarded as outliers if any of the measures (RT1, RT2) deviated more than 2.5 standard deviations from the respective cell mean (4.5%). RT1 was then analyzed in an ANOVA with Prime response type (rule-based vs. violation) as within-subjects factor, whereas RT2 was analyzed in a 2×2×2 ANOVA with Prime response type (rule-based vs. violation), Probe target valence (positive vs. negative), and Probe target authority-relation (strong vs. weak) as within-subjects factors (see Figure 8).

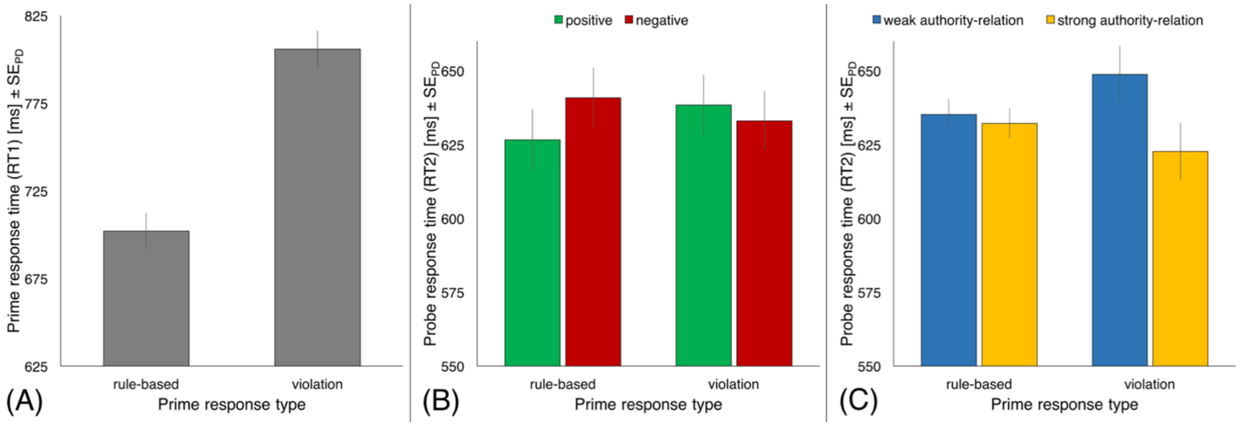

Figure 8. Results of Experiment 5.
Prime response times (RT1; panel A) and Probe response times (RT2, panels B & C) as a function of Prime response type (abscissa), Probe target valence (panel B: left, green bars for positive targets; right, red bars for negative targets), and Probe target authority-relation (panel C: left, blue bars for weakly authority-related targets; right, yellow bars for strongly authority-related targets). Error bars represent standard errors of paired differences (SE_{PD}), for the interactions calculated separately for each instance of Prime response type (Pfister & Janczyk, 2013).

Prime responses.

A significant effect of Prime response type emerged, $F(1,27) = 103.06$, $p < .001$, $\eta_p^2 = .79$, driven by slower responses for violations (806ms) than for rule-based behavior (702ms, Figure 8A).

Probe responses.

A significant effect of Probe target authority-relation emerged, $F(1,27) = 5.97$, $p = .021$, $\eta_p^2 = .18$, indicating faster responses to target words with a strong relation to authority (627ms) compared to target words with a weak relation (642ms). There was an interaction between Prime response type and Probe target valence, $F(1,27) = 7.39$, $p = .011$, $\eta_p^2 = .22$, with response costs for negative targets after rule-based responses ($\Delta = 14$ms), and a response benefit for negative targets after rule violations ($\Delta = -5$ms, Figure 8B). An interaction between Prime response type and Probe target authority-relation, $F(1,27) = 5.66$, $p = .025$, $\eta_p^2 = .17$, indicated weak benefits for authority-related words after rule-based responses ($\Delta = 3$ms), but strong benefits after rule violations ($\Delta = 26$ms, Figure 8C). Further, there was an interaction between the Probe target valence and the Probe target authority-relation, $F(1,27) = 6.31$, $p = .018$, $\eta_p^2 = .19$, 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), but a strong benefit for positive words ($\Delta = 27$ms).

Discussion.

In Experiment 5, I conceptually replicated Experiment 4, but doubled the number of Probe words. Thereby, I added an additional factor to the experiment, namely the authority-relation of the Probe. Importantly, the new set of Probe words allowed for an independent manipulation of valence and authority-relation. This enabled me not only to test for an affective component of rule violations, but also for authority-related processes that rule violations might entail.

First, I replicated the interaction between the Prime response type and Probe target valence: Negative target words were again categorized faster than after rule violations than after rule-based responses, providing further evidence that rule violations seem to be considered a negative event. Further, I found an interaction between the Prime response type and the Probe target authority-relation, with a much stronger benefit for the categorization of authority-related over authority-unrelated words after rule violations compared to rule-based responses. Rule violations seem to sensitize towards authority-related stimuli in the environment, consequently, they are categorized much faster after a violation. As outlined above, this special sensitivity towards authority-related stimuli could reflect latent expectations of negative feedback or punishment (Pfister et al., 2016). Before drawing any further conclusions from these results, Experiment 6 provides an important control condition by using a rule inversion task rather than a rule violation task in the Prime.

3.3 Experiment 6.

In Experiment 2, I tested whether having committed a violation modulates the categorization of valent and authority-related words. I found that violations lead to a faster categorization of both, negative and authority-related Probe words. As outlined in the Introduction, this difference may relate to one of three differences between rule-based responses and violation responses: The labelling of the responses as either rule-based or violation, the presentation frequency, and the additional negation of the instructed mapping rule in case of violations. To exclude the latter two of these three factors, I replicated Experiment 5 with a slight variation: Instead of asking participants to *follow* or *break* the instructed rule, I asked them to either *follow* or *invert* the rule (akin to Experiment 3). Also, inversions required the same negation of the instructed mapping rule as violations, and they were presented as often as violations in Experiment 5. Thereby, I realized a control condition that produced the exact same responses as Experiment 5, but instructed an operation that is more neutral than the violation of a rule but still akin to traditional conflict tasks. By employing the inversion task as the Prime and testing its effects on the Probe task, I could test whether the obtained pattern of results also emerged in this setting.

Methods.

Participants.

A new set of twenty-eight participants was recruited (mean age = 28.0 years, *SD* = 10.1, 7 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. Three participants were removed from the sample due to high error rates and were replaced.

Apparatus, stimuli and procedure.

The experiment was mostly identical to Experiment 2. But instead of instructing participants to break a given rule in one out of four trials, participants were asked to either “follow the rule” or “invert the rule” in the Prime. The Probe was not changed at all, and participants again categorized the target words as positive or negative without mentioning their authority-relation. This way, Experiment 6 required the exact same responses as Experiment 5, although participants were not explicitly instructed to break a rule, but confronted with a rather neutral operation.

Results.

Data treatment and analyses.

The data was treated exactly as in Experiments 4 and 5. I omitted trials in which participants failed to act according to the instruction (Prime: 8.4%, with more errors for inversions than for rule-based responses, $t(27) = 7.06$, $p < .001$, $d = 1.13$; Probe: 7.2%, irrespective of Probe target valence and Probe target authority-relation, $|t|\text{s} < 1.17$, $ps > .253$) and the immediately following trials (Prime: 7.1%, Probe: 6.4%). Trials were discarded as outliers if any of the measures (RT1, RT2) deviated more than 2.5 standard deviations from the respective cell mean (4.3%). RT1 was then analyzed in an ANOVA with Prime response type (rule-based vs. inversion) as within-subjects factor, whereas RT2 was analyzed in a 2×2×2 ANOVA with Prime response type (rule-based vs. inversion), Probe target valence (positive vs. negative), and Probe target authority-relation (strong vs. weak) as within-subjects factors (see Figure 9).

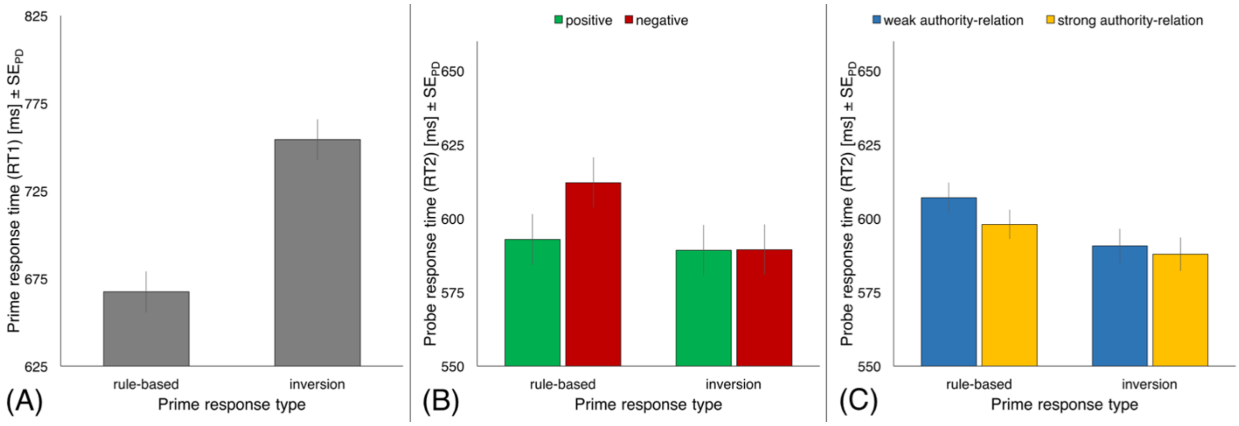

Figure 9. Results of Experiment 6.
Prime response times (RT1; panel A) and Probe response times (RT2, panels B & C) as a function of Prime response type (abscissa), Probe target valence (panel B: left, green bars for positive targets; right, red bars for negative targets), and Probe target authority-relation (panel C: left, blue bars for weakly authority-related targets; right, yellow bars for strongly authority-related targets). Error bars represent standard errors of paired differences (SE_{PD}), for the interactions calculated separately for each instance of Prime response type (Pfister & Janczyk, 2013).

Prime responses.

A significant effect of Prime response type emerged, $F(1,27) = 57.05$, $p < .001$, $\eta_p^2 = .68$, driven by slower responses for inversions (754ms) than for rule-based behavior (667ms, Figure 9A).

Probe responses.

A significant effect of Prime response type emerged, $F(1,27) = 8.59$, $p = .007$, $\eta_p^2 = .24$, indicating faster responses after inversion responses (589ms) compared to rule-based responses (603ms). There was an interaction between Prime response type and Probe target valence, $F(1,27) = 4.61$, $p = .041$, $\eta_p^2 = .15$, with response costs for negative targets after rule-based responses ($\Delta = 19$ms), and no response costs for negative targets after rule inversions (Δ = 0ms, Figure 9B). Further, there was a marginally significant interaction between the Probe target valence and the Probe target authority-relation, $F(1,27) = 2.93$, $p = .099$, $\eta_p^2 = .10$, with no benefit for target words that have a strong authority-relation over target words with a weak relation for negative words (Δ = -3ms), but a benefit for positive words (Δ = 14ms). Notably, the interaction between Prime response type and Probe target authority-relation returned non-significant results, $F(1,27) = 1.16$, $p = .291$, $\eta_p^2 = .04$ (Figure 9C).

Discussion.

In Experiment 6, I changed the Prime task to an inversion task instead of a violation task. This new inversion task required the exact same negation operation and produced the exact same responses as the violation task, but was labeled neutrally compared to the violation instruction. And indeed, the obtained pattern of result seems to differ from the results of Experiment 5: While both experiments return an interaction between the Prime response type and the Probe target valence, only the violation instruction produced an interaction between the Prime response type and the Probe target authority-relation. The instructional variation in the Prime task seems to influence subsequent word categorization.

To compare the effects of the instructional variation on the Prime- and the Probe-task between all experiments, I conducted between-experiments analyses. As both, violations and inversions were presented with the same frequency, and they both required a negation of the instructed mapping rule, differences between the experiments can be unequivocally attributed to the labeling of the response options.

3.4 Between-experiment analyses.

Results.

For the between-experiment analysis of RT1, I conducted a 2×2 split-plot ANOVA with experiment as between-subjects factor and the within-subject factor Prime response type (rule-based vs. deviant, with deviant corresponding to rule violations in Experiments 4-5, and to rule inversions in Experiment 6). For this analysis, Experiments 4 – 6 could be considered, as, apart from the instructional manipulation in Experiment 6, no additional experimental factors were introduced here. Experiments 4 & 5 were therefore pooled to provide a better estimate of the effects of rule violations and then contrasted against the rule inversions of Experiment 6. Also, as both violations and inversions proved to produce higher error rates compared to rule-based responses, the error rates of Experiments 4 – 6 were compared to test whether the comparison of violation and inversion trials is affected by a speed-accuracy tradeoff.

RT2 was analyzed with a 2×2×2×2 split-plot ANOVA with Prime response type (rule-based vs. deviant), Probe target valence (positive vs. negative), Probe target authority-relation (strong vs. weak) as within-subjects factors and experiment as between-subjects factor. Here, only Experiments 5 & 6 could be considered, as Experiment 4 did not include the factor Probe target authority-relation. Again, error rates were compared between experiments to account for possible tradeoffs.

To reduce redundancy, I only focused on interactions that included the factor experiment. As I expected the effects of rule violations to exceed those of rule inversions (Chapter 2), all follow-up tests are reported as one-tailed.

Prime responses.

A marginally significant interaction between Prime response type and experiment, $F(1,78) = 2.37$, $p = .064$, $\eta_p^2 = .06$, was driven by lager effects of violations ($\Delta = 108$ms), compared to inversions ($\Delta = 87$ms). The analysis of the error rates did not yield any significant effects, $F < 1$, $p = .885$, with violations and inversions producing comparable error rates.

Probe responses.

The three-way interaction between Prime response type, Probe target authority-relation, and experiment returned significant results, $F(1,54) = 6.74$, $p = .006$, $\eta_p^2 = .11$, with a significant interaction between Prime response type and Probe target authority-relation in Experiment 2 (Figure 8C), and no interaction in Experiment 3 (Figure 9C). The analysis of the error rates did not yield any significant results, $F\text{s} < 1.16$, $p\text{s} > .286$.

Discussion.

The between-experiment analyses, revealed the specific effects of the instructional manipulation that was introduced in Experiment 3. First, we see that it is indeed harder to commit a violation response compared to an inversion response. From this result (as from those reported in Chapter 2), we could only derive quantitative differences between rule violations and similar control conditions, showing 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. This qualitative difference between the two conditions can be found by analyzing the Probe trials, here we see that violations seem to sensitize towards authority-related stimuli, while inversions do not. The implications of this result will be discussed in the Preliminary Discussion of this chapter.

3.5 Experiment 7.

In Experiments 4-6, I 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 behavior 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 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. For Experiment 7, I 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 4-6 were simply due to semantic priming, the same effects should emerge again. However, if the effects were 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. 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 Experiment 5 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 475ms (mean RT1 in Experiments 5-6) instead of the Prime target, which was then followed by the 100ms blank. This setup ensured that the temporal structure between Experiments 4-7 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 Experiments 4-6), 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.

Even though there was no Prime response in Experiment 7, I still abbreviate Probe response times as RT2 to remain consistent with the terminology of Experiments 4-6. For the following analyses, I only used trials from the experimental blocks. I 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%). RT2 was then analyzed in a 2×2×2 ANOVA with Prime cue (rule-based vs. violation), Probe valence (positive vs. negative) and Probe authority-relation (strong vs. weak) as within-subjects factors.

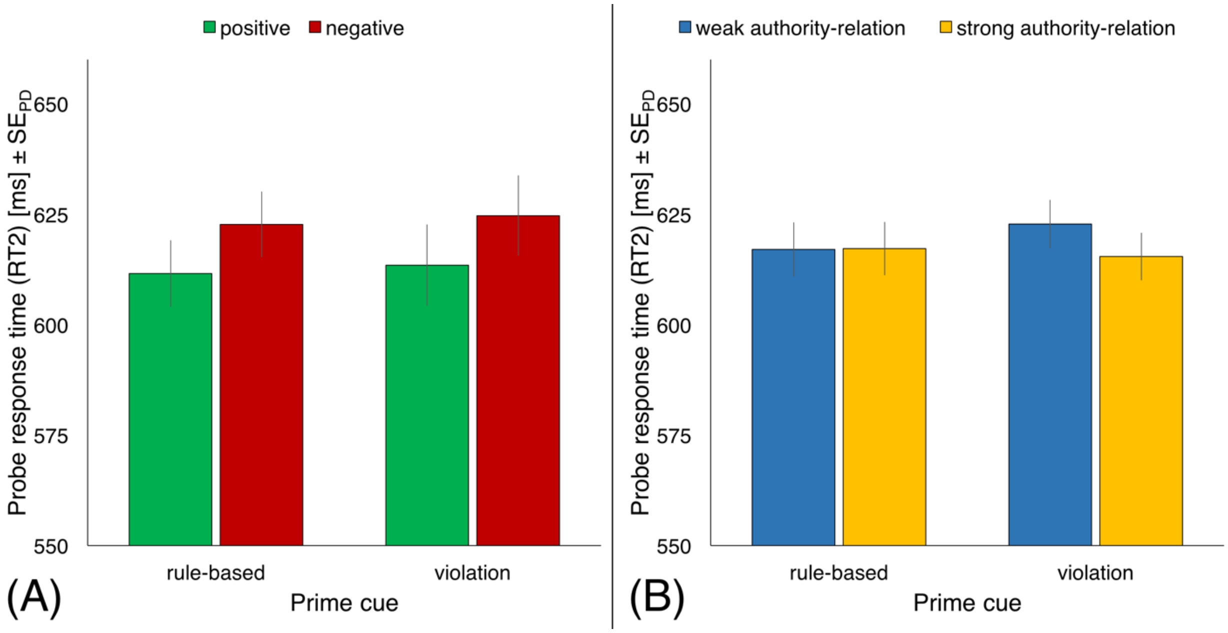

Figure 10. Results of Experiment 7.
Probe response times (RT2) as a function of Prime cue (abscissa), Probe valence (panel A: left, green bars for positive targets; right, red bars for negative targets), and Probe authority-relation (panel B: left, blue bars for weakly authority-related targets; right, yellow bars for strongly authority-related targets). Error bars represent standard errors of paired differences (SE_{PD}), calculated separately for each instance of Prime cue (Pfister & Janczyk, 2013).

Probe responses.

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, $t(27) = 2.30$, $p = .029$, $d = 0.44$). Neither Prime cue, $F < 1$, nor any interaction involving

Prime cue, *F*s < 1.38, *p*s > .250 (Figure 10), was significant. No other effects or higher-order interactions turned significant, *F*s < 2.71, *p*s > .111.

Discussion.

In Experiment 7, I tested whether the effects obtained in Experiments 4-6 can be explained by semantic priming (Meyer & Schvaneveldt, 1971). I adapted the procedure of Experiment 5 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 5. 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 5, and thus semantic priming is unlikely to account for both, the affective and the authority-related aftereffects 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 setu*p* is unable to identify such an effect, as Prime and Probe task are not presented in sufficient temporal proximity. We can, however, conclude that the affective component found in the Probe trials of Experiment 5 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 7, again with positive authority-related target words categorized 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.

3.6 Preliminary Discussion.

In Experiments 4-7, I 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 categorization task as the Probe, I could identify how having committed a violation response modulated the sensitivity towards valent (Experiments 4) and authority-related stimuli (Experiment 5). The aftereffects of rule inversions were tested the same way as a control condition (Experiment 6). Finally, the possibly confounding factor of semantic priming was ruled out (Experiment 7).

The data of the Prime responses shows that it is indeed harder to commit a violation compared to a rule inversion, replicating previous results (Chapter 2). 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 labeling of the response influences the difficulty of these responses: violations produce bigger effect sizes, they are 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, I tested how violations and inversion modulated a subsequent categorization of valent and authority-related words. The data of these Probe responses showed that both violations and inversions sensitize towards negative stimuli: while after rule-based responses, positive target words were categorized faster, a violation and an inversion seem to promote the processing of negative target words, which were consequently categorized faster afterwards. This result stresses the conflicting nature and aversive quality of violations (Aarts et al., 2012; Dreisbach & Fischer, 2012). However, this is not unique to violations, but is also true for inversions. So while rule violations seem to entail an affective component, it can be attributed to the simultaneous activation of two responses, the default, rule-based response and the deviant response, as this is also the case for inversion responses. This double activation makes these responses more difficult and triggers an aversive signal afterwards, which in return promotes the processing of negative stimuli.

The analysis of the authority-related dimension of the Probe target words, however, tells a different story. Here, I 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. Other than sensitizing towards negative stimuli, violations can also act as a prime for further authority-related stimuli.

Heightened attention towards authority-related stimuli that is specific to violations might reflect latent expectations of sanctions and punishment (Pfister et al., 2016): Even though I explicitly omitted this in my 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 behavior (Kohlberg, 1963, 1977; Piaget, 1932) and also is 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 these 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 be suspended by local, situational factors.

To conclude this chapter, here I show that violation responses trigger processes that sensitize 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 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 qualitatively different.

4. How.

Overall, the empirical evidence presented so far suggests that cognitive costs are an inevitable burden of rule violations. However, there are individuals who might be more efficient than others at violating rules. Take, for example, criminals convicted for theft, fraud, swindle, or forgery. When these individuals are asked to break rules in an experimental setting, they show significantly reduced response costs for violations when compared to a control grou*p* with no criminal history (Jusyte et al., in press). They seem to suffer less from the burdens of non-conformity, which ultimately enables them to break rules more easily (“law of less work”, Kool, McGuire, Rosen, & Botvinick, 2010). Equally, lying is considered the socially disregarded alternative to being honest, and for most people, telling lies is associated with cognitive effort (Duran, Dale, & McNamara, 2010; Foerster, Wirth, Kunde, & Pfister, in press; Spence et al., 2001). Still, the majority of lies are told only by a few prolific liars, while most people are honest most of the time (Serota & Levine, 2014). The enhanced cognitive effort that comes with lying, which might be reduced for prolific liars, could drive our tendency to be customarily honest. Further, some so-called countercultures (e.g., punks) even advertise sympathy for deviance and non-conformity, combined with a healthy disrespect for the dominant value system, as their defining feature (Yinger, 1982; Fox, 1987).

What is still unclear is whether individuals who are less subject to the response costs of rule violations are “born this way”, with a cognitive system that is hard-wired to be afflicted less by the struggles of overcoming rule-based behavior, or whether they manage to circumvent these burdens by any means. If the latter were true, the following questions are in order: Is there a way to enable anyone to violate rules efficiently without being thwarted by their distinct behavioral signature? What circumstances allow us to become capable and skillful rule breakers?

Despite curiosity, why would attempts to facilitate rule violation be desirable? Next to the negative examples discussed so far, non-conformity can have an immediate positive spin: Prosocial behavior also falls within the realm of non-conformity, where people do something that is unusual, extraordinary, creative, different from what the others do, where they speak up instead of remaining silent, where they help instead of just standing by (Csíkszentmihályi, 1996; Darley & Latané, 1968; Dovidlo, Piliavin, Schroeder, & Penner, 2006). In these cases, we also have to overcome our default tendency to adhere to the group norms to give way for the prosocial behavior. Innovation, per definition, includes the deviation from common ways of solving problems as well. So again: if we broke rules more efficiently, we might more easily behave in a prosocial and innovative manner. In this chapter, I will approach this subject by testing whether response costs for rule violations can be reduced by controlled, situational variations.

In Experiment 8, we will first have a closer look on how negations respond to these experimental variations, as our working model (see Chapter 2.5) assumes that rule violations are special instances of negations with an add-on. Therefore, in Experiment 8, I will first review the field of negation processing in more depth.

Experiment 9 will then approach rule violations with the same manipulations, and finally Experiment 10 will test whether the transfer of a separate task can additionally influence the burdens of non-conformity.

The data that this work is based on was used for two separate manuscripts that are currently submitted to specialized journals.

4.1 Experiment 8.

"Don't use no double negatives", states a classic guideline on scientific writing (Trigg, 1979). And rightfully so: Already a single negation requires effortful cognitive processing that may not only fail to reach the intended outcome, but may even produce the exact opposite result. Such detrimental effects of negations are also known as the white bear effect, which lends its name to findings that participants who actively tried not to think of a white bear actually found themselves to be haunted by precisely this mental image (Wegner, Schneider, Carter, & White, 1987).

Similar ironic effects do not only occur during thought suppression, but they also apply to overt behavior. Imagine a college student starting their computer to do coursework. If it was not for the coursework, the student might be inclined to read through recent posts on Facebook, but this behavior would jeopardize any coursework-related plans. At first sight, it seems as if an explicit implementation intention ("*if the computer has started, I will not visit Facebook"*) might help the student (Gollwitzer & Sheeran, 2006). However, recent research suggests that holding this intention may, in fact, increase the student's likelihood of falling back to their habit (Adriaanse, Van

Oosten, De Ridder, De Wit, & Evers, 2011). That is: Intending not to do something may at times promote a mental representation of precisely the unwanted behavior (Gollwitzer & Oettingen, 2012). The student seeking to overcome an unwanted habit would therefore be well-advised to employ different strategies, such as replacing the habit with a more desirable behavior ("*if the computer has started, I will immediately start my word processor"*) or simply intending to ignore habit-related cues ("*if the computer has started, I will ignore the Facebook icon"*). These strategies might be especially promising, because processes such as ignoring certain stimuli can be trained to further improve performance (Cunningham & Egeth, 2016).

There might be situations, however, in which negation is the only sensible option. For instance, certain stereotypes may not be easily ignored or countered by wanted behavior. These situations pose a considerable challenge, because even extended training of stereotype negation has been shown to enhance rather than reduce stereotyping (Gawronski, Deutsch, Mbirkou, Seibt, & Strack, 2008). These findings suggest that, in contrast to deliberate ignoring (Cunningham & Egeth, 2016), the cognitive requirements of negation processing cannot be mitigated easily by high-frequency training.

Theoretical models of how negations are represented seem to agree with this notion (Gilbert, 1991; Wegner, 2009). Intuitively, one might follow the *Cartesian approach*, which assumes that mental representations are managed by two separate and serial processes: *Comprehension* recollects the pure semantic content of a representation, followed by an *assessment* of the semantic content as true or false.

However, empirical evidence suggests that the human mind is better described by the *Spinozan approach* (Gilbert, 1991; Gilbert, Krull, Malone, 1990; Wegner et al., 1985). In this view, the comprehension of an idea inherently entails that its semantic content is accepted as true. Comprehension and assessment are conceptualized in a single, joint process. False statements therefore require an additional process that rejects the idea and relabels the automatically accepted content as false. The unacceptance of an idea requires time and effort, and as the semantic content of a negation is always automatically accepted in a first step, this unacceptance process is required for every single instance of a negation, irrespective of its frequency. Therefore, high-frequency training of negations alone can hardly reduce their ironic effects (Gawronski et al., 2008).

So far, it seems as if the processing of negations is difficult and, ironically, attempts to mitigate negation effects may even result in the complete opposite. However, I propose that a potent strategy to counter negation effects emerges when negation processing is viewed from the perspective of cognitive conflict (Schroder et al., 2012). This perspective is motivated by structural similarities of how negations and cognitive conflict are processed: The resolution of a negation involves two competing representations, one of which is activated automatically while the other requires effortful processing, and this process mirrors the resolution of conflict induced by tasks that require the inhibition of a prepotent response (such as incongruent Stroop stimuli or NoGo stimuli). Recent evidence indeed suggests negations to rely on precisely this type of response inhibition (de Vega et al., 2016).

The literature on cognitive conflict and control also offers two clear methods to reduce cognitive conflict: Conflict effects are minimized if conflict is experienced both, frequently and recently (Botvinick et al., 2001). In frequency manipulations, the proportion of conflicting trials is raised, and as a consequence, response costs for the conflicting trials decrease (Logan & Zbrodoff, 1979; Funes, Lupiáñez, & Humphreys, 2010). In recency manipulations, conflicting and non-conflicting trials are analyzed as a function of the immediately preceding trial, resulting in reduced response costs in conflicting trials after having just experienced a conflicting trial (Gratton et al., 1992). Both these factors seem to reduce conflict effects independently of each other (Torres-Quesada, Funes, & Lupiáñez, 2013). Furthermore, effects of conflict frequency may at least partly be explained as being due to a higher probability of benefits due to conflict recency (Botvinick et al., 2001; Torres-Quesada, Lupiáñez, Milliken, & Funes, 2014).

These findings suggest that previous attempts to mitigate negation effects fell just short of providing a powerful solution: The costs of negation processing might be reduced if negations are not only applied frequently (as in previous training studies; Gawronski et al., 2008) but rather if a particular negation has been processed both, frequently and very recently. Such an impact of recency is also conceivable within the *Spinozan model*: When a mental representation has been recollected and negated, repeating that same process for a subsequent negation shortly after might not be necessary, because traces of the negated representation might still be in working memory and can be used rather than recollecting and rejecting the mental representation's semantic content anew.

Therefore, I set out to investigate the impact of frequency and recency on negation effects in a combined design. Participants were either confronted with a high or low frequency of negations and I analyzed the impact of negation frequency and recency on the costs incurred by negation processing. To measure these costs, I used a motion tracking design to analyze the spatial deviation of movement trajectories of negation responses relative to standard, affirmative responses. Such movement trajectories have been shown to be particularly sensitive to negation processing (Dale & Duran, 2011; Wirth, Pfister, Foerster, Huestegge, & Kunde, 2016) and conflict processing alike (Calderon, Verguts, & Gevers, 2015; Dshemuchadse, Scherbaum, & Goschke, 2013). I expected these measures to yield strong negation costs but, crucially, negation costs should be reduced or even absent if participants could benefit from both, frequent and recent negation processing. However, if only a low frequency of negations has been experienced, no adaptation effects should emerge (see Experiment 3).

Methods.

Participants.

Eighty participants were recruited (mean age = 25.5 years, *SD* = 4.7, 28 male, 8 left-handed) and received either course credit or €10 monetary compensation. Because the literature did not allow for estimating a possible effect size *a priori*, I chose to recruit sufficient participants to detect a medium-sized effect of *d* = 0.50 with high power (1-β = .99), while at the same time providing ample chances to detect even smaller effects (1-β ≥ .8 for *d*s > 0.32).

All participants gave informed consent, were naïve to the purpose of the experiment and were debriefed after the session. The data of one participant was removed due to technical difficulties during testing, and five participants were removed from the sample due to high error rates (> 25%).

Stimuli and procedure.

The Experiment was derived from the setup of Experiment 1 with the following changes: I used two shapes (square and triangle) as target stimuli to prompt movements to the left or to the right target (Figure 11). In between trials, the two shapes were displayed to the left and right of the screen center to remind participants of the stimulus-response mapping. In between the two shapes, an exclamation mark (!) instructed standard responses based on the displayed mapping rule, a circular arrow (↺) prompted participants to negate the displayed mapping. Error feedback was displayed if participants reached the wrong target area or failed to hit one of the designated target areas at all.

Between blocks, the proportion of negation trials was manipulated: in blocks with a low proportion of negations (low-PN), the displayed mapping rule had to be negated in one out of four trials. In blocks with a high proportion of negations (high-PN), the mapping rule had to be negated in three out of four trials. The proportion of negations within a block changed after half of the experiment, the order of presentation (first half: low-PN, second half: high-PN vs. first half: high-PN, second half: low-PN) was manipulated between participants. Participants completed 20 blocks of 64 trials each.

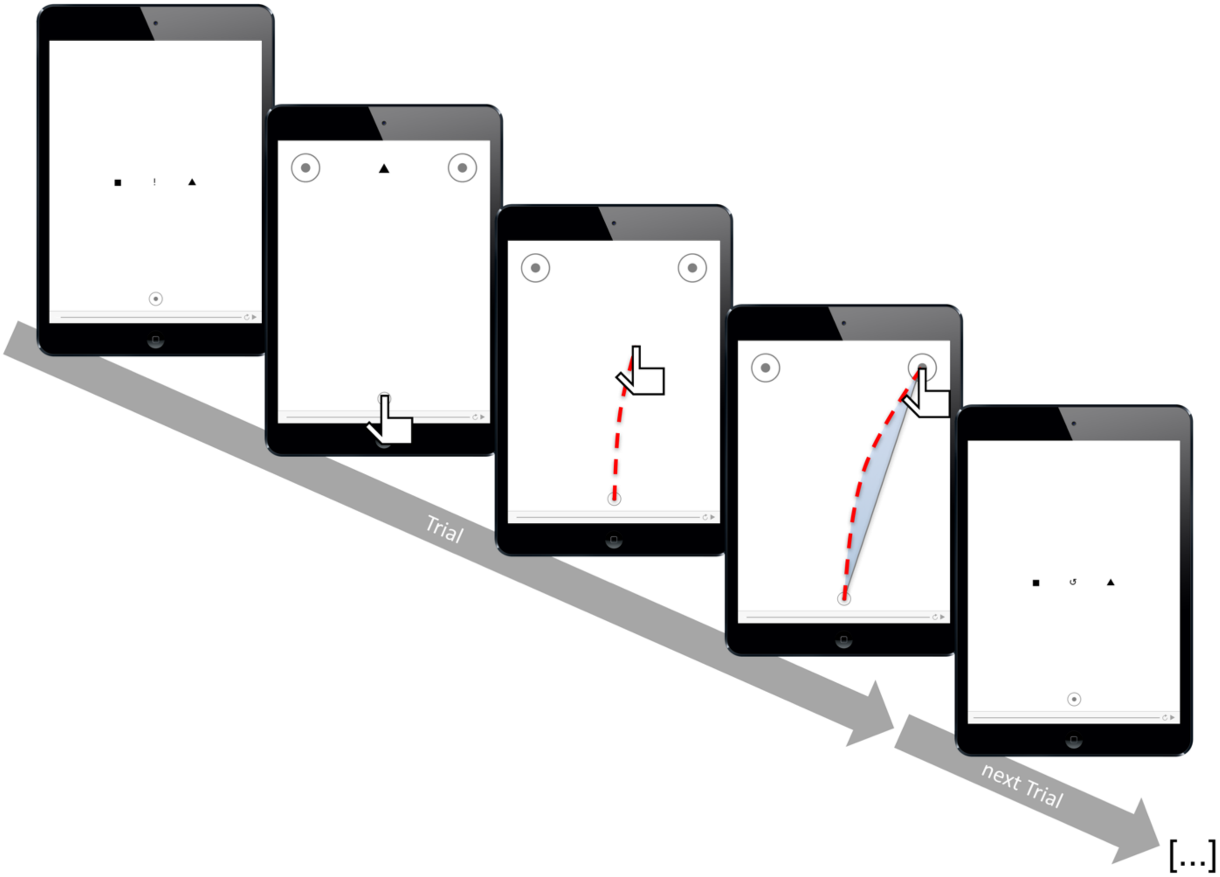

Figure 11. Procedure of Experiment 8.
Before each trial, participants were reminded of the mapping rule, together with the instruction to either perform a standard response according to the displayed mapping rule or to negate this mapping rule in the next trial. As soon as participants put their finger on the starting area, the mapping rule disappeared and the two target areas and the target symbol appeared, prompting movements to the left or the right. The target symbol disappeared when the finger left the starting area. A trial was completed when the finger was lifted from the screen inside one of the two target areas, and the next trial started immediately with the corresponding standard or negation instructions.

Results.

Data selection and analyses.

For all analyses, the first block of each PN condition was considered practice and removed. I then omitted trials in which participants failed to act according to the instruction or failed to hit any of the two target areas at all (6.0%) and trials following errors (5.1%). Trials were discarded as outliers if any of the measures (IT, MT, AUC) deviated more than 2.5 standard deviations from the respective cell mean (5.3%). Each measure was then analyzed in a separate 2×2×2×2 analysis of variance (ANOVA) with current response type (standard vs. negation), preceding response type, and proportion negation (low-PN vs. high-PN) as within-subject factors, and proportion order (low-PN-first vs. high-PN-first) as a between-subjects factor.

Initiation times.

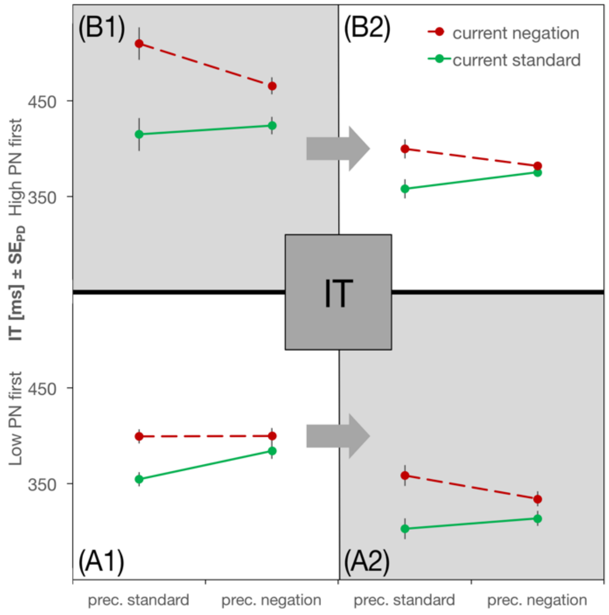

Figure 12. Results of Experiment 8 (ITs).
Initiation times (IT) are plotted as a function of preceding response type (abscissa), current response type (continuous, green line for standard responses; dashed, red line for negation responses), and the current proportion of negations (PN; white background for low-PN, gray background for high-PN). Further, the figure is split by proportion order: The lower panels (A) represent the low-PN-first condition, the upper panels (B) represent the high-PN-first condition. Panels with the number one represent the first half of the experiment per proportion order, panels with the number two represent the second half. Error bars represent standard errors of paired differences (SE_{PD}), calculated separately for each instance of preceding response type (Pfister & Janczyk, 2013).

A significant effect of current response type, $F(1,72) = 68.54$, $p < .001$, $\eta_p^2 = .49$, was driven by faster response initiation for standard responses (367ms) than for negations (407ms). A marginally significant effect of proportion negation, $F(1,72) = 2.79$, $p = .099$, $\eta_p^2 = .04$, described response initiations in the low-PN condition as faster (382ms) compared to those in the high-PN condition (392ms). Proportion order interacted with proportion negation, $F(1,72) = 85.83$, $p < .001$, $\eta_p^2 = .54$, with costs for low-PN blocks relative to high-PN blocks for participants who started with the low-PN condition (Δ = -56ms), but benefits for those who started with the high-PN condition (Δ = 81ms). An interaction between preceding response type and proportion negation, $F(1,72) = 9.02$, $p = .004$, $\eta_p^2 = .11$, indicated post-negation slowing in the low-PN condition (Δ = 7ms), but post-negation speeding in the high-PN condition (Δ = -12ms). Similarly, there was an interaction between current response type and proportion negation, $F(1,72) = 22.34$, $p < .001$, $\eta_p^2 = .24$, with a smaller negation effect in low-PN blocks (Δ = 27ms) compared to high-PN blocks (Δ = 54ms). Also, there was an interaction between current response type and preceding response type, $F(1,72) = 24.27$, $p < .001$, $\eta_p^2 = .25$, with a stronger negation effect after standard responses (Δ = 59ms) than after negation responses (Δ = 21ms, Figure 12). There was a three-way interaction between the factors current response type, preceding response type, and proportion negation, $F(1,72) = 4.62$, $p = .035$, $\eta_p^2 = .06$, as well as an interaction between current response type, proportion negation, and proportion order, $F(1,72) = 13.03$, $p = .001$, $\eta_p^2 = .15$. None of the remaining effects were significant, *F*s < 2.50, *p*s > .118.

To break down this complex pattern of results, i.e., to follow up on the significant higher order interactions, I split the analysis and report the data separately for each proportion order. For this follow-u*p* test I thus conducted two 2×2×2 ANOVAs with current response type (standard vs. negation), preceding response type, and proportion negation (low-PN vs. high-PN) as within-subject factors.

Initiation times, low-PN-first.

A significant effect of current response type, $F(1,37) = 33.15$, $p < .001$, $\eta_p^2 = .47$, was driven by slower response initiations for negations (374ms) than for standard responses (341ms). Response initiation after negation responses was overall slower (360ms) than after standard responses (355ms), $F(1,37) = 4.74$, $p = .036$, $\eta_p^2 = .11$. Further, response initiation was slower in the low-PN condition (386ms) relative to the high-PN condition (329ms), $F(1,37) = 31.39$, $p < .001$, $\eta_p^2 = .46$. Response benefits after standard responses emerged for the low-PN condition (Δ = 15ms), and response costs emerged for the high-PN condition (Δ = -6ms), as qualified by the interaction of response type and proportion negation, $F(1,37) = 16.96$, $p < .001$, $\eta_p^2 = .31$. Finally, the interaction between preceding response type and current response type was significant, $F(1,37) = 11.12$, $p = .002$, $\eta_p^2 = .23$, with a stronger effect of negations after standard responses (Δ = 50ms) compared to after negation responses (Δ = 17ms, Figure 12A). Current negations did not benefit from previous negations relative to previous standard responses in low-PN blocks (Δ = 0ms, $|t| < 1$), but they did benefit in the later high-PN blocks (Δ = 24ms, $t(37) = 3.66$, $p = .001$, $d = 0.61$). None of the remaining effects were significant, Fs < 1.14, ps > .292.

Initiation times, high-PN-first.

A significant effect of current response type, $F(1,35) = 35.63$, $p < .001$, $\eta_p^2 = .50$, was driven by slower response initiations for negations (441ms) than for standard responses (394ms). A significant effect of preceding response type, $F(1,35) = 5.06$, $p = .031$, $\eta_p^2 = .13$, described responses following standard responses as slower (423ms) compared to responses following negations (413ms). Further, response initiation was slower in the high-PN condition (458ms) relative to the low-PN condition (377ms), $F(1,35) = 54.96$, $p < .001$, $\eta_p^2 = .61$. Response costs after standard responses emerged for the high-PN condition (Δ = -19ms), but not for the low-PN condition (Δ = -1ms), $F(1,35) = 5.10$, $p = .030$, $\eta_p^2 = .13$. The interaction between current response type and proportion negation was significant, $F(1,35) = 23.10$, $p < .001$, $\eta_p^2 = .40$, with a stronger effect of negations in high-PN blocks (Δ = 71ms) compared to low-PN blocks (Δ = 24ms). The interaction between preceding response type and current response type was significant, $F(1,35) = 13.03$, $p = .001$, $\eta_p^2 = .27$, with a stronger effect of negations after standard responses (Δ = 70ms) compared to after negation responses (Δ = 25ms). Finally, the three-way interaction was significant, $F(1,35) = 4.86$, $p = .034$, $\eta_p^2 = .12$, indicating a significant interaction between preceding and current response type for high-PN blocks, $F(1,35) = 8.82$, $p = .005$, $\eta_p^2 = .20$ (Figure 12B1), and a significantly stronger interaction for low-PN blocks, $F(1,37) = 13.92$, $p = .001$, $\eta_p^2 = .29$ (Figure 12B2). Current negations benefitted from previous negations relative to previous standard responses in both, high-PN and low-PN blocks (Δs > 18ms, ts > 1.99, ps < .054, ds > 0.33). None of the remaining effects were significant, Fs < 2.44, ps > .127.

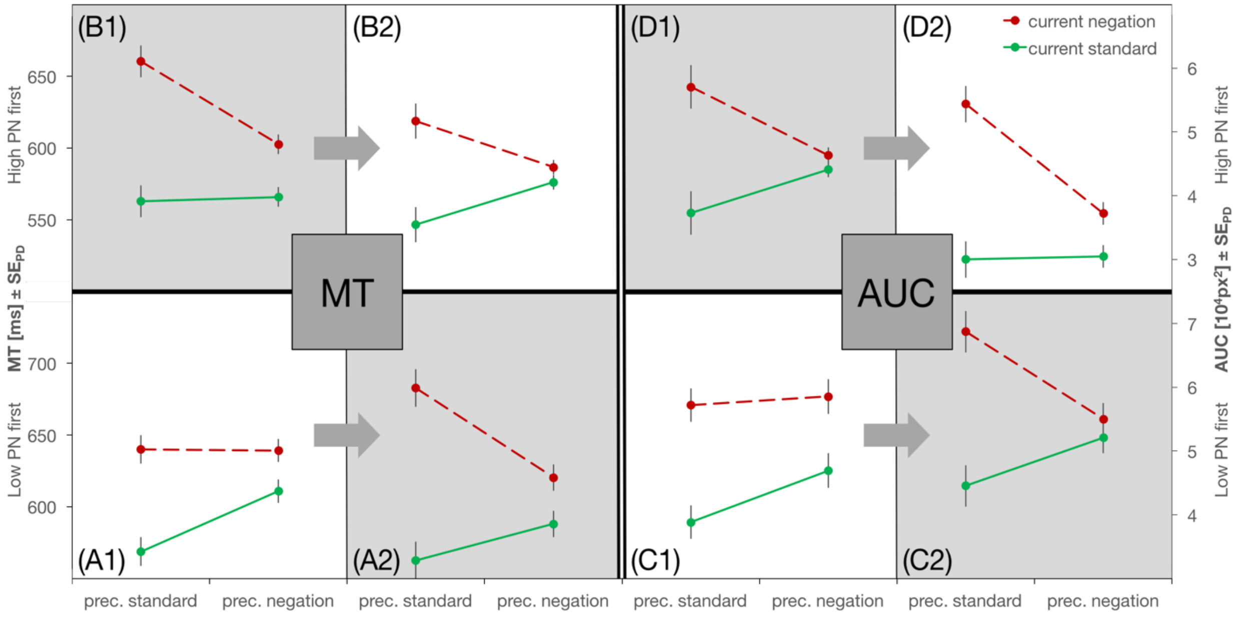

Figure 13. Results of Experiment 8 (MTs & AUCs).
Movement times (MT; left) and areas under the curve (AUC; right) are plotted as a function of preceding response type (abscissa), current response type (continuous, green line for standard responses; dashed, red line for negation responses), and the current proportion of negations (PN; white background for low-PN, gray background for high-PN). Further, the figure is split by proportion order: The lower panels (A & C) represent the low-PN-first condition, the upper panels (B & D) represent the high-PN-first condition. Panels with the number one represent the first half of the experiment per proportion order, panels with the number two represent the second half. Error bars represent standard errors of paired differences (SE_{PD}), calculated separately for each instance of preceding response type (Pfister & Janczyk, 2013).

Movement times.

A significant effect of current response type, $F(1,72) = 171.68$, $p < .001$, $\eta_p^2 = .71$, indicated standard responses (572ms) to be faster than negations (631ms). Responses were slightly faster after negation responses (598ms) relative to after standard responses (605ms), $F(1,72) = 6.69$, $p = .012$, $\eta_p^2 = .09$. An interaction between preceding response type and proportion order, $F(1,72) = 9.55$, $p = .003$, $\eta_p^2 = .12$, indicated post-negation speeding in the high-PN-first group (Δ = -15ms), and no difference in the low-PN-first group (Δ = 1ms). Current response type interacted with proportion negation, $F(1,72) = 24.11$, $p < .001$, $\eta_p^2 = .25$, with a smaller negation effect in low-PN blocks (Δ = 45ms) compared to high-PN blocks (Δ = 72ms). A similar interaction emerged between preceding response type and proportion negation, $F(1,72) = 64.69$, $p < .001$, $\eta_p^2 = .47$, with slower negation responses relative to standard responses in low-PN blocks (Δ = 9ms), and the reversed effect in high-PN blocks (Δ = -23ms). Also, there was an interaction between current response type and preceding response type, $F(1,72) = 68.33$, $p < .001$, $\eta_p^2 = .49$, with a stronger negation effect after standard responses (Δ = 90ms) than after negation responses (Δ = 27ms, Figure 13). There was a three-way interaction between the factors current response type, preceding response type, and proportion negation, $F(1,72) = 4.59$, $p = .035$, $\eta_p^2 = .06$, as well as a four-way interaction between all factors, $F(1,72) = 4.24$, $p = .043$, $\eta_p^2 = .06$. Accordingly, I again conducted follow-up analyses for each proportion order. None of the remaining effects were significant, *F*s < 2.50, *p*s > .118.

Movement times, low-PN-first.

A significant effect of current response type, $F(1,37) = 118.58$, $p < .001$, $\eta_p^2 = .76$, was driven by slower responses for negations (641ms) than for standard responses (577ms). Response benefits after standard responses emerged for the low-PN condition (Δ = 21ms), and response costs emerged for the high-PN condition (Δ = -18ms), $F(1,37) = 44.40$, $p < .001$, $\eta_p^2 = .55$. The interaction between current response type and proportion negation was also significant, $F(1,37) = 9.84$, $p = .003$, $\eta_p^2 = .21$, with a stronger effect of negations in high-PN blocks (Δ = 77ms) compared to low-PN blocks (Δ = 50ms). The interaction between preceding response type and current response type was significant, $F(1,37) = 36.19$, $p < .001$, $\eta_p^2 = .49$, with a stronger effect of negations after standard responses (Δ = 96ms) compared to after negation responses (Δ = 32ms). Finally, the three-way interaction between preceding response type, current response type and proportion negation was significant, $F(1,37) = 7.02$, $p = .012$, $\eta_p^2 = .16$, with a significant interaction between preceding and current response type for high-PN blocks, $F(1,37) = 29.60$, $p < .001$, $\eta_p^2 = .44$ (Figure 13A2), and a significantly smaller interaction for low-PN blocks, $F(1,37) = 18.16$, $p < .001$, $\eta_p^2 = .32$ (Figure 13A1). Especially, current negations did not benefit from pervious negations relative to previous standard responses in low-PN blocks (Δ = 1ms, $|t| < 1$), but did benefit in the later high-PN blocks (Δ = 60ms, $t(37) = 5.55$, $p < .001$, $d = 0.90$). None of the remaining effects were significant, Fs < 1, ps > .696.

Movement times, high-PN-first.

A significant effect of current response type, $F(1,35) = 61.69$, $p < .001$, $\eta_p^2 = .64$, was driven by slower responses for negations (621ms) than for standard responses (568ms). A significant effect of preceding response type, $F(1,35) = 13.46$, $p = .001$, $\eta_p^2 = .28$, described responses following standard responses as slower (602ms) compared to responses following negations (587ms). The interaction between current response type and proportion negation was significant, $F(1,35) = 18.24$, $p < .001$, $\eta_p^2 = .34$, with a stronger effect of negations in high-PN blocks (Δ = 66ms) compared to low-PN blocks (Δ = 41ms). Responses costs after standard responses emerged for the high-PN condition (Δ = -28ms), but not for the low-PN condition (Δ = -2ms), $F(1,35) = 22.01$, $p < .001$, $\eta_p^2 = .39$. The interaction between preceding response type and current response type was significant, $F(1,35) = 32.26$, $p < .001$, $\eta_p^2 = .48$, with a stronger effect of negations after standard responses (Δ = 84ms) compared to after negation responses (Δ = 22ms). Finally, the three-way interaction was not significant, $F < 1$, with similar interactions for both, low-PN and high-PN conditions (Figure 13B). Current negations benefitted from previous negations relative to previous standard responses in both, high-PN and low-PN blocks (Δs > 33ms, ts > 3.71, ps < .001, ds > 0.62). None of the remaining effects were significant, Fs < 2.44, ps > .127.

Areas under the curve.

A significant effect of current response type, $F(1,72) = 135.48$, $p < .001$, $\eta_p^2 = .65$, was driven by more curved trajectories for negations (54022px^2) than for standard responses (40274px^2). There was a significant effect of preceding response type, $F(1,72) = 11.79$, $p = .001$, $\eta_p^2 = .14$, with more curved responses after a standard response (48301px^2) compared to after a negation response (45995px^2). In the high-PN condition, responses were descriptively less curved (46106px^2) than in the low-PN condition (48190px^2), $F(1,72) = 3.88$, $p = .053$, $\eta_p^2 = .05$. An interaction between preceding response type and proportion order, $F(1,72) = 21.30$, $p < .001$, $\eta_p^2 = .23$, indicated slight post-negation effects for participants in the low-PN-first group (Δ = 823px^2), but a post-negation benefit in the high-PN-first group (Δ = -5610px^2). Proportion order interacted with proportion negation, $F(1,72) = 36.35$, $p < .001$, $\eta_p^2 = .34$, with benefits in low-PN blocks for participants who started with the low-PN condition (Δ = 4683px^2) but costs for those who started with the high-PN condition (Δ = -9226px^2). Response benefits after negation responses emerged in the high-PN condition (Δ = -5822px^2) relative to response costs in the low-PN condition (Δ =

1208px^2), F(1,72) = 42.65, p < .001, η_p^2 = .37. Also, there was an interaction between current response type and preceding response type, F(1,72) = 53.25, p < .001, η_p^2 = .43, with a stronger negation effect after standard responses (Δ = 21696px^2) than after negation responses (Δ = 5799px^2, Figure 13). Finally, there were three-way interactions between the factors current response type, preceding response type, and proportion negation, F(1,72) = 6.82, p = .011, η_p^2 = .09, as well as between current response type, proportion negation, and proportion order, F(1,72) = 4.10, p = .047, η_p^2 = .05, and a four-way interaction between all factors, F(1,72) = 7.41, p = .008, η_p^2 = .09. Accordingly, I again conducted follow-up analyses for each proportion order. None of the remaining effects were significant, Fs < 1.94, ps > .168.

Areas under the curve, low-PN-first.

A significant effect of current response type, F(1,37) = 84.22, p < .001, η_p^2 = .70, was driven by more contorted responses for negations (60430px^2) than for standard responses (46138px^2). Similarly, a significant main effect of proportion negation, F(1,35) = 7.96, p = .008, η_p^2 = .18, marked responses in the low-PN-condition as less contorted (50942px^2) than in the high-PN-condition (55626px^2). Response benefits after standard responses emerged for the low-PN condition (Δ = 4633px^2), but response costs emerged for the high-PN condition (Δ = -2987px^2), F(1,37) = 27.72, p < .001, η_p^2 = .43. The interaction between preceding response type and current response type was significant, F(1,37) = 19.17, p < .001, η_p^2 = .34, with a stronger effect of negations after standard responses (Δ = 20921px^2) compared to after negation responses (Δ = 7665px^2). Finally, the three-way interaction between preceding response type, current response type and proportion negation was significant, F(1,37) =

9.79, p = .003, η_p^2 = .21, with a significant interaction between preceding and current response type for high-PN blocks, F(1,37) = 21.86, p < .001, η_p^2 = .37 (Figure 13C2), and a significantly smaller interaction for low-PN blocks, F(1,37) = 4.32, p = .045, η_p^2 = .11 Figure 13C1). Especially, current negations were descriptively hindered by pervious negations relative to previous standard responses in low-PN blocks (Δ = -1440px^2, t(37) = -0.64, p = .525, d = 0.10), but benefit in the later high-PN blocks (Δ = 13049px^2, t(37) = 5.29, p < .001, d = 0.86). None of the remaining effects were significant, Fs < 1.05, ps > .313.

Areas under the curve, high-PN first.

A significant effect of current response type, F(1,35) = 54.65, p < .001, η_p^2 = .61, was driven by more contorted responses for negations (47257px^2) than for standard responses (34085px^2). A significant effect of preceding response type, F(1,35) = 23.69, p < .001, η_p^2 = .41, described responses following standard responses as more contorted (43476px^2) compared to responses following negations (37866px^2). Similarly, a significant main effect of proportion negation, F(1,35) = 33.40, p < .001, η_p^2 = .49, marked responses in the low-PN-condition as more contorted (45284px^2) compared to the high-PN-condition (36058px^2). Responses costs after standard responses emerged for the low-PN condition (Δ = -2407px^2), but were relatively small compared to the response costs that emerged for the high-PN condition (Δ = -8814px^2), F(1,35) = 16.19, p < .001, η_p^2 = .32. Further, the interaction between current response type and proportion negation was significant, F(1,35) = 5.79, p = .022, η_p^2 = .14, with a stronger effect of negations in high-PN blocks (Δ = 15370px^2) compared to low-PN blocks (Δ = 10975px^2). The interaction between preceding response type and current response type was significant, $F(1,35) = 34.87$, $p < .001$, $\eta_p^2 = .50$, with a stronger effect of negations after standard responses ($\Delta = 22515px^2$) compared to after negation responses ($\Delta = 3830px^2$). Finally, the three-way interaction was not significant, $F < 1$, with similar interactions for both, low-PN and high-PN conditions (Figure 13D). Current negations benefitted from previous negations relative to previous standard responses in both, high-PN and low-PN blocks ($\Delta s > 11819px^2$, ts > 4.78, ps < .001, ds > 0.80).

Discussion.

In Experiment 8, I tested whether the cognitive costs of negation processing can be reduced by a combined manipulation of negation frequency and recency. Previous attempts to reduce the impact of negation processing show that high-frequency training alone actually increases rather than reduces the impact of negations (Gawronski et al., 2008). On the other hand, recency alone did not reduce the impact of negations either (see Experiment 3). My findings are compatible with both views, importantly, they showed a remarkable effect when frequency and recency manipulations are combined. That is: Negations can indeed be countered effectively, if negation operations are performed frequently and if a particular negation has also been applied very recently.

Despite the effectiveness of combined frequency and recency of negations, the cognitive costs of negation processing did not vanish entirely. In line with the *Spinozan model*, these findings suggest that every single activation of a negation requires that its semantic content is initially affirmed, whereas it is negated only in a second step. This second negation step cannot be bypassed by high-frequency training, negations always produce ironic effects (Wegner, 2009). This result also suggests a qualitative difference between the process of negating a response rule on the one hand and consistent ignoring of certain stimuli on the other hand (Cunningham & Egeth, 2016).

Further, I found that negation effects were strongly reduced after previous negations relative to previous affirmations. While the impact of recency has largely been neglected in the negation literature (c.f., Chapter 2), here I describe recency as a crucial factor to mitigate the ironic effects of negations. Again, this notion is compatible with the *Spinozan approach*, if we assume that negations leave a trace in working memory, and a subsequent negation can profit from the already negated semantic content.

Finally, I found that the ironic effects of negations were smallest when both, frequency and recency, play in concert. While a recent negation alone significantly reduced negation effects (as in the first blocks of the low-PN-first group), the greatest benefit came when both, a high frequency and recency, have been experienced, and in combination, they managed to almost eliminate the burdens of negations. Whereas in conflict tasks, frequency and recency have been shown to work independently (Funes et al., 2010), for negations, they might interact: While experiencing (or having experienced) a high frequency seems to signal the necessity for adaptation, recency seems to provide the mechanism for adaptation, possibly via working memory traces.

The conceptualization of negations in the present paradigm closely mirrors typical designs that are employed in research on cognitive conflict. This aspect of the experimental design allowed us to access the cognitive architecture underlying negation processing and their mediating processes in a highly controlled setting. How these results related to more externally valid approaches, e.g., the negation of stereotypes (Gawronski et al. 2008), still has to be demonstrated. But for now, the combined influence of frequency and recency seems to be the most successful and promising attempt to mitigate ironic negation effects on overt behavior.

4.2 Experiment 9.

In Experiment 8, I found that only the combination of the factors frequency and recency produced an adaptation for the response execution parameters of the negation task. It seems as if these two mechanisms do not work independently for negations (in contrast to conflict tasks, Torres-Quesada et al., 2013), but that a high frequency of negations has to be experienced first for recency effects to emerge. In a sense, it might be that a high frequency of negations signals the necessity for adaptation, and recency only then provides the mechanism to do so.

In Experiment 9, I will now test whether this is also true for rule violations, which have been argued to be a special instance of a negation (see Chapter 2). I therefore adapted the setup of Experiment 8 to feature the violation task that was used in Experiment 1. Again, the proportion of violations was manipulated between blocks, and instruction was done via the written labels between trials (follow the rule vs. break the rule). If indeed, rule violations pose as a special instance of a negation task (with an add-on), then this setup should basically replicate the results of Experiment 8. Again, if only a low frequency of violations has been experienced, no adaptation effects should emerge (see Experiment 1).

Methods.

Participants.

Twenty-four participants were recruited (mean age = 26.3 years, *SD* = 7.8, 7 male, 4 left-handed) and received either course credit or €8 monetary compensation. All participants gave informed consent, were naïve to the purpose of the experiment and were debriefed after the session.

Stimuli and procedure.

The experiment was modeled Experiment 1 (see Figure 1). Only now, the proportion of violation trials was manipulated between blocks: in blocks with a low proportion of rule violations (low-PV), the displayed mapping rule had to be violated in one out of four trials. In blocks with a high proportion of rule violations (high-PV), the mapping rule had to be violated in three out of four trials. The proportion of violations within a block changed after half of the experiment, the order of presentation (first half: low-PV, second half: high-PV vs. first half: high-PV, second half: low-PV) was manipulated between participants. Participants completed 20 blocks of 64 trials each.

Results.

Data selection and analyses.

For all analyses, the first block of each PV condition was considered practice and removed. I then omitted trials in which participants failed to act according to the instruction or failed to hit any of the two target areas at all (6.2%) and trials following an error (4.8%). Trials were discarded as outliers if any of the measures (IT, MT, AUC) deviated more than 2.5 standard deviations from the respective cell mean (5.9%). Each measure was then analyzed in a separate 2×2×2×2 analysis of variance (ANOVA) with current response type (rule-based vs. violation), preceding response type, and proportion violation (low-PV vs. high-PV) as within-subject factors, and proportion order (low-PV-first vs. high-PV-first) as a between-subjects factor.

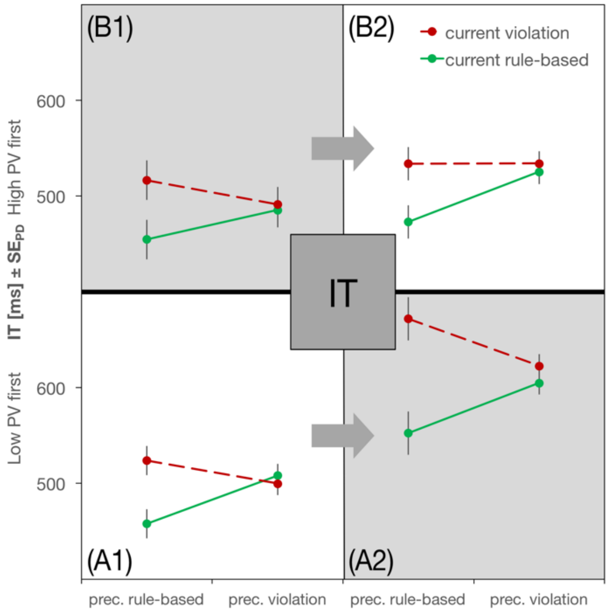

Figure 14. Results of Experiment 9 (ITs).
Initiation times (IT) are plotted as a function of preceding response type (abscissa), current response type (continuous, green line for rule-based responses; dashed, red line for violation responses), and the current proportion of violations (PV; white background for low-PV, gray background

for high-PV). Further, the figure is split by proportion order: The lower panels (A) represent the low-PV-first condition, the upper panels (B) represent the high-PV-first condition. Panels with the number one represent the first half of the experiment per proportion order, panels with the number two represent the second half. Error bars represent standard errors of paired differences (SE_{PD}), calculated separately for each instance of preceding response type (Pfister & Janczyk, 2013).

Initiation times.

A significant effect of current response type, $F(1,46) = 42.45$, $p < .001$, $\eta_p^2 = .66$, was driven by faster response initiation for rule-based behavior (508ms) than for violations (549ms). Also, there was an effect of preceding response type, $F(1,46) = 8.04$, $p = .010$, $\eta_p^2 = .27$, marking response initiations following a rule-based response as faster (523ms) than those following a violation response (532ms). Similarly, an effect of proportion violation, $F(1,46) = 9.08$, $p = .006$, $\eta_p^2 = .29$, described response initiations in the low-PV condition as faster (507ms) compared to those in the high-PV condition (550ms). Proportion order interacted with proportion violation, $F(1,46) = 25.86$, $p < .001$, $\eta_p^2 = .54$, with costs for low-PV blocks for participants who started with the low-PV condition (Δ = -30ms) but benefits for those who started with the high-PV condition (Δ = 115ms). An interaction between preceding response type and proportion violation, $F(1,46) = 5.14$, $p = .034$, $\eta_p^2 = .19$, indicated post-violation slowing in the low- PV condition (Δ = 20ms), but a smaller effect in the high-PV condition (Δ = 2ms). Similarly, there was an interaction between current response type and proportion violation, $F(1,46) = 7.91$, $p = .010$, $\eta_p^2 = .26$, with a smaller violation effect in low-PV blocks (Δ = 32ms) compared to high-PV blocks (Δ = 51ms). Also, there was an interaction between current response type and preceding response type, $F(1,46) = 16.02$, $p = .001$, $\eta_p^2 = .42$, with a stronger violation effect after rule-based responses (Δ = 77ms) than after violation responses (Δ = 6ms). This interaction held true for all combinations of proportion violation and proportion order (Figure 14), as indicated by all higher-order interactions including both factors returning non-significant results, Fs < 1.42, ps > .246. None of the remaining effects were significant, Fs < 1.25, ps > .275.

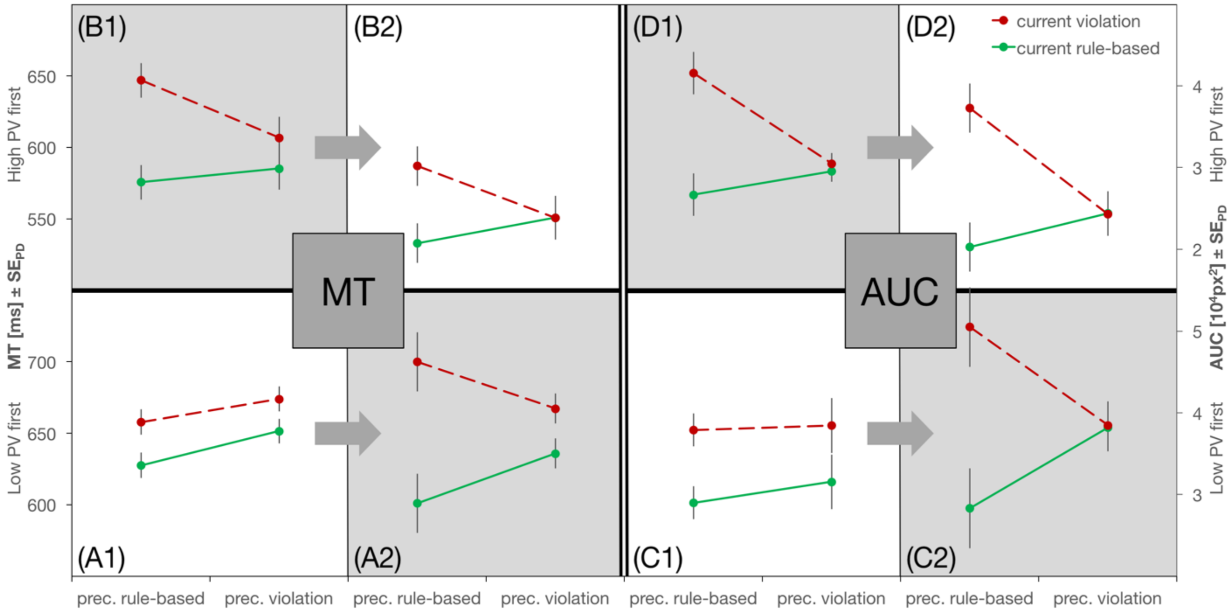

Figure 15. Results of Experiment 9 (MTs & AUCs).
Movement times (MT; left) and areas under the curve (AUC; right) are plotted as a function of preceding response type (abscissa), current response type (continuous, green line for rule-based responses; dashed, red line for violation responses), and the current proportion of violations (PV; white background for low-PV, gray background for high-PV). Further, the figure is split by proportion order: The lower panels (A & C) represent the low-PV-first condition, the upper panels (B & D) represent the high-PV-first condition. Panels with the number one represent the first half of the experiment per proportion order, panels with the number two represent the second half. Error bars represent standard errors of paired differences (SE_{PD}), calculated

separately for each instance of preceding response type (Pfister & Janczyk, 2013).

Movement times.

A significant effect of current response type, $F(1,22) = 50.83$, $p < .001$, $\eta_p^2 = .70$, was driven by faster responses for rule-based behavior (589ms) than for violations (628ms). There was a marginally significant effect of proportion violation, $F(1,22) = 3.87$, $p = .062$, $\eta_p^2 = .15$, with faster responses in the low-PV blocks (597ms) compared to the high-PV blocks (619ms). An interaction between preceding response type and proportion order, $F(1,22) = 6.95$, $p = .015$, $\eta_p^2 = .24$, indicated post-violation slowing for participants in the low-PV-first condition (Δ = 7ms), but post-violation speeding in the high-PV-first group (Δ = -12ms). Current response type interacted with proportion violation, $F(1,22) = 17.15$, $p < .001$, $\eta_p^2 = .44$, with a smaller violation effect in low-PV blocks (Δ = 23ms) compared to high-PV blocks (Δ = 55ms). Also, there was an interaction between current response type and preceding response type, $F(1,22) = 15.63$, $p = .001$, $\eta_p^2 = .42$, with a stronger violation effect after rule-based responses (Δ = 61ms) than after violation responses (Δ = 17ms, Figure 15). There was a three-way interaction between the factors current response type, preceding response type, and proportion violation, $F(1,22) = 4.30$, $p = .050$, $\eta_p^2 = .16$, as well as a four-way interaction between all factors, $F(1,22) = 7.55$, $p = .012$, $\eta_p^2 = .26$. None of the remaining effects were significant, *F*s < 2.02, *p*s > .169.

To break down this complex pattern of results, i.e., to follow up on the significant four-way interaction, I split the analysis and report the data separately for each proportion order. For this follow-up test I thus conducted two 2×2×2 ANOVAs with current response type (rule-based vs. violation), preceding response type, and proportion violation (low-PV vs. high-PV) as within-subject factors.

Movement times, low-PV-first.

A significant effect of current response type, $F(1,11) = 53.93$, $p < .001$, $\eta_p^2 = .83$, was driven by slower responses for violations (665ms) than for rule-based behavior (622ms). The interaction between current response type and proportion violation was also significant, $F(1,11) = 13.47$, $p = .004$, $\eta_p^2 = .55$, with a stronger effect of violations in high-PV blocks ($\Delta = 64$ms) compared to low-PV blocks ($\Delta = 21$ms). The interaction between preceding response type and current response type was significant, $F(1,11) = 5.55$, $p = .038$, $\eta_p^2 = .34$, with a stronger effect of violations after rule-based responses ($\Delta = 62$ms) compared to after violation responses ($\Delta = 23$ms). Finally, the three-way interaction between preceding response type, current response type and proportion violation was significant, $F(1,11) = 13.64$, $p = .004$, $\eta_p^2 = .55$, with a significant interaction between preceding and current response type for high-PV blocks, $F(1,11) = 8.86$, $p = .013$, $\eta_p^2 = .45$ (Figure 15 A2), but not for low-PV blocks, $F(1,11) = 0.77$, $p = .399$, $\eta_p^2 = .07$ (Figure 15A1). None of the remaining effects were significant, Fs < 2.65, ps > .132.

Movement times, high-PV-first.

A significant effect of current response type, $F(1,11) = 14.61$, $p = .003$, $\eta_p^2 = .57$, was driven by slower responses for violations (590ms) than for rule-based behavior (555ms). A significant effect of preceding response type, $F(1,11) = 5.73$, $p = .036$, $\eta_p^2 = .34$, described responses following rule-based behavior as slower (579ms) compared to responses following violations (567ms). Similarly, a significant main effect of proportion violation, *F*(1,11) = 10.13, *p* = .009, η_p^2 = .48, marked responses in the low-PV condition as faster (550ms) compared to the high-PV condition (596ms). The interaction between current response type and proportion violation was marginally significant, *F*(1,11) = 4.35, *p* = .061, η_p^2 = .28, with a stronger effect of violations in high-PV blocks (Δ = 46ms) compared to low-PV blocks (Δ = 25ms). The interaction between preceding response type and current response type was significant, *F*(1,11) = 11.01, *p* = .007, η_p^2 = .50, with a stronger effect of violations after rule-based responses (Δ = 59ms) compared to after violation responses (Δ = 12ms). Finally, the three-way interaction was not significant, *F*(1,11) = 0.20, *p* = .665, η_p^2 = .02, with similar interactions for both, low-PV and high-PV conditions (Figure 15B). None of the remaining effects were significant, Fs < 1, ps > .707.

Areas under the curve.

A significant effect of current response type, *F*(1,22) = 47.98, *p* < .001, η_p^2 = .69, was driven by more direct responses for rule-based behavior (27179px^2) than for violations (35868px^2). There was a significant effect of preceding response type, *F*(1,22) = 5.37, *p* = .030, η_p^2 = .20, with more curved responses after a rule-based response (32473px^2) compared to after a violation response (30574px^2). An interaction between preceding response type and proportion order, *F*(1,22) = 6.95, *p* = .015, η_p^2 = .24, indicated no post-violation effects for participants in the low-PV-first group (Δ = 32px^2), but a post-violation benefit in the high-PV-first group (Δ = -3830px^2). Proportion order interacted with proportion violation, *F*(1,22) = 10.24, *p* = .004, η_p^2 = .32, with benefits in low-PV blocks for participants who started with the low-PV condition (Δ = 4442px^2) but costs for those who started with the high-PV condition (Δ = -5568px^2). Also, there was an interaction between current response type and preceding response type, F(1,22) = 28.68, p < .001, η_p^2 = .57, with a stronger violation effect after rule-based responses (Δ = 14844px^2) than after violation responses (Δ = 2536px^2). As with MTs, there was a three-way interaction between the factors current response type, preceding response type, and proportion violation, F(1,22) = 10.92, p = .003, η_p^2 = .33, as well as a four-way interaction between all factors, F(1,22) = 7.69, p = .011, η_p^2 = .26. Accordingly, I again conducted follow-up analyses for each proportion order. None of the remaining effects were significant, Fs < 1.51, ps > .232.

Areas under the curve, low-PV-first.

A significant effect of current response type, F(1,11) = 23.03, p = .001, η_p^2 = .68, was driven by more contorted responses for violations (39794px^2) than for rule-based behavior (30245px^2). Similarly, a marginally significant main effect of proportion violation, F(1,11) = 4.51, p = .057, η_p^2 = .29, marked responses in the low-PV-condition as less contorted (32799px^2) than in the high-PV-condition (37240px^2). The interaction between preceding response type and current response type was significant, F(1,11) = 9.69, p = .010, η_p^2 = .47, with a stronger effect of violations after rule-based responses (Δ = 15323px^2) compared to after violation responses (Δ = 3774px^2). Finally, the three-way interaction between preceding response type, current response type and proportion violation was significant, F(1,11) = 11.27, p = .006, η_p^2 = .51, with a significant interaction between preceding and current response type for high-PV blocks, F(1,11) = 15.47, p = .002, η_p^2 = .58 (Figure 15C2), but not for low-PV blocks,

$F(1,11) = 0.16$, $p = .692$, $\eta_p^2 = .02$ (Figure 15C1). None of the remaining effects were significant, Fs < 3.01, ps > .111.

Areas under the curve, high-PV first.

A significant effect of current response type, $F(1,11) = 26.25$, $p < .001$, $\eta_p^2 = .71$, was driven by more contorted responses for violations (31942px^2) than for rule-based behavior (24112px^2). A significant effect of preceding response type, $F(1,11) = 9.61$, $p = .010$, $\eta_p^2 = .47$, described responses following rule-based behavior as more contorted (29942px^2) compared to responses following violations (26112px^2). Similarly, a significant main effect of proportion violation, $F(1,11) = 5.73$, $p = .036$, $\eta_p^2 = .34$, marked responses in the low-PV-condition as more contorted (30811px^2) compared to the high-PV-condition (25243px^2). The interaction between preceding response type and current response type was significant, $F(1,11) = 23.21$, $p = .001$, $\eta_p^2 = .68$, with a stronger effect of violations after rule-based responses ($\Delta = 14356$px^2) compared to after violation responses ($\Delta = 1297$px^2). Finally, the three-way interaction was not significant, $F(1,11) = 0.39$, $p = .547$, $\eta_p^2 = .03$, with similar interactions for both, low-PV and high-PV conditions (Figure 15D). None of the remaining effects were significant, Fs < 1, ps > .951.

Discussion.

In Experiment 9, I tested how rule violation performance changes as a function of frequency and recency of rule violations. Participants were confronted with blocks that contained either 25% or 75% violation trials, and at first sight, the pattern of results seems rather complex. However, when regarded through the lens of adaptation processes to frequency (in terms of interactions between current response type and proportion violation) and to recency (in terms of interactions between preceding and current response type), the data allows for interesting conclusions. First, the frequency manipulation had no direct effect on spatial parameters of response execution, while the temporal measure was unexpectedly affected in a way opposite to what I anticipated. In blocks with a high frequency of violations, response costs for violations increased compared to blocks with a low frequency of violations.

Next, I found that the order in which proportions of violations were experienced, shaped the way in which violation recency affected responding. Participants who started with a low proportion of violations did not show adaptations to recent violations, replicating my previous results (Experiment 1). However, as soon as these participants encountered a high proportion of violations (in the second half), they did also show adaptation to immediately preceding violations. On the other hand, participants who started with a high proportion of violations showed adaptation to recent violations right away, and they continued to do so when proportion of violation dropped. It seems as if also for violation tasks (as for the negation task, Experiment 8), frequency and recency adaptations are not independent mechanisms, but recency adaptations only emerge once a high frequency of violations has been experienced.

4.3 Experiment 10.

Next to replicating the interplay between frequency and recency of violations, I aimed at testing for another factor that might reduce response costs for rule violations: transfer effects that are triggered by a related, but separate task.

Numerous studies show that conflict effects in one task are reduced when following a different task that recruits executive control, and transfer between separate tasks is strong when they are maximally similar or maximally dissimilar (Notebaert & Verguts, 2008; for a review, see Braem, Abrahamse, Duthoo, & Notebaert, 2014).

The reasons for such transfer between tasks are not entirely settled, but they might relate to negative affect which comes with both, interference (Dreisbach & Fischer, 2012) as well as with rule breaking (Chapter 3). Negative affect is thought to serve as internal signal which prompts a stronger focus on task-relevant information. Consequently, if subjects experience conflict in a different task, this might result in a stronger focus on task-relevant information such that, when subsequently breaking a rule, performance is less affected by the irrelevant option of obeying to the rule in terms of delays of responding and spatial attraction towards the rule-consistent response location.

We designed a Simon-task that closely resembles the Rule-task of Experiment 9 (Simon, 1990). In these trials, one of the target areas changed to either red or green, and the color (not the location) indicated if a movement to the left or to the right had to be executed. This resulted in congruent trials (moving towards the colored target area) and incongruent trials (moving away from the colored target area). If conflict adaptation in a Simon-task transfers to the Rule-task, response costs for violations should be smaller after incongruent rather than congruent Simon trials. Of course this transfer might work the other way round as well, such that breaking a rule could reduce the cost of spatial incongruency in a subsequent Simon-task.

Methods.

Participants.

A new set of forty-eight participants was recruited (mean age = 26.1 years, *SD* = 5.3, 16 male, 4 left-handed) and received either course credit or €8 monetary compensation. All participants gave informed consent, were naïve to the purpose of the experiment and were debriefed after the session.

Apparatus, stimuli and procedure.

The experiment was mostly identical to the first experiment. But intermixed with the rule task, half of the trials now employed a Simon-task. In these trials, one of the target areas turned either red or green as soon as they appeared, and participants had to respond to the color by a movement to the left or the right. The location of the colored target area was irrelevant to the task, which resulted in either S-R congruent trials (moving towards the color stimulus) or S-R incongruent trials (moving away from the color stimulus). Again, the S-R mapping for the Simon trials was displayed before a movement started, together with the written instruction “Color” (German: “Farbe”). Simon trials employed 50% congruent and 50% incongruent trials throughout the experiment, whereas the original Rule-task trials were still subject to the proportion violation manipulation. All trials within a block were presented in randomized order.

Results.

Data treatment and analyses.

The data was treated exactly as in Experiment 9. Accordingly, I again omitted trials in which participants failed to act according to the instruction or failed to hit any of the two target areas at all (4.4%), trials following an error (3.5%), and outliers (5.1%). I then analyzed each measure separately for each possible trial sequence (Rule task → Rule-task, Rule-task → Simon-task, Simon-task → Rule-task, Simon-task → Simon-task; see Figure 17 - Figure 23). Analyses were performed in terms of separate 2×2×2×2 ANOVAs with current response type (rule-based vs. violation for the Rule-task; congruent vs. incongruent for the Simon-task), preceding response type, and proportion violation (low-PV vs. high-PV) as within-subject factors, and PV order (low-PV-first vs. high-PV-first) as a between-subjects factor. Again, if an ANOVA produced a higher-order interaction, the analysis was broken up into less complex analyses (as in Experiment 9) to improve accessibility of the data.

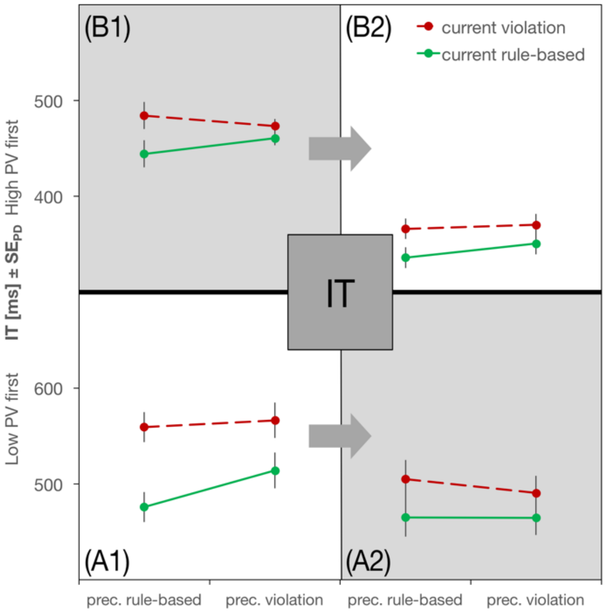

Figure 16. Results of Experiment 10 (Rule-task→Rule-task, ITs).
Initiation times (IT) are plotted as a function of preceding response type (abscissa), current response type (continuous, green line for rule-based responses; dashed, red line for violation responses), and the current proportion of violations (PV; white background for low-PV, gray background for high-PV). Further, the figure is split by proportion order: The lower panels

(A) represent the low-PV-first condition, the upper panels (B) represent the high-PV-first condition. Panels with the number one represent the first half of the experiment per proportion order, panels with the number two represent the second half. Error bars represent standard errors of paired differences (SE_{PD}), calculated separately for each instance of preceding response type (Pfister & Janczyk, 2013).

Rule-task→Rule-task sequences, initiation times.

A significant effect of current response type, $F(1,46) = 18.56$, $p < .001$, $\eta_p^2 = .29$, was driven by faster response initiation for rule-based behavior (437ms) than for violations (475ms). Also, there was an effect of preceding response type, $F(1,46) = 4.98$, $p = .010$, $\eta_p^2 = .29$, marking response initiations following a rule-based response as faster (453ms) than those following a violation response (459ms). Similarly, a marginally significant effect of proportion violation, $F(1,46) = 3.74$, $p = .059$, $\eta_p^2 = .08$, described response initiations in the low-PV condition as faster (441ms) compared to those in the high-PV condition (471ms). Proportion order interacted with proportion violation, $F(1,46) = 24.57$, $p < .001$, $\eta_p^2 = .35$, with costs for low-PV blocks for participants who started with the low-PV condition (Δ = -48ms) but benefits for those who started with the high-PV condition (Δ = 109ms). An interaction between preceding response type and proportion violation, $F(1,46) = 14.10$, $p < .001$, $\eta_p^2 = .24$, indicated post-violation slowing in the low-PV condition (Δ = 16ms), but post-violation speeding in the high-PV condition (Δ = -3ms). Similarly, there was a marginally significant interaction between current response type and proportion violation, $F(1,46) = 3.54$, $p = .066$, $\eta_p^2 = .07$, with a larger violation effect in low-PV blocks (Δ = 46ms) compared to high-PV blocks (Δ =

29ms). Also, there was an interaction between current response type and preceding response type, $F(1,46) = 7.83$, $p = .007$, $\eta_p^2 = .15$, with a stronger violation effect after rule-based responses (Δ = 47ms) than after violation responses (Δ = 27ms). This interaction held true for all combinations of proportion violation and proportion order (Figure 16), as indicated by all higher-order interactions including both factors returning non-significant results, Fs < 1.22, ps > .276. None of the remaining effects were significant, Fs < 2.36, ps > .132.

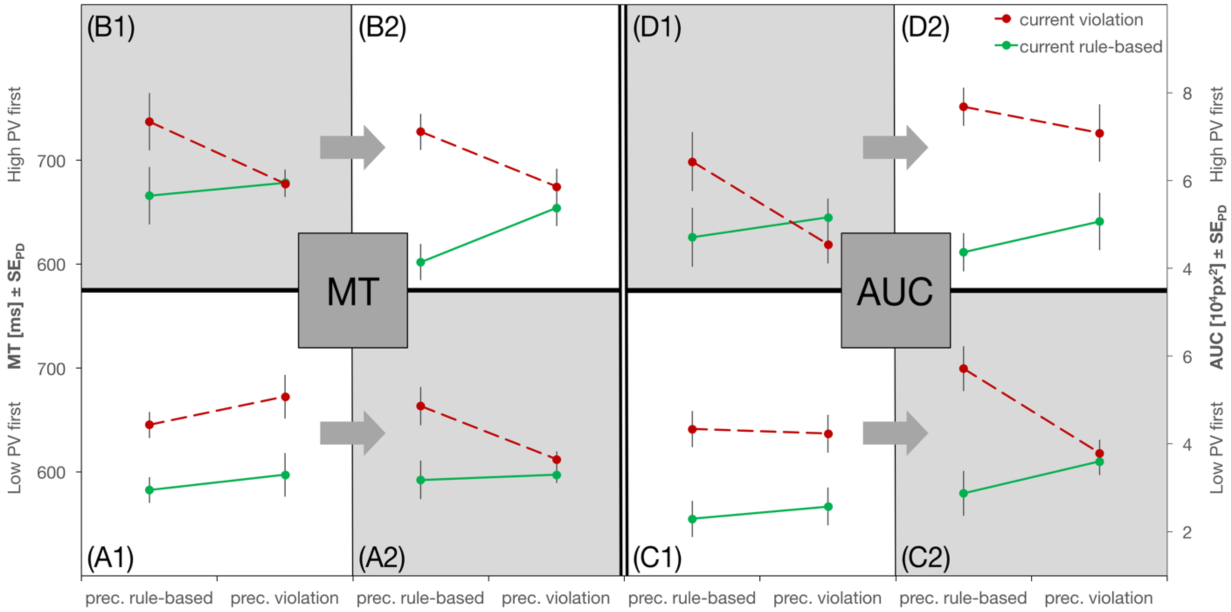

Figure 17. Results of Experiment 10 (Rule-task→Rule-task, MTs & AUCs).

Movement times (MT; left) and areas under the curve (AUC; right) are plotted as a function of preceding response type (abscissa), current response type (continuous, green line for rule-based responses; dashed, red line for violation responses), and the current proportion of violations (PV; white background for low-PV, gray background for high-PV). Further, the figure is split by proportion order: The lower panels (A & C) represent the low-PV-first condition, the upper panels (B & D) represent the high-PV-first condition. Panels with the number one represent the first half of the experiment per proportion order, panels with the number two represent the second half. Error bars represent standard errors of paired differences (SE_{PD}), calculated separately for each instance of preceding response type (Pfister & Janczyk, 2013).

Rule-task→Rule-task sequences, movement times.

A significant effect of current response type, $F(1,46) = 67.42$, $p < .001$, $\eta_p^2 = .59$, was driven by faster responses for rule-based behavior (621ms) than for violations (676ms). An interaction between preceding response type and proportion order, $F(1,46) = 11.47$, $p = .001$, $\eta_p^2 = .20$, indicated no effect of preceding response type for participants in the low-PV-first group (Δ = -1ms), but post-violation speeding in the high-PV-first group (Δ = -12ms). Current response type interacted with proportion violation, $F(1,46) = 7.73$, $p = .008$, $\eta_p^2 = .14$, with a larger violation effect in low-PV blocks (Δ = 71ms) compared to high-PV blocks (Δ = 39ms). Also, there was an interaction between current response type and preceding response type, $F(1,46) = 20.35$, $p < .001$, $\eta_p^2 = .31$, with a stronger violation effect after rule-based responses (Δ = 83ms) than after violation responses (Δ = 27ms). There was a three-way interaction between the factors current response type, preceding response type, and proportion order, $F(1,46) = 7.45$, $p = .009$, $\eta_p^2 = .14$, as well as a four-way interaction between all factors, $F(1,46) = 4.30$, $p = .044$, $\eta_p^2 = .09$, which will again be explained by splitting the analysis into separate ANOVAs for each proportion order. None of the remaining effects were significant, *F*s < 2.36, *p*s > .132.

Rule-task→Rule-task sequences, movement times, low-PV-first.

A significant effect of current response type, $F(1,23) = 30.37$, $p < .001$, $\eta_p^2 = .57$, was driven by slower responses for violations (648ms) than for rule-based behavior

(592ms). The interaction between current response type and proportion violation was marginally significant, *F*(1,23) = 3.33, *p* = .081, η_p^2 = .13, with a smaller effect of violations in high-PV blocks (Δ = 43ms) compared to low-PV blocks (Δ = 69ms). Similarly, the interaction between preceding response type and proportion violation was significant, *F*(1,23) = 9.87, *p* = .005, η_p^2 = .30, with response costs following violations in the low-PV blocks (Δ = 21ms), but a response benefit in the high-PV blocks (Δ = -23ms). Finally, the three-way interaction between preceding response type, current response type and proportion violation was significant, *F*(1,23) = 4.74, *p* = .040, η_p^2 = .17, with a significant interaction between preceding and current response type for high-PV blocks, *F*(1,23) = 6.76, *p* = .016, η_p^2 = .23 (Figure 17A2), but not for low-PV blocks, *F*(1,23) = 0.33, *p* = .574, η_p^2 = .01 (Figure 17A1). None of the remaining effects were significant, Fs < 2.10, ps > .161.

Rule-task→Rule-task sequences, movement times, high-PV-first.

A significant effect of current response type, *F*(1,23) = 38.24, *p* < .001, η_p^2 = .62, was driven by slower responses for violations (704ms) than for rule-based behavior (650ms). A significant effect of preceding response type, *F*(1,23) = 5.03, *p* = .035, η_p^2 = .18, described responses following rule-based behavior as slower (683ms) compared to responses following violations (671ms). The interaction between current response type and proportion violation was significant, *F*(1,23) = 4.42, *p* = .047, η_p^2 = .16, with a smaller effect of violations in high-PV blocks (Δ = 35ms) compared to low-PV blocks (Δ = 73ms). Also, the interaction between preceding response type and current response type was significant, *F*(1,23) = 21.07, *p* < .001, η_p^2 = .48, with a stronger effect of violations after rule-based responses (Δ = 98ms) compared to after violation responses

(Δ = 10ms). Finally, the three-way interaction was not significant, $F(1,23) = 0.77$, $p = .390$, $\eta_p^2 = .03$, with similar interactions for both, low-PV and high-PV conditions (Figure 17B). None of the remaining effects were significant, Fs < 2.71, ps > .113.

Rule-task→Rule-task sequences, areas under the curve.

A significant effect of current response type, $F(1,46) = 66.35$, $p < .001$, $\eta_p^2 = .59$, was driven by more direct responses for rule-based behavior (38280px^2) than for violations (54724px^2). There was a significant effect of preceding response type, $F(1,46) = 6.61$, $p = .013$, $\eta_p^2 = .13$, with more curved responses after a rule-based response (47974px^2) compared to after a violation response (45030px^2). An interaction between current response type and proportion violation, $F(1,46) = 10.04$, $p = .003$, $\eta_p^2 = .18$, further indicated larger violation effects in the low-PV condition (Δ = 22567px^2), compared to the high-PV condition (Δ = 10323px^2). Similarly, an interaction between preceding response type and proportion violation, $F(1,46) = 7.31$, $p = .010$, $\eta_p^2 = .14$, indicated only small post-violation effects for participants in the low-PV-first condition (Δ = 711px^2), but a post-violation benefit in the high-PV-first group (Δ = -6599px^2). Proportion order interacted with proportion violation, $F(1,46) = 10.53$, $p = .002$, $\eta_p^2 = .19$, with benefits in low-PV blocks for participants who started with the low-PV condition (Δ = 6347px^2) but costs for those who started with the high-PV condition (Δ = -8382px^2). Also, there was an interaction between current response type and preceding response type, $F(1,46) = 31.25$, $p < .001$, $\eta_p^2 = .41$, with a stronger violation effect after rule-based responses (Δ = 47974px^2) than after violation responses (Δ = 45030px^2). Finally, there was a three-way interaction between the factors current response type, preceding response type, and proportion violation, $F(1,46) = 7.59$, $p = .008$, $\eta_p^2 = .14$, as well as another three-way interaction between current response type, proportion violation and proportion order, $F(1,46) = 5.27$, $p = .026$, $\eta_p^2 = .10$, which is why the data will again be reanalyzed separately for each proportion order. None of the remaining effects were significant, *F*s < 1.06, *p*s > .308.

Rule-task→Rule-task sequences, areas under the curve, low-PV-first.

A significant effect of current response type, $F(1,23) = 48.61$, $p < .001$, $\eta_p^2 = .68$, was driven by more contorted responses for violations (45179px^2) than for rule-based behavior (28359px^2). Similarly, an significant main effect of proportion violation, $F(1,23) = 5.89$, $p = .024$, $\eta_p^2 = .20$, marked responses in the low-PV condition as less contorted (33595px^2) than in the high-PV condition (39942px^2). Proportion violation and preceding response type combined in a marginally significant interaction, $F(1,23) = 3.11$, $p = .091$, $\eta_p^2 = .12$, with only small response costs after a violation in the low-PV condition (Δ = 888px^2) but a response benefit after a violation in the high-PV condition (Δ = -6029px^2). The interaction between preceding response type and current response type was significant, $F(1,23) = 27.87$, $p < .001$, $\eta_p^2 = .55$, with a stronger effect of violations after rule-based responses (Δ = 24400px^2) compared to after violation responses (Δ = 9241px^2). Finally, the three-way interaction between preceding response type, current response type and proportion violation was significant, $F(1,23) = 10.14$, $p = .004$, $\eta_p^2 = .31$, with a significant interaction between preceding and current response type for high-PV blocks, $F(1,23) = 34.26$, $p < .001$, $\eta_p^2 = .60$ (Figure 17C2), but not for low-PV blocks, $F(1,23) = 0.68$, $p = .417$, $\eta_p^2 = .03$ (Figure 17C1). None of the remaining effects were significant, Fs < 2.49, ps > .128.

Rule-task→Rule-task sequences, areas under the curve, high-PV-first.

A significant effect of current response type, $F(1,23) = 24.63$, $p < .001$, $\eta_p^2 = .52$, was driven by more contorted responses for violations (64269px^2) than for rule-based behavior (48200px^2). A marginally significant effect of preceding response type, $F(1,23) = 4.25$, $p = .051$, $\eta_p^2 = .16$, described responses following rule-based behavior as more contorted (57894px^2) compared to responses following violations (54576px^2). Similarly, a significant main effect of proportion violation, $F(1,23) = 5.10$, $p = .034$, $\eta_p^2 = .18$, marked responses in the low-PV condition as more contorted (60426px^2) compared to the high-PV condition (52044px^2). There was an interaction between current response type and proportion violation, $F(1,23) = 13.54$, $p = .001$, $\eta_p^2 = .37$, with a larger violation effect in the low-PV condition ($\Delta = 26626$px^2) compared to the high-PV condition ($\Delta = 5512$px^2). Similarly, preceding response type and proportion violation entered an interaction, $F(1,23) = 4.28$, $p = .050$, $\eta_p^2 = .16$, with only small response costs after a violation in the low-PV condition ($\Delta = 533$px^2) but a response benefit after a violation in the high-PV condition ($\Delta = -7169$px^2). The interaction between preceding response type and current response type was significant, $F(1,23) = 12.08$, $p = .002$, $\eta_p^2 = .34$, with a stronger effect of violations after rule-based responses ($\Delta = 25138$px^2) compared to after violation responses ($\Delta = 7000$px^2). Finally, the three-way interaction was not significant, $F(1,23) = 1.15$, $p = .295$, $\eta_p^2 = .05$, with similar interactions for both, low-PV and high-PV conditions (Figure 17D).

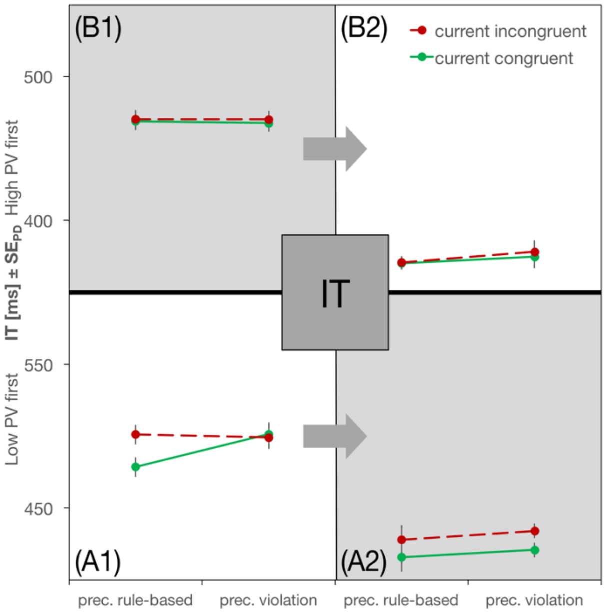

Figure 18. Results of Experiment 10 (Rule-task→Simon-task, ITs). Initiation times (IT) are plotted as a function of preceding response type (abscissa), current response type (continuous, green line for rule-based responses; dashed, red line for violation responses), and the current proportion of violations (PV; white background for low-PV, gray background for high-PV). Further, the figure is split by proportion order: The lower panels (A) represent the low-PV-first condition, the upper panels (B) represent the high-PV-first condition. Panels with the number one represent the first half of the experiment per proportion order, panels with the number two represent the second half. Error bars represent standard errors of paired differences (SE_{PD}), calculated separately for each instance of preceding response type (Pfister & Janczyk, 2013).

Rule-task→Simon-task sequences, initiation times.

A significant effect of current response type, *F*(1,46) = 5.14, *p* = .028, η_p^2 = .10, was driven by faster response initiations for congruent (412ms) than for incongruent responses (419ms). Also, there was a marginally significant effect of preceding response type, *F*(1,46) = 3.74, *p* = .059, η_p^2 = .08, with slower response initiations following violations (418ms) compared to rule-based responses (413ms).

There was a significant interaction between proportion violation and proportion order, $F(1,46) = 48.06$, $p < .001$, $\eta_p^2 = .51$, with costs in low-PV blocks for participants who started with the low-PV condition (Δ = -70ms) but benefits for those who started with the high-PV condition (Δ = 96ms). The interaction between current response type and preceding response type was not significant, $F(1,46) = 1.38$, $p = .246$, $\eta_p^2 = .03$ (Figure 18), and neither were any of the higher-order interactions including both factors, Fs < 2.67, ps > .108. None of the remaining effects were significant, Fs < 2.53, ps > .119.

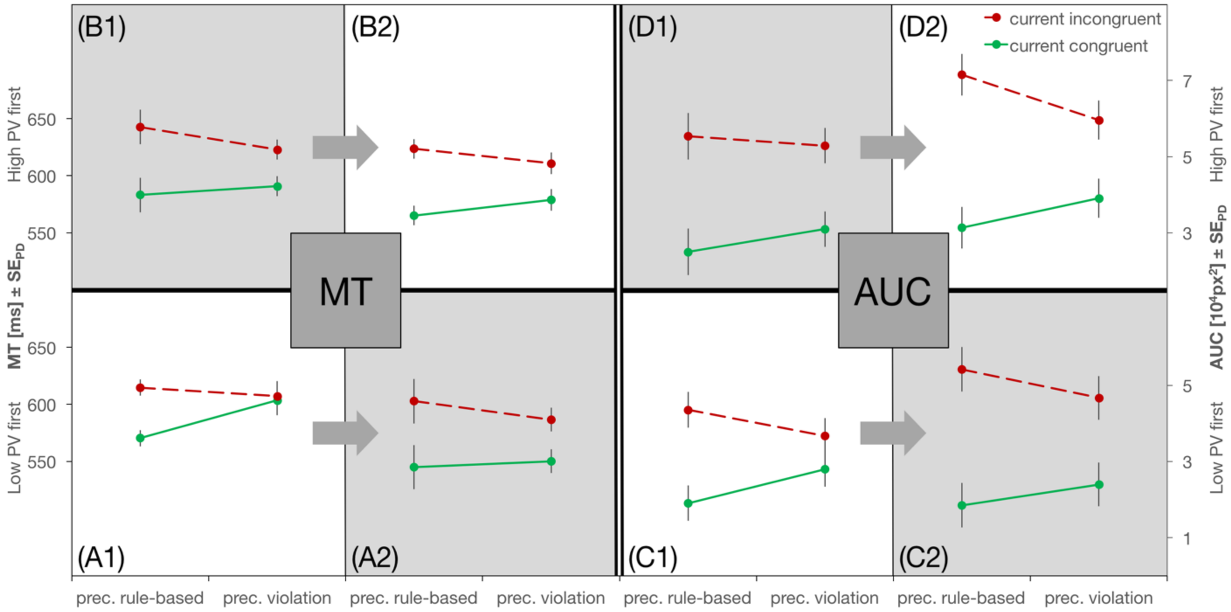

Figure 19. Results of Experiment 10 (Rule-task→Simon-task, MTs & AUCs)

Movement times (MT; left) and areas under the curve (AUC; right) are plotted as a function of preceding response type (abscissa), current response type (continuous, green line for congruent; dashed, red line for incongruent), and the current proportion of violations (PV; white background for low-PV, gray background for high-PV). Further, the figure is split by proportion order: The lower panels (A & C) represent the low-PV-first condition, the upper panels (B & D) represent the high-PV-first condition. Panels with the number one represent the first half of the experiment per proportion order, panels with the number two represent the second half. Error bars represent standard errors of paired differences (SE_{PD}), calculated separately for each instance of preceding response type (Pfister & Janczyk, 2013).

Rule-task→Simon-task sequences, movement times.

A significant effect of current response type, $F(1,46) = 53.78$, $p < .001$, $\eta_p^2 = .53$, was driven by faster responses for congruent (573ms) than for incongruent responses (614ms). There was a marginally significant interaction between proportion violation and proportion order, $F(1,46) = 3.56$, $p = .066$, $\eta_p^2 = .71$, with costs in low-PV blocks for participants who started with the low-PV condition (Δ = -28ms) but benefits for those who started with the high-PV condition (Δ = 15ms). Also, there was an interaction between current response type and preceding response type, $F(1,46) = 14.27$, $p < .001$, $\eta_p^2 = .24$, with a stronger congruency effect after rule-based responses (Δ = 55ms) than after violation responses (Δ = 26ms). This interaction held true for all combinations of proportion violation and proportion order (Figure 19, A & B), as indicated by all higher-order interactions including both factors returning non-significant results, *F*s < 1, *p*s > .451. None of the remaining effects were significant, *F*s < 1.71, *p*s > .198.

Rule-task→Simon-task sequences, areas under the curve.

A significant effect of current response type, $F(1,46) = 69.18$, $p < .001$, $\eta_p^2 = .60$, was driven by more direct responses for congruent (26996px^2) than for incongruent trials (52593px^2). Proportion order interacted with proportion violation, $F(1,46) = 9.62$, $p = .003$, $\eta_p^2 = .17$, with benefits in low-PV blocks for participants who started with the low-PV condition (Δ = 4005px^2) but costs for those who started with the high-PV condition (Δ = -9322px^2). Also, there was an interaction between current response type and preceding response type, $F(1,46) = 31.22$, $p < .001$, $\eta_p^2 = .40$, with a stronger congruency effect after rule-based responses ($\Delta = 32702px^2$) than after violation responses ($\Delta = 18492px^2$). This interaction held true for all combinations of proportion violation and proportion order (Figure 19, C & D), as indicated by all higher-order interactions including both factors returning non-significant results, *F*s < 1.21, *p*s > .278. None of the remaining effects were significant, *F*s < 2.12, *p*s > .149.

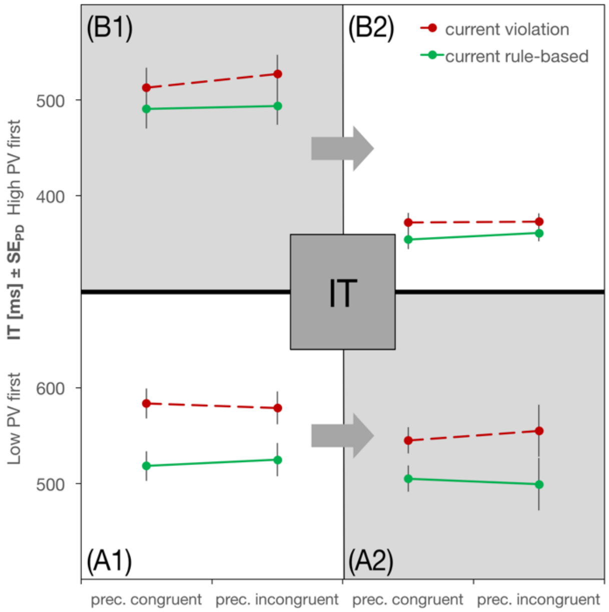

Figure 20. Results of Experiment 10 (Simon-task→Rule-task, ITs). Initiation times (IT) are plotted as a function of preceding response type (abscissa), current response type (continuous, green line for rule-based responses; dashed, red line for violation responses), and the current proportion of violations (PV; white background for low-PV, gray background for high-PV). Further, the figure is split by proportion order: The lower panels (A) represent the low-PV-first condition, the upper panels (B) represent the high-PV-first condition. Panels with the number one represent the first half of the experiment per proportion order, panels with the number two represent the second half. Error bars represent standard errors of paired differences (SE_{PD}), calculated separately for each instance of preceding response type (Pfister & Janczyk, 2013).

Simon-task→Rule-task sequences, initiation times.

A significant effect of current response type, $F(1,46) = 16.92$, $p < .001$, $\eta_p^2 = .27$, was driven by faster response initiations for rule- based (469ms) than for violation responses (506ms). There was an effect of proportion violation, $F(1,46) = 9.73$, $p = .003$, $\eta_p^2 = .18$, with faster response initiations in the low-PV blocks (458ms) compared to the high-PV blocks (516ms). Current response type and proportion order entered a marginally significant interaction, $F(1,46) = 3.13$, $p = .084$, $\eta_p^2 = .06$, with larger violation effects for participants who started with the low-PV condition (Δ = 54ms) compared to those who started with the high-PV condition (Δ = 21ms). There was a significant interaction between proportion violation and proportion order, $F(1,46) = 20.11$, $p < .001$, $\eta_p^2 = .30$, with costs in low-PV blocks for participants who started with the low-PV condition (Δ = -25ms) but benefits for those who started with the high-PV condition (Δ = 141ms). The interaction between current response type and preceding response type was not significant, $F(1,46) = 0.08$, $p = .776$, $\eta_p^2 = .00$, and this held true for all combinations of proportion violation and proportion order (Figure 20), as indicated by all higher-order interactions including both factors returning non-significant results, Fs < 1.64, ps > .206. None of the remaining effects were significant, Fs < 1.40, ps > .243.

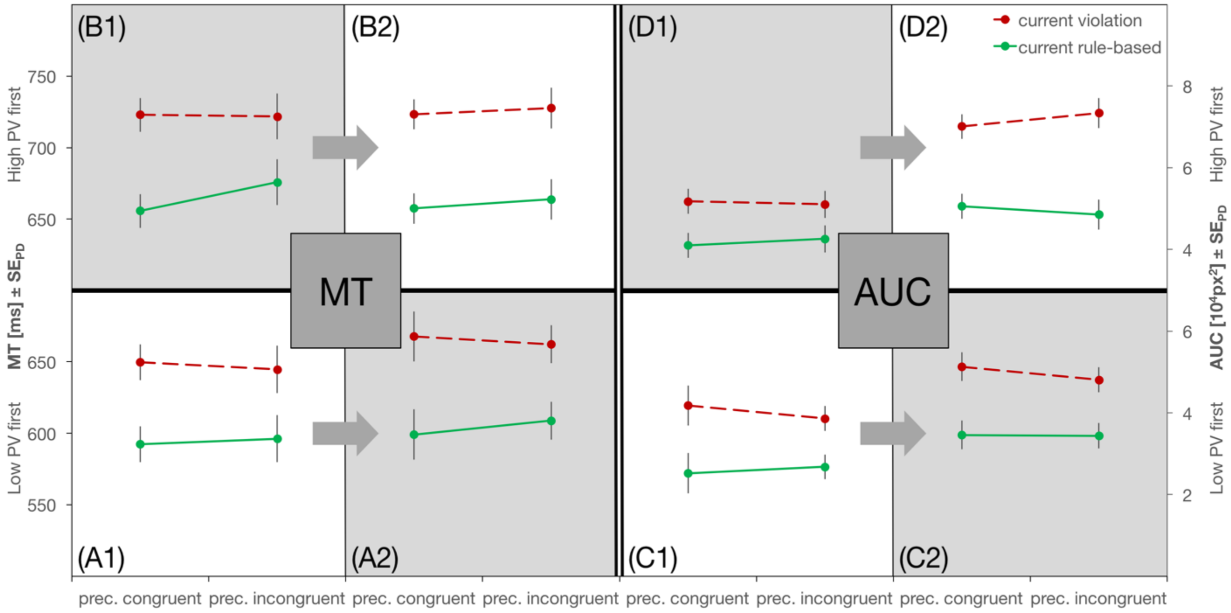

Figure 21. Results of Experiment 10 (Simon-task→Rule-task, MTs & AUCs).

Movement times (MT; left) and areas under the curve (AUC; right) are plotted as a function of preceding response type (abscissa), current response type (continuous, green line for rule-based responses; dashed, red line for violation responses), and the current proportion of violations (PV; white background for low-PV, gray background for high-PV). Further, the figure is split by proportion order: The lower panels (A & C) represent the low-PV-first condition, the upper panels (B & D) represent the high-PV-first condition. Panels with the number one represent the first half of the experiment per proportion order, panels with the number two represent the second half. Error bars represent standard errors of paired differences (SE$_{PD}$), calculated separately for each instance of preceding response type (Pfister & Janczyk, 2013).

Simon-task→Rule-task sequences, movement times.

A significant effect of current response type, $F(1,46) = 73.26$, $p < .001$, $\eta_p^2 = .61$, was driven by faster responses for rule-based (631ms) than for violation responses (690ms). The interaction between current response type and preceding response type was not significant, $F(1,46) = 1.97$, $p = .167$, $\eta_p^2 = .04$, and this held true for all combinations of proportion violation and proportion order (Figure 21, A & B), as indicated by all higher-order interactions including both factors returning non-significant results, *F*s < 1, *p*s > .460. None of the remaining effects were significant, *F*s < 1.06, *p*s > .309.

Simon-task→Rule-task sequences, areas under the curve.

A significant effect of current response type, *F*(1,46) = 69.32, *p* < .001, η_p^2 = .60, was driven by more direct responses for rule-based (37959px^2) than for violation trials (53276px^2). Proportion order interacted with proportion violation, *F*(1,46) = 21.91, *p* < .001, η_p^2 = .32, with benefits in low-PV blocks for participants who started with the low-PV condition (Δ = 8990px^2) but costs for those who started with the high-PV condition (Δ = -14054px^2). Also, there was an interaction between current response type and proportion violation, *F*(1,46) = 8.26, *p* = .006, η_p^2 = .15, with a stronger violation effect in low-PV blocks (Δ = 18212px^2) compared to high-PV blocks (Δ = 12422px^2). The interaction between current response type and preceding response type was not significant, *F*(1,46) = 0.34, *p* = .562, η_p^2 = .01, and this held true for all combinations of proportion violation and proportion order (Figure 21, C & D), as indicated by all higher-order interactions including both factors returning non-significant results, *F*s < 1.90, *p*s > .174. None of the remaining effects were significant, *F*s < 1.43, *p*s > .237.

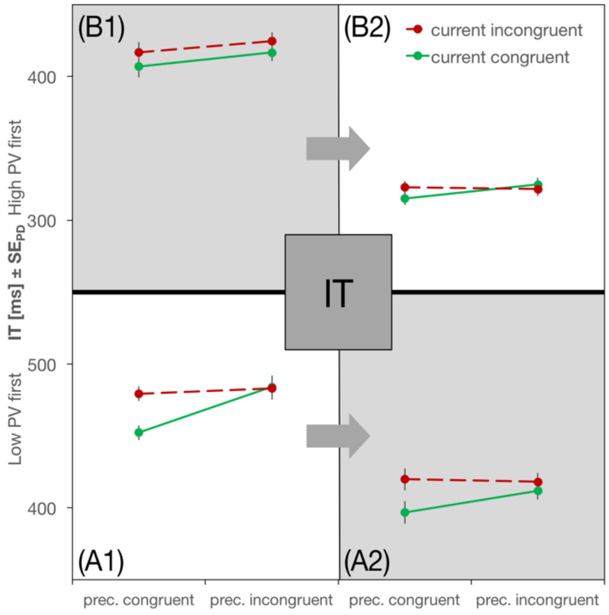

Figure 22. Results of Experiment 10 (Simon-task→Simon-task, ITs). Initiation times (IT) are plotted as a function of preceding response type (abscissa), current response type (continuous, green line for rule-based responses; dashed, red line for violation responses), and the current proportion of violations (PV; white background for low-PV, gray background for high-PV). Further, the figure is split by proportion order: The lower panels (A) represent the low-PV-first condition, the upper panels (B) represent the high-PV-first condition. Panels with the number one represent the first half of the experiment per proportion order, panels with the number two represent the second half. Error bars represent standard errors of paired differences (SE$_{PD}$), calculated separately for each instance of preceding response type (Pfister & Janczyk, 2013).

Simon-task→Simon-task sequences, initiation times.

A significant effect of current response type, $F(1,46) = 16.40$, $p < .001$, $\eta_p^2 = .26$, was driven by faster response initiations for congruent (400ms) than for incongruent responses (410ms). There was a significant effect of preceding response type, $F(1,46) = 18.75$, $p < .001$, $\eta_p^2 = .29$, with slower response initiations following incongruent (409ms) compared to congruent responses (400ms). Also, there was a significant interaction between proportion violation and proportion order, $F(1,46) = 54.13$, $p < .001$, $\eta_p^2 = .54$, with costs in low-PV blocks for participants who started with the low-PV condition (Δ = -63ms) but benefits for those who started with the high-PV condition (Δ = 92ms). There was an interaction between current response type and preceding response type, $F(1,46) = 9.51$, $p = .003$, $\eta_p^2 = .17$, with a stronger congruency effect after congruent responses (Δ = 17ms) than after incongruent responses (Δ = 3ms). This interaction held true for all combinations of proportion violation and proportion order (Figure 22), as indicated by all higher-order interactions including both factors returning non-significant results, Fs < 1.48, ps > .231. None of the remaining effects were significant, Fs < 2.42, ps > .127.

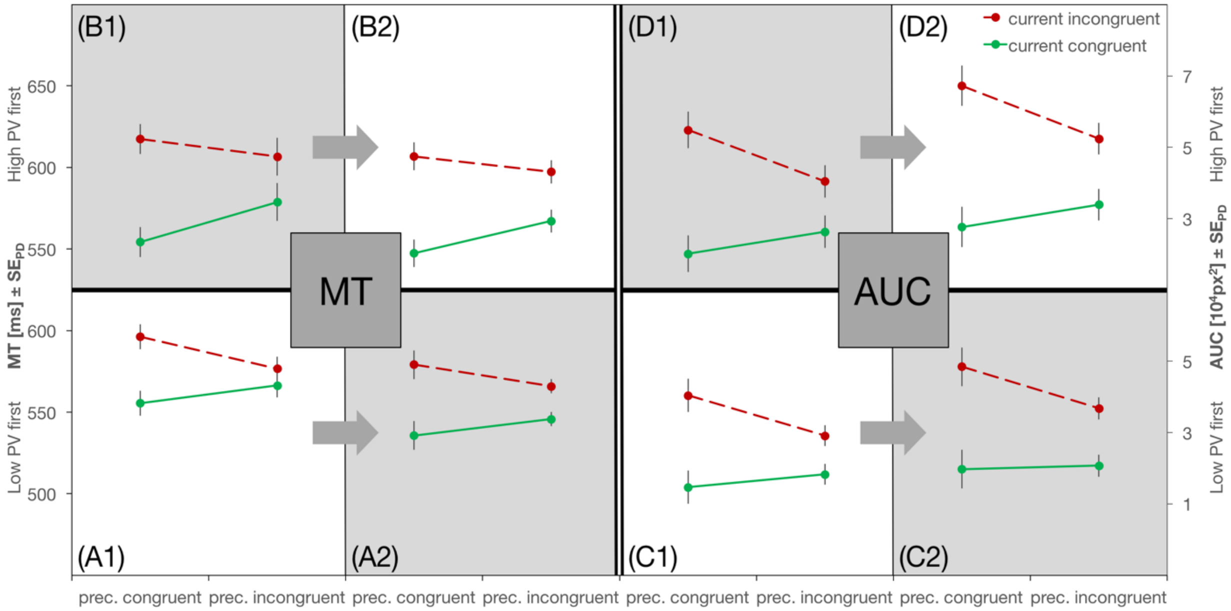

Figure 23. Results of Experiment 10 (Simon-task→Simon-task, MTs & AUCs).

Movement times (MT; left) and areas under the curve (AUC; right) are plotted as a function of preceding response type (abscissa), current response type (continuous, green line for congruent; dashed, red line for incongruent), and the current proportion of violations (PV; white background for low-PV, gray background for high-PV). Further, the figure is split by proportion order: The lower panels (A & C) represent the low-PV-first condition, the upper panels (B & D) represent the high-PV-first condition. Panels with the number one represent the first half of the experiment per proportion order, panels with the number two represent the second half. Error bars represent standard errors of paired differences (SE_{PD}), calculated separately for each instance of preceding response type (Pfister & Janczyk, 2013).

Simon-task→Simon-task sequences, movement times.

A significant effect of current response type, $F(1,46) = 66.03$, $p < .001$, $\eta_p^2 = .59$, was driven by faster responses for congruent (556ms) than for incongruent responses (593ms). There was a marginally significant interaction between current response type and proportion order, $F(1,46) = 3.32$, $p = .075$, $\eta_p^2 = .07$, with smaller congruency effects for participants who started with the low-PV condition ($\Delta = 29$ms) compared to participants who started with the high-PV condition ($\Delta = 45$ms). Similarly, a marginally significant interaction between preceding response type and proportion order emerged, $F(1,46) = 3.34$, $p = .074$, $\eta_p^2 = .07$, with slight post-conflict slowing for participants that started with the low-PV ($\Delta = 3$ms), but post-conflict speeding for participants who started with the high-PV ($\Delta = -6$ms). Also, there was an interaction between current response type and preceding response type, $F(1,46) = 43.90$, $p < .001$, $\eta_p^2 = .49$, with a stronger congruency effect after congruent responses ($\Delta = 52$ms) than after incongruent responses ($\Delta = 22$ms). This interaction held true for all combinations of proportion violation and proportion order (Figure 23, A & B), as indicated by all higher-order interactions including both factors returning non-significant results, *F*s < 1, *p*s > .421. None of the remaining effects were significant, *F*s < 1.85, *p*s > .181.

Simon-task→Simon-task sequences, areas under the curve.

A significant effect of current response type, $F(1,46) = 66.38$, $p < .001$, $\eta_p^2 = .59$, was driven by more direct responses for congruent (22666px^2) than for incongruent trials (46194px^2). Also, there was a significant effect of preceding response type, $F(1,46) = 24.04$, $p < .001$, $\eta_p^2 = .34$, with faster responses after incongruent (32232px^2) compared to after congruent responses (36628px^2). Proportion order interacted with proportion violation, $F(1,46) = 18.75$, $p < .001$, $\eta_p^2 = .29$, with benefits in low-PV blocks for participants who started with the low-PV condition (Δ = 5777px^2) but costs for those who started with the high-PV condition (Δ = -9866px^2). Also, there was an interaction between current response type and preceding response type, $F(1,46) = 61.05$, $p < .001$, $\eta_p^2 = .57$, with a stronger congruency effect after congruent responses (Δ = 32191px^2) than after incongruent responses (Δ = 14864px^2). This interaction held true for all combinations of proportion violation and proportion order (Figure 23, C & D), as indicated by all higher-order interactions including both factors returning non-significant results, *F*s < 2.56, *p*s > .116. None of the remaining effects were significant, *F*s < 1.28, *p*s > .264.

Discussion.

In Experiment 10, my first aim was to replicate the results of Experiment 9. With increased power I had a look at the four-way interaction that again emerged in Rule-task → Rule-task sequences. Now, the frequency manipulation emerged as expected, with lower response costs for violations in blocks with a high frequency of violations. It is difficult to align the results of the frequency manipulation of Experiments

9 and 10, and there is neither reason nor model that helps to reconcile these divergent results, other than that Experiment 10 gives a better estimate with its bigger sample size and returns a result that plays well with what is found in the literature. In both cases, however, violations came with notable costs even if rule violations were more frequent than rule-based responses. And again, recency adaptations only emerged when a high frequency of violations had already been experienced. This strongly suggests that frequency and recency adaptations are not independent mechanisms in the Rule-task, but that recency adaptations only occur if a violation task set is or has been used with sufficient frequency. On the other hand, frequency adaptations do not seem to depend on recency, as frequency manipulations even emerged when no recency adaptations from a preceding rule violation were possible (e.g., visible in the AUC data in Simon-task → Rule-task sequences).

Our next goal for Experiment 10 was to test whether transfer effects from a separate task could modulate the response costs for violations. To do so, I designed a Simon-task that closely resembled the Rule-task. The data is much less complex here: After a violation, responses to incongruent Simon trials are facilitated compared to after a rule-based response (in Rule-task → Simon-task sequences). And even though the Simon-task in principal produced the well-known within task adaptation effects (in Simon-task → Simon-task sequences, as a manipulation check), these adaptation effects do not transfer to the Rule-task (in Simon-task → Rule-task sequences). So the transfer between tasks is asymmetric, such that only violations seem to affect subsequent Simon responses, and not the other way.

One might wonder what exactly transfers between the two tasks. Even though both tasks were designed to share a maximum of features, the cognitive processes that they require differ strongly. Incongruent Simon trials require the translation of a relevant perceptual information (color) into a motor response while shielding this process from the task-irrelevant location of the stimulus. Here, the relevant features have to be activated while simultaneously inhibiting the irrelevant features. In comparison, violations are thought to entail a dual-activation of the rule-based and the violation task set, with an inhibition of the violation task set afterwards to proactively reduce task set competition in the next trial. It might be that the inhibition after a violation can improve the subsequent inhibition of task-irrelevant features in a Simon-task, but the inhibition that is exercised during a Simon-task cannot benefit a rule violation, as producing a violation response does not entail an inhibition. Only after the violation response has been completed, an inhibition process is required, but at that point, a transfer benefit can no longer emerge. Further, the transfer asymmetry could be explained by assuming that rule violations require two processes to allow for a response selection: an inhibition and a modulation (e.g., negation of the original rule), while incongruent Simon trials require only an inhibition process to arrive at the correct response. So after a violation, an inhibition process has already been recruited and performance in a subsequent incongruent Simon trial can improve, but after an incongruent Simon trial, no modulation process is at work, which is required to violate a rule.

However, it might not be the violation task itself that causes the adaptation in the Simon-task, but rather the affect that comes with violating a rule (Chapter 3).

Violations have been shown to entail a negative affective component, and adaptation effects emerge especially in negative settings (van Steenbergen, Band, & Hommel, 2009, 2010; Wirth, Pfister, & Kunde, 2016). This could also explain why, for participants that start with the low-PV condition, there is no benefit after a violation for a subsequent violation (Figure 17, A1), but for a subsequent Simon-task (Figure 19, A1). If not the violation task itself causes the adaptation, but the affective signal that it triggers, this could explain why the Simon-task, which has been demonstrated to respond to mood manipulations, shows an adaptation effect after a violation, but a subsequent violation, that might be more robust towards modulations by affect, does not.

4.4 Preliminary Discussion.

In this chapter, I tested how experimental manipulations that have been shown to reduce the impact of conflicting stimuli, namely the frequency and recency of conflicts, can also modulate the burdens of rule violations. I first started by investigating how negations respond to these manipulations, as negations are believed to be at the core of every rule violation (see Chapter 1). I found that only a combined influence of frequency and recency mitigated the ironic effects of negations. When applied to rule violations, I found the same pattern of results, thereby also replicating Experiments 1 & 3. It seems as if recent exposure to a violation or negation can only be taken into account once a high frequency of violations or negations has been experienced. A high frequency might therefore signal the necessity for adaptation, and recency might provide the mechanism (see the General Discussion for a model).

Finally, I found that the transfer to/from a separate, but closely matched task is asymmetric. While a violation can reduce the impact of subsequent spatial incongruency, spatial incongruency does not reduce the burdens of a subsequent rule violation.

These results are not compatible with the two-step activation model that I initially discussed (Chapter 2). If, as assumed, the task set of violations (and negations) indeed decays right after response execution, then frequency manipulations should not render subsequent violations (and negations) any easier. Further, the model does not assume transfer from or to another task. Therefore, in the General Discussion, I will refine the two-step activation model to account for these new empirical findings.

5. General Discussion.

Rule violations are difficult to plan and execute (Pfister et al., 2016), and in Chapter 2, we found that rule violations seem to represent a special instance of a negation. Further, the planning of a rule violation produced sequential modulations, with repeated violations benefitting from prior violations, the parameters that mirror the execution of the response did not adapt. Based on this result, I introduced a rudimentary two-step activation model, which holds that in order to violate a rule, the rule has to be activated first, and only manipulated in a second step. The original rule has to be held active constantly, causing the ironic effects when violating rules.

In Chapter 3, I then investigated what exactly differentiates rule violations from simple rule reversals/negations. I found that next to an affective component, rule violations additionally trigger an authority-related process that sensitized towards subsequent authority-related stimuli. This additional sensitivity is special to violations and does not occur with negations, and these results cannot be explained by semantic priming.

Finally, in Chapter 4, I followed the idea that the burdens of non-conformity can be overcome (Jusyte et al., in press). Rather than testing special populations, such as convicted criminals, I instead explored whether the difficulty to break rules can be overcome or at least reduced for a given individual as a function of experience.

Instead of further summarizing the results, I want to break presentation conventions and first discuss a revised version of the two-step activation model, for which I also assess convergence with the empirical results presented here. I label this model the "Decision-Implementation-Mandatory switch-Inhibition" (DIMI) model (see Figure 24).

5.1 A modeling approach.

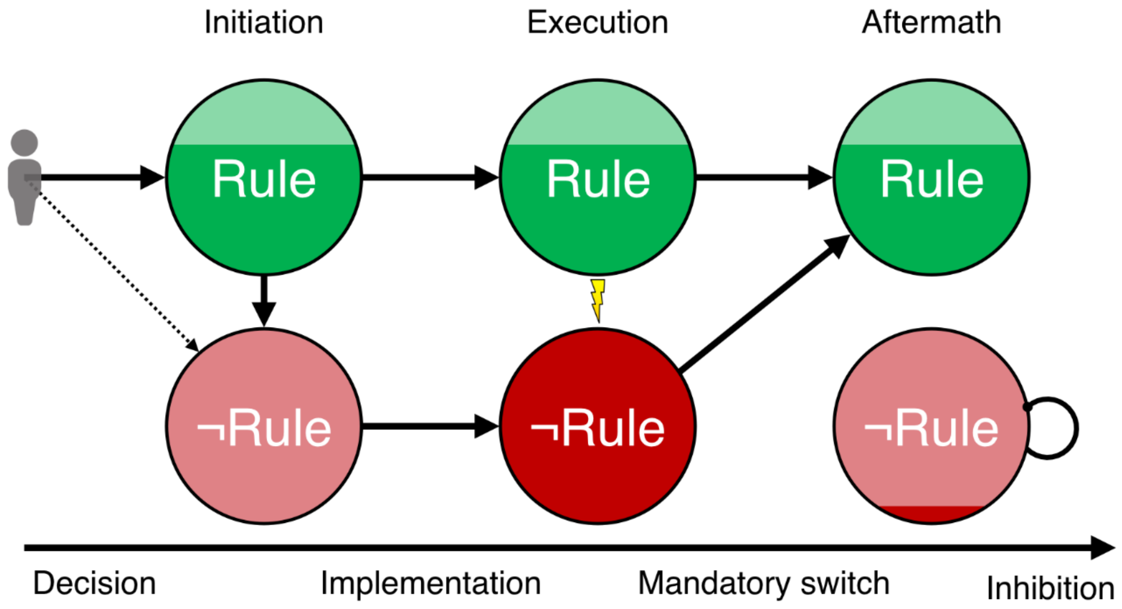

Figure 24. The 'Decision – Implementation – Mandatory Switch – Inhibition' (DIMI) model.

Following the arrows, the model describes the two routes for following or breaking a rule. The circles represent the task sets for following or breaking a rule, the level within the circles depicts the degree of implementation of each task set. First, participants decide whether to follow or break a rule. In this step, either the pre-implemented task set for rule-based responses is chosen, or a new task set for a violation response is created (by means of inversion, negation, or transformation). Next, the chosen task set has to be fully implemented to allow for the selection of a response. While the task set for rule-based responses can stand on its own, the violation task set is only represented as a transformation of the rule-based task set and therefore cannot be active alone. Breaking a rule thereby inherently creates conflict between both task sets. To avoid this conflict in the next trials, one of the task sets has to be inhibited. But as the violation task set cannot be entertained on its own, a mandatory switch back to the rule-based task set is required, and the violation task set is actively inhibited afterwards.

The DIMI model makes the following assumptions:

- Rule-based and violation responses rely on two distinct task sets. The violation task set does not stand on its own, though, but it consists of of the original rule plus a modulator that alienates its meaning (Chapter 2). Consequently, violations require that the original rule is still accessible.
- Choosing one or the other task set typically takes place before response initiation. Implementation, by contrast, is not necessarily completed before response initiation and can continue even during response execution.
- Humans are generally prepared to abide by the rules, so the task set for rule-based behavior can be construed as the default (Asch, 1956; Milgram, 1963, Chapter 3). Therefore, the task set for rule-based responding is partially pre-implemented.
- The simultaneous implementation of two task sets causes interference (Hsieh, Chang, & Meiran, 2012; Meiran, Hsieh, & Dimov, 2010).

Let us first consider the case of rule-consistent behavior. Choosing the pre-implemented rule-based task set is relatively effortless, so responses can be initiated quickly. The implementation of this task set is easier and faster than violating rules, therefore rule-based responses are completed faster. After its use, the strength of implementation levels off to its initial state over time.

Let's now consider rule violations: If one decides to break a rule, the task set for rule violations first has to be created by modulating the original rule-based task set. This modulation can consist of any operator that alienates the meaning of the original task set (in our case, participants probably used a negation), which ultimately creates a new, dependent task set for rule violations. Dependent here means that the task set is represented as a combination of the original task set plus the modulating operator (*with strings attached* to the original rule). This process takes some time, consequently violation responses are initiated comparably slow. This new task set now has to be implemented to allow for response selection, but as it has only just been derived, its implementation takes far longer. Simultaneously, the original rule must be active so that its content can be accessed. This dual implementation during rule violations could explain the persisting influence of the original rule (Pfister et al., 2016). However, implementing two task sets at once is difficult and might lead to interference (Hsieh, et al., 2012; Kuhns, Lien, & Ruthruff, 2007; Meiran et al., 2010), so one of the task sets is best inhibited before the next trial (Koch, Gade, Schuch, & Philipp, 2010). But as the violation task set cannot stand on its own, it is the violation task set is actively inhibited after use, and a mandatory task switch back to the rule-based task set is triggered after violating a rule. This inhibition of the violation task set might also be driven by the negative valence that it is associated with (Chapter 3). Hence, violating the rules cannot become the default (which is assumed to be an evolutionary feature that rewards social behavior; Hoffman, 1981). And as the violation task set is only required rarely, it is inhibited strongly.

This model accounts for the behavioral signature of rule violations within a trial, as well as for sequential effects. For a subsequent rule violation, the decision process (which occurs during ITs) would benefit from a recent violation, as the violation task set would not have to be derived anew. Instead, choosing between following or breaking the rule would follow general task switching logic: repeating the currently active task set (which, after the mandatory switch back, is rule-based responding) would be easier than switching to the currently inhibited task set (represented by the dotted arrow in Figure 24). However, if the violation task set had not been implemented for a longer period of time, it is deallocated to reduce competition between the two task sets. In this case, the decision to follow a rule would again be very fast, and violating a rule would again entail the derivation process, which would be very slow. This modeled pattern of results is actually backed up by empirical data, showing that the decision to follow or break a rule (reflected by ITs) produces sequential adaptation effects, with large costs for violations after a rule-based response, and smaller costs for violations after a violation response.

However, when it comes to the actual execution of the response, the model predicts no repetition benefits for violations: As every trial includes a mandatory switch back to the rule-based task set and an inhibition of the violation task set, choosing the inhibited task set becomes faster, but it has to be implemented anew as if it had not been used before. The implementation process takes longer for violations than for the pre-implemented rule-based task set (reflected by MTs), and the inhibition process afterwards annuls any chance for residual activation of the task set to improve a subsequent violation. Also, the relative degree of implementation of the competing task set allows for predictions of the spatial attraction towards the alternative response (reflected by AUCs). Again, these predictions are reinforced by the empirical data that suggests that neither, the temporal or spatial measures of the response execution are the subject to sequential modulation.

Our current results show that this is only true for participants that start with a low proportion of violations (which probably is the most externally valid scenario, see Experiment 1). The frequency manipulation that I introduced in Experiments 9 and 10 still has to be addressed in the model. To account for this within the model, the following assumptions are added:

- The proportion of violations directly influences the self-inhibition of the violation task set after use, the more it is required, the less it is inhibited.
- Once participants have attenuated the self-inhibition process, the strength of inhibition is fixed, and even with a later low proportion of violations, the self-inhibition is not enlarged (which might reflect a strategic trade-off).

With a high frequency of violations, the violation task set is required more often, and consequently it is inhibited less strongly after use. The inhibition process is attenuated to facilitate a likely subsequent violation that could now benefit from residual activity from the previous trial. This marks a trade-off: Residual activation in the violation task set improves a subsequent violation, but increases the chance of interference between the two task sets. Thus, a subsequent rule-based response should be more difficult. Taken together, this model predicts that with a high proportion of violations, ITs should produce a smaller violation effect, because both task sets are constantly implemented to a certain degree. Also, the execution parameters should now produce smaller violation effects, and even adaptation effects, as repeated violations can benefit from residual activation from the previous trial. This should improve the time of implementation (reflected by shorter times to complete a violation, MTs), and reduce the competing influence of the rule-based task set during violations, allowing for more efficient spatial responses (AUCs).

Crucially, after a violation there is still a mandatory switch back to the rule-based task set. This allows for the odd prediction that even with a high proportion of violations, the infrequent rule-based responses should still be faster and more efficient than frequent violations (again stressing that violations cannot become our default, maybe due to its affective component). And again, all these predictions are met by the empirical data of Experiments 9 and 10. Violation effects diminish with a higher frequency, still violations never become faster or more efficient than rule-based responses. Even for response execution sequential modulations now emerged. However, while this shows that recency adaptations strongly depend on the factor frequency (at least for the response execution), frequency has an effect even when recency cannot be involved: In Experiment 10, when switching from a Simon-task to a Rule-task, recency adaptations could not emerge, but still frequency modulated the results, which can be explained by the attenuated inhibition process in the condition with a high proportion of violations.

What is now left to explain within the model is the transfer effects between the Rule-task and the Simon-task. While the frequency manipulation within the Rule-task only affected the Rule-task and had no influence on the Simon-task, recency adaptations in the Simon-task to previous rule violations emerged. However, these transfer effects were asymmetrical, only after a violation, a response to an incongruent Simon stimulus was facilitated, but after an incongruent Simon trial, no adaptation effects emerged in the Rule-task. This asymmetry can be explained by an affective account that has already been discussed (Chapter 3), but can also be accounted for by the presented model. First, let us summarize the processing steps that are assumed for completing a Simon-task. With stimulus onset, the stimulus' location and color can be processed. The color always indicates the required spatial response, however extracting this information is not automatic. By contrast, extracting the required response from the location is easy and relatively automatic. In a congruent trial, both features hold the same information, so both features are considered in response selection to arrive at a fast decision. In an incongruent trial, this strategy would be detrimental, because the location of the stimulus provokes the commission of an error. Here, the location has to be inhibited to give way for the processing of the stimulus color. In the next trial, this inhibition can be maintained so that response selection in congruent trials is slower, but response selection in incongruent trials is now less affected by the stimulus location. So both, the Simon- and the Rule-task, require the inhibition of information that may cause interference. However, the point in time at which this inhibition is required differs. While the Simon-task requires the incongruent feature to be inhibited prior to response selection, the Rule-task does not allow for the rule-based task set to be inhibited prior to response selection. The violation task set is only represented as a transformation of the original rule and – without the original rule in mind – has no meaning of its own. That might explain why the transfer between the two tasks is asymmetrical. After a rule-violation, there is an inhibition process, which might transfer to the Simon-task and suppresses the irrelevant location of the stimulus. But in an incongruent trial, there is also an inhibition process, but it cannot be transferred to the Rule-task, as the Rule-task only employs an inhibition process after all is set and done. There might be transfer from the Simon-task to the inhibition at the end of a violation, but if this was the case, it cannot be measured by parameters that emerge during a violation.

How can we reduce the burdens that come with non-conformity? This is how to be a rule breaker: Do it often, and then do it repeatedly. Having violated a rule recently only improves the planning of a further violation, the execution is still heavily crippled. And while training alone diminished the response costs for violation, the greatest benefit results from combining both, training and accessibility. Accessibility to a task that presumably resembles the rule violation does not help, however training to break one rule might transfer to the violation of a second rule. With two rules that have to be broken, both tasks require similar operations (modulation and inhibition) that might allow for transfer. These questions still have to be addressed in further research. For now, the best advice to violate rules efficiently is to keep the corresponding task set implemented as strongly as possible, and that can best be done by using it frequently and having it used recently.

5.2 Concluding Remarks.

In the present experiments, I identified specific behavioral markers of rule violations and rule negations. Not only do they come with immediate ironic effects, but they also produce unique sequential modulations. However, I have shown that rule violations are not simply quantitatively different from rule inversions, but that there is a qualitative difference in what cognitive processes they trigger after execution. Finally, I found that the combination of frequency and recency, having violated rules often and shortly before, is the best strategy to mitigate the ironic effects of rule violations, even though they never reach the level of rule-based responses.

Where does that leave us? Now that you have broken the rules and read the whole dissertation, even though the cover warned you not to, you fulfill both criteria of occupying yourself with rule violations frequently during the last minutes, and having just now finished it. Now that you are in the flow, it should be easiest to start anew. But keep an eye out for authorities…

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.

Adriaanse, M. A., Van Oosten, J. M. F., De Ridder, D. T. D., De Wit, J. B. F., & Evers, C. (2011). Planning what not to eat: Ironic effects of implementation intentions negating unhealthy habits. *Personality and Social Psychology Bulletin, 37*(1), 69-81.

Allport, D. A., Styles, E. A., & Hsieh, S. (1994). Shifting intentional set: Exploring the dynamic control of tasks. In C. Umilta & M. Moscovitch (Eds.), *Conscious and nonconscious information processing: Attention and performance XV* (pp. 421–452). Cambridge, MA: MIT Press.

Arrington, C. M., & Logan, G. D. (2004). The cost of a voluntary task switch. *Psychological Science, 15*(9), 610-615.

Arrington, C. M., Weaver, S. M., & Pauker, R. L. (2010). Stimulus-based priming of task choice during voluntary task switching. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(4), 1060-1067.

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied, 70*, 1-70.

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

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review, 108*(3), 624-652.

Braem, S., Abrahamse, E. L., Duthoo, W., and Notebaert, W. (2014). What determines the specificity of conflict adaptation? A review, critical analysis, and proposed synthesis. *Frontiers in Psychology*, 5:1134.

Calderon CB, Verguts T, Gevers W (2015) Losing the boundary: Cognition biases action well after action selection. *Journal of Experimental Psychology: General 144*(4), 737–743

Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology, 3*(3), 472-517.

Clark, H. H., & Chase, W. G. (1974). Perceptual coding strategies in the formation and verification of descriptions. *Memory & Cognition, 2*(1), 101-111.

Csikszentmihalyi, M. (1996). *Flow and the psychology of discovery and invention*. New York: Harper Collins.

Cunningham, C. A., & Egeth, H. E. (2016). Taming the white bear. Initial costs and eventual benefits of distractor inhibition. *Psychological Science, 27*(4), 476-485.

Dale, R., & Duran, N. D. (2011). The cognitive dynamics of negated sentence verification. *Cognitive Science, 35*(5), 983-996.

Darley, J. M., & Latane, B. (1968). Bystander intervention in emergencies: diffusion of responsibility. *Journal of Personality and Social Psychology, 8*(4), 377-383.

de Vega, M., Morera, Y., León, I., Beltrán, D., Casado, P., & Martín-Loeches, M. (2016). Sentential negation might share neurophysiological mechanisms with action inhibition. Evidence from frontal theta rhythm. *Journal of Neuroscience, 36*(22), 6002-6010.

Dovidio, J. F., Piliavin, J. A., Schroeder, D. A., & Penner, L. A. (2006). *The social psychology of prosocial behavior*. Mahwah, NJ: Erlbaum.

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

Dshemuchadse, M., Scherbaum, S., & Goschke, T. (2013). How decisions emerge: Action dynamics in intertemporal decision making. *Journal of Experimental Psychology: General, 142*(1), 93-100.

Duran, N. D., Dale, R., & McNamara, D. S. (2010). The action dynamics of overcoming the truth. *Psychonomic Bulletin & Review, 17*(4), 486-491.

Everett, J. A., Pizarro, D. A., & Crockett, M. J. (2016). Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology: General, 145*(6), 772.

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.

Foerster, A., Wirth, R., Kunde, W., & Pfister, R. (in press). The dishonest mind set in sequence. *Psychological Research*, (ahead of print), 1-22.

Fox, K. J. (1987). Real Punks and Pretenders The Social Organization of a Counterculture. *Journal of Contemporary Ethnography, 16*(3), 344-370.

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.

Funes, M. J., Lupiáñez, J., & Humphreys, G. (2010). Sustained vs. transient cognitive control: Evidence of a behavioral dissociation. *Cognition, 114*(3), 338-347.

Gawronski, B., Deutsch, R., Mbirkou, S., Seibt, B., & Strack, F. (2008). When “just say no” is not enough: Affirmation versus negation training and the reduction of automatic stereotype activation. *Journal of Experimental Social Psychology, 44*(2), 370-377.

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

Gilbert, D. T., Krull, D. S., & Malone, P. S. (1990). Unbelieving the unbelievable: Some problems in the rejection of false information. *Journal of Personality and Social Psychology, 59*(4), 601-613.

Gollwitzer, P. M. & Oettingen, G. (2012). *Goal pursuit*. In R. M. Ryan (Ed.), The Oxford handbook of human motivation (pp. 208-231). New York: Oxford University Press.

Gollwitzer, P. M., & Sheeran, P. (2006). Implementation intentions and goal achievement: A meta-analysis of effects and processes. *Advances in Experimental Social Psychology, 38*, 69-119.

Gratton, G., Coles, M. G., & Donchin, E. (1992). Optimizing the use of information: Strategic control of activation of responses. *Journal of Experimental Psychology: General, 121*(4), 480-506.

Hasson, U., Simmons, J. P., & Todorov, A. (2005). Believe it or not. On the possibility of suspending belief. *Psychological Science, 16*(7), 566-571.

Hoffman, M. L. (1981). Is altruism part of human nature? *Journal of Personality and Social Psychology, 40*(1), 121-137.

Hsieh, S., Chang, C. C., & Meiran, N. (2012). Episodic retrieval and decaying inhibition in the competitor-rule suppression phenomenon. *Acta Psychologica, 141*(3), 316-321.

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.

Kessler, Y., Shencar, Y., & Meiran, N. (2009). Choosing to switch: Spontaneous task switching despite associated behavioral costs. *Acta Psychologica, 131*(2), 120-128.

Kim, D., & Hommel, B. (2015). An event-based account of conformity. *Psychological Science, 26*(4), 484-489.

Koch, I., Gade, M., Schuch, S., & Philipp, A. M. (2010). The role of inhibition in task switching: A review. *Psychonomic Bulletin & Review, 17*(1), 1-14.

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

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

Kool, W., McGuire, J. T., Rosen, Z. B., & Botvinick, M. M. (2010). Decision making and the avoidance of cognitive demand. *Journal of Experimental Psychology: General, 139*(4), 665.

Kuhns, D., Lien, M. C., & Ruthruff, E. (2007). Proactive versus reactive task-set inhibition: Evidence from flanker compatibility effects. *Psychonomic Bulletin & Review, 14*(5), 977-983.

Liefooghe, B., Demanet, J., & Vandierendonck, A. (2010). Persisting activation in voluntary task switching: It all depends on the instructions. *Psychonomic Bulletin & Review, 17*(3), 381-386.

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.

Logan, G. D., & Zbrodoff, N. J. (1979). When it helps to be misled: Facilitative effects of increasing the frequency of conflicting stimuli in a Stroop-like task. *Memory & Cognition, 7*(3), 166-174.

Mayo, R., Schul, Y., & Burnstein, E. (2004). “I am not guilty” vs. “I am innocent”: Successful negation may depend on the schema used for its encoding. *Journal of Experimental Social Psychology, 40*(4), 433-449.

Meiran, N., Hsieh, S., & Dimov, E. (2010). Resolving task rule incongruence during task switching by competitor rule suppression. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(4), 992-1002.

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.

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology, 67*, 371-378.

Monsell, S. (2003). Task switching. Trends in Cognitive Sciences, 7(3), 134-140.

Notebaert, W., & Verguts, T. (2008). Cognitive control acts locally. *Cognition, 106*(2), 1071-1080.

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., Janczyk, M., Wirth, R., Dignath, D., & Kunde, W. (2014). Thinking with portals: Revisiting kinematic cues to intention. *Cognition, 133*(2), 464-473.

Pfister, R., Wirth, R., Schwarz, K. A., 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. *The Quarterly Journal of Experimental Psychology, 29*(4), 727-743.

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

Reason, J. (1995). Understanding adverse events: human factors. *Quality in Health Care, 4*(2), 80-89.

Rogers, R. D., & Monsell, S. (1995). Costs of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology: General, 124*(2), 207-231.

Scherbaum, S., Dshemuchadse, M., Fischer, R., & Goschke, T. (2010). How decisions evolve: The temporal dynamics of action selection. *Cognition, 115*(3), 407-416.

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.

Schroder, H. S., Moran, T. P., Moser, J. S., & Altmann, E. M. (2012). When the rules are reversed: Action-monitoring consequences of reversing stimulus–response mappings. *Cognitive, Affective, & Behavioral Neuroscience, 12*(4), 629-643.

Serota, K. B., & Levine, T. R. (2015). A few prolific liars: Variation in the prevalence of lying. *Journal of Language and Social Psychology, 34*(2), 138-157.

Simon, J. R. (1990). The effects of an irrelevant directional cue on human information processing. *Advances in Psychology, 65*, 31-86.

Spence, S. A., Farrow, T. F. D., Herford, A. E., Wilkinson, I. D., Zheng, Y., & Woodruff, P. W. R. (2001). Behavioural and functional anatomical correlates of deception in humans. *NeuroReport, 12*(13), 2849–2853.

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

Torres-Quesada, M., Funes, M. J., & Lupiáñez, J. (2013). Dissociating proportion congruent and conflict adaptation effects in a Simon–Stroop procedure. *Acta Psychologica, 142*(2), 203-210.

Torres-Quesada, M., Milliken, B., Lupiáñez, J., & Funes, M. J. (2014). Proportion congruent effects in the absence of sequential congruent effects. *Psicologica: International Journal of Methodology and Experimental Psychology, 35*(1), 101-115.

Trigg, G. L. (1979). Grammar. *Physical Review Letters, 42*(12), 747-748.

van Steenbergen, H., Band, G. P., & Hommel, B. (2009). Reward counteracts conflict adaptation evidence for a role of affect in executive control. *Psychological Science, 20*(12), 1473-1477.

van Steenbergen, H., Band, G. P., & Hommel, B. (2010). In the mood for adaptation how affect regulates conflict-driven control. *Psychological Science, 21*(11), 1629-1634.

Vandierendonck, A., Demanet, J., Liefooghe, B., & Verbruggen, F. (2012). A chain-retrieval model for voluntary task switching. *Cognitive Psychology, 65*(2), 241-283.

Wason, P. C. (1959). The processing of positive and negative information. *Quarterly Journal of Experimental Psychology, 11*(2), 92-107.

Wegner, D. M. (2009). How to think, say, or do precisely the worst thing for any occasion. *Science, 325*(5936), 48-50.

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.

Wegner, D. M., Schneider, D. J., Carter, S. R., & White, T. L. (1987). Paradoxical effects of thought suppression. *Journal of Personality and Social Psychology, 53*(1), 5-13.

Wirth, R., Foerster, A., Rendel, H., Kunde, W., & Pfister, R. (in press). Rule-violations sensitise towards negative and authority-related stimuli. *Cognition & Emotion*, (ahead of print), 1-14.

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

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.

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.

Yinger, J. M. (1982). *Countercultures*. New York, NY: Free Press.

List of Figures.