Searching for truth in dishonesty:
The cognitive architecture of lying

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

Vorgelegt von Anna Förster
aus Würzburg

Würzburg, 2019

Erstgutachter (und Erstbetreuer): Prof. Dr. Wilfried Kunde

Zweitgutachter (und Zweitbetreuer): Prof. Dr. Matthias Gamer

Tag des Kolloquiums: 07.05.2020

It all started with an introduction to empirical and experimental research methods in a yellow building. Thank you for taking me on the ride, Roland and Wilfried.

× Zusammenfassung

Menschen handeln und interagieren in der Regel entsprechend dem was sie als wahr erachten. Allerdings muss diese Präferenz hin und wieder unehrlichen Handlungen weichen. Die dafür notwendige Überwindung initial ehrlicher Antworttendenzen erweist sich als kognitiv aufwendig. Diese Thesis ergründet in drei Experimentalserien die Eigenschaften dieses Wettstreits ehrlicher und unehrlicher Antworttendenzen für offen ausgeführte Lügen. Damit reihen sie sich in jüngste Bestrebungen ein, Lügen nicht nur oberflächlich als schwierigere der beiden Handlungen zu beschreiben, sondern zu einer präzisen Charakterisierung der beteiligten kognitiven Prozesse zu gelangen. Die Forschungsfragen und das methodische Vorgehen dieser Thesis basieren dafür auf der gemeinsamen Betrachtung kognitiver Theorien, empirischer Evidenz und Paradigmen aus der Forschung zum Lügen, zur kognitiven Kontrolle und zu sensomotorischen Stadienmodellen der Informationsverarbeitung.

Die Experimente lokalisieren den dem Lügen inhärenten Handlungskonflikt in der zentralen, kapazitätslimitierten Phase der Informationsverarbeitung (Experimente 1 bis 4), zeigen jedoch auch, dass dieser Konflikt sowohl durch kognitive Kontrollmechanismen (Experimente 5 bis 7) als auch durch das Verinnerlichen falscher Alibis (Experimente 8 bis 11) reduziert bzw. vollständig eliminiert werden kann. Die Daten offenbaren eine starke Flexibilität in der kognitiven Verarbeitung unehrlicher Handlungen: Einerseits scheint die Ausführung einer Lüge und die Überwindung wahrheitsgemäßer Handlungstendenzen besonders auf kapazitätslimitierte Selektionsprozesse zurückzugreifen, begleitet von vor- und nachgelagerten Aktivierungs- und Überwachungsprozessen. Andererseits können kognitive Kontrollmechanismen und falsche Alibis diese aufwändigen Prozesse entscheidend eindämmen. Diese Ergebnisse untermauern und erweitern bestehende Theorien zu den kognitiven Grundlagen des Lügens. Für angewandte Vorhaben im Bereich der Lügendetektion ist die beobachtete Flexibilität der kognitiven Verarbeitung angesichts falscher Alibis alarmierend. Ein vielversprechender Ansatz zur Weiterentwicklung in diesem Bereich wäre eine genaue Unterscheidung von Prozessen der Aktivierung, des passiven Zerfalls und der aktiven Inhibition wahrheitsgemäßer Repräsentationen beim Lügen und eine Bewertung der Anpassungsfähigkeit dieser Prozesse.

× Summary

Honest actions predominate human behavior. From time to time, this general preference must yield to dishonest actions, which require an effortful process of overcoming initial honest response activation. This thesis presents three experimental series to elucidate this tug-of-war between honest and dishonest response tendencies in overtly committed instances of lies, thereby joining recent efforts to move from a sheer phenomenological perspective on dishonest responding as being more difficult than honest responding to a precise description of the underlying cognitive processes. The consideration of cognitive theories, empirical evidence, and paradigms from different research fields – dishonesty, cognitive control and sensorimotor stage models of information processing – lay the groundwork for the research questions and methodological approach of this thesis.

The experiments pinpoint the underlying conflict of dishonest responding in the central, capacity-limited stage of information processing (Experiments 1 to 4), but they also demonstrate that cognitive control processes (Experiments 5 to 7) and the internalization of false alibis (Experiments 8 to 11) can reduce or even completely eliminate this conflict. The data reveals great flexibility at the cognitive basis of dishonest responding: On the one hand, dishonest responding appears to rely heavily on capacity-limited processes of response selection to overcome honest response tendencies alongside up- and downstream consequences of response activation and monitoring. On the other hand, agents have powerful tools to mitigate these effortful processes through control adaptation and false alibis. These results support and expand current theorizing of the cognitive underpinnings of dishonest responding. Furthermore, they are alerting from an applied perspective on the detection of lies, especially when considering the flexibility of even basic cognitive processes in the face of false alibis. A promising way to move forward from here would be a fine-grained discrimination of response activation, passive decay and active inhibition of honest representations in dishonest responding and the assessment of the adaptiveness of these processes.

Honesty and dishonesty up close

1 Facets of lies

I start with a harmless truth about me: I despise raw tomatoes but love them processed. In general, I would share this information without hesitation, but in the following specific situation, I might make an exception. I would like to buy tomatoes on the market to prepare a sauce for pasta. The farmer is happy about my choice of vegetables and claims that the quality of these tomatoes is outstanding, in fact, she is sure that these are the best tomatoes she has ever harvested. Therefore, she encourages me to enjoy them in the only way that she deems appropriate: on its own and neither cooked nor seasoned. I might counter this with: “Thank you for the advice, I will do that”.

We can agree from face value that this would be an outright lie. But what exactly constitutes a lie? One way to approach this question is to consult philosophical analyses of the concept whereas a second way to answer it is to collect subjective judgments from lay people. From the latter perspective my initial example can be construed as a prototypical lie as it contains information that the liar, in this case me, believes to be false and this information is communicated with the intention to mislead someone (Coleman & Kay, 1981; see also Bloomquist, 2010 for a similar perspective on cheating and stealing). However, people classify actions still as lies if these actions do not comprise these prototypical elements. Moreover, they do not provide dichotomous decisions whether actions are or are not lies but give more fine-grained gradual assessments of the extent of (dis)honesty[1]. In the following, I will give a brief introduction into the peculiarities of lies along these prototypical elements.

1.1 Intending false belief

It is not a unique consequence of lies that false information appears as being true, but this can also happen, amongst other reasons, by mistake. For example, I could misunderstand the farmer, thinking that she encouraged me to enjoy the tomatoes well cooked and seasoned. I would happily agree with this advice and this erroneous communication would leave the farmer with a false belief about how I consume the

[1] I use honesty and truth-telling as well as dishonesty and lying interchangeably throughout this thesis.

tomatoes. So what distinguishes dishonest responding from errors, acting, and ironic or sarcastic jokes, and makes it similar to cheating and stealing? It often comes with an intention to mislead someone else by providing information that the liar himself deems to be misleading (Bloomquist, 2010; Coleman & Kay, 1981).

Although deceptive intentions are present in most instances of lies, they can be absent in some cases, e.g., when it is obvious for every conversational partner that a lie is being told or when a lie is stated in private. People classify false statements in such contexts still as lies (Arico & Fallis, 2013; Coleman & Kay, 1981) even if they acknowledge explicitly that misleading intentions are absent or weak (e.g., Rutschmann & Wiegmann, 2017; J. Turri & Turri, 2016). A person might also want to provide misleading information but it turns out to be true. Applied to the tomato example: Is it still a lie if I later eat a piece of tomato raw and unseasoned confusing it with bell pepper? The debate about this issue is still ongoing, however instances, where people want to provide false information but deliver actually true information are judged as being less of a lie than instances where information is actually false (Coleman & Kay, 1981; A. Turri & Turri, 2015, 2019; Wiegmann, Samland, & Waldmann, 2016). Relatedly, people judged truthful statements that are provided in a misleading way to be lies (Rogers, Zeckhauser, Gino, Norton, & Schweitzer, 2017). Keeping with the initial example of the conversation between me and the farmer, this could translate to such a response: “Thank you for the tip!” The goal of this statement would still be to mislead the farmer about my preference for raw tomatoes, in this case, without communicating false information.

1.2 Means of inducing a false belief

The former example also points to another characteristic of lies. Namely, whether the liar delivers false information about the inquired content or omits information on that matter. The initial example about eating the raw tomatoes is a lie in the form of an explicit, false statement contradicting the truth entirely. However, the described situation would not even require such an explicit statement of falsity. The farmer would probably assume anyway that I would follow her advice after I already bought tomatoes and most people like them raw. As such, “Thank you for the tip!”, from my side would omit any explicit information on the consumption of raw tomatoes but I would still attempt to mislead the farmer to believe that I would comply with her advice (cf., Ekman, 1997, Coleman & Kay, 1981).

Although the mentioned differences between these instances of lying are quite subtle and bring about the same consequences for the target's belief, initial evidence points to a different assessment of lies by omission and lies by commission (see E. Levine et al., 2018 for a comparison of these types of lies in a prosocial context). Giving participants the opportunity to lie in the lab by either withholding true information that would have corrected falsity or to directly deliver false information showed that they lie more frequently by omitting rather than delivering information (Pittarello, Rubaltelli, & Motro, 2016). Participants observed a virtual coin flip and knew that they would gain a reward if one side turned up whereas the other side would leave them without a reward. In fact, the unrewarded side of the coin turned up for everyone. One half of the participants then received correct feedback that they did not gain a reward, the other half received incorrect feedback that they gained a reward. Both groups had to choose whether this feedback was correct or incorrect in a drop-down menu. As such, lies of both groups entailed a motor action, which might be construed as lying by commission. However, the group that received correct feedback had to claim that the feedback was incorrect whereas the group that received incorrect feedback merely confirmed a false premise, which encourages a distinction of lying into commissions and omissions for the two instances. Participants lied less by means of committing rather than omitting information in this study. Such a pattern of results also emerged when commissions and omissions dissociated more strongly in regard to the execution of motor actions (Teper & Inzlicht, 2011). Participants were more prone to cheat on a math test by looking into the results before they arrived at a solution by themselves when the cheating opportunity afforded no action in contrast to when it did.

These studies demonstrate that there is not one definite concept of a lie for lay people but that certain characteristics determine the level to which actions are judged as being lies. The experiments of the current thesis feature lies by commission only that agree with two of the characteristics: participants provided a motor response and they knew that this response was false. Relying on one instance of lying throughout all experiments rather than introducing different concepts was a feasible approach for clear-cut tests of the current research questions. For the same reason, I manipulated the intention of actions meticulously, instructing participants to respond honestly and dishonestly and giving accuracy feedback for their actions. This approach is well in line with a large body of research on the cognitive underpinnings of dishonest responding (e.g., Furedy, Davis, &

Gurevich, 1988; Suchotzki, Verschuere, Van Bockstaele, Ben-Shakhar, & Crombez, 2017).

2 The truth will out

Diversity does not only seem to be a property of the definition of lies but also of the demeanor that accompanies dishonest responding. The bad news is that people in general and even persons who must deal with deception frequently (e.g., judges), struggle to distinguish lies from truthful statements any better than chance (e.g., Bond & DePaulo, 2008; Ekman & O'Sullivan, 1991). The good news is that the truth predominates human behavior; or: *The truth will out.* Two central perspectives motivate this claim that I will elaborate on in the following.

2.1 Truth trumps lie

People report to tell lies about one to two times a day (e.g., Debey, De Schryver, Logan, Suchotzki, & Verschuere, 2015; DePaulo, Kashy, Kirkendol, Wyer, & Epstein, 1996; Halevy, Shalvi, & Verschuere, 2014; Serota, Levine, & Boster, 2010). Cheating tasks in the laboratory corroborate these self-reports, showing that people take opportunities to cheat to some extent, but not exhaustively, with large interindividual differences (e.g., Fischbacher & Föllmi-Heusi, 2013; Gerlach, Teodorescu, & Hertwig, 2019; Halevy et al., 2014; Hilbig & Thielmann, 2017; Lohse, Simon, & Konrad, 2018; Pfister, Wirth, Weller, Foerster, & Schwarz, 2018; Serota et al., 2010; Tabatabaeian, Dale, & Duran, 2015). The *die under cup* paradigm is one prominent approach to assess unsolicited and anonymous lying experimentally. In this task, participants roll a die secretly and only they know the outcome of their rolls. Crucially, they earn money based on their self-reported outcome (e.g., higher earnings for higher reports), which encourages them to report deceptive outcomes. Lying cannot be attributed to an individual participant in this case, but can be assessed on a group level by analyzing whether the reported die rolls differ from the expected value. People lie in this task, but not all necessarily lie in a way that would maximize their benefit (Fischbacher & Föllmi-Heusi, 2013). This was evident because not only the die roll with the highest-earning outcome was reported more frequently than chance level, but also the one with the second-highest outcome. Furthermore, some participants still reported low outcomes, indicating that they stuck to the truth. A considerable fraction of people seem to prefer honesty even when lying

promises high incentives while others are ready to lie even for small benefits (Hilbig & Thielmann, 2017). The latter study suggested a classification of participants into groups of brazen liars who lied consistently independent of the size of incentives, corruptible participants who lied for the prospect of high incentives and incorruptible participants who were mostly honest (cf., Fischbacher & Föllmi-Heusi, 2013; Hilbig & Zettler, 2015; Pfister et al., 2018).

Despite these interindividual differences, truthful behavior clearly prevails on average. Nevertheless, a lively debate emerged in recent years about whether people are indeed initially biased to tell the truth or rather lie by default when tempted to cheat in an anonymous setting as the *die under cup* paradigm (e.g., Bereby-Meyer & Shalvi, 2015; Capraro, Schulz, & Rand, 2019; Foerster, Pfister, Schmidts, Dignath, & Kunde, 2013; Shalvi, Eldar, & Bereby-Meyer, 2012, 2013; Tabatabaeian et al., 2015; Verschuere & Shalvi, 2014; see also Rand, Greene, & Nowak, 2012 for related findings on cooperative actions). A clever procedural tweak in a recent study allowed researchers to leave participants in the dark about the opportunity to cheat until they actually had to respond (dis)honestly (Lohse et al., 2018). Participants randomly received a low or high income in a lottery and to their surprise, had to self-report this outcome. As such, participants with a low outcome could benefit from falsely claiming the high income. These participants, however, preferred to respond honestly than participants who responded dishonestly. Both, dishonesty and awareness of the cheating opportunity decreased further when participants responded under external time pressure. The authors drew the intriguing conclusion that people are intuitively honest because they do not easily recognize cheating opportunities.

Truthful actions do not only trump lies in numbers but also in processing efficiency. While the formerly described methodology suffices to assess the prevalence of lies mostly on an aggregated group level, it is not precise enough to pinpoint the efficiency of lying or the exact cognitive processes that are involved in generating a lie (for this argument in the context of rule violations, see for example Wirth, Foerster, Rendel, Kunde, & Pfister, 2018). This is however possible in paradigms that I term *instructed intention paradigms*[2],

[2] These are also known as *Differentiation-of-Deception paradigm* (Furedy, Davis, & Gurevich, 1988) or *Sheffield lie test* (Verschuere, Spruyt, Meijer, & Otgaar, 2011, of which the latter name appears to be a reference to the affiliation of the researchers who introduced the described method (Spence et al., 2001).

thereby emphasizing the experimental manipulation of honest and dishonest responding. In most cases, (dis)honest responding varies randomly on a trial-by-trial basis of repeated measurements within participants. Participants deliver speeded (dis)honest responses on a computer to questions about general knowledge (e.g. “Does the Main [Elbe] river run through Würzburg?”; honest response: yes [no]), about autobiographical information (e.g., “Are you a PhD student [professor]?”; honest response in my case: yes [no]) or about experimentally controlled activities that were conducted in the laboratory. The consequences of correct honest and dishonest responses are the same and mostly neutral for both intentions. As such, differences between both intentions can be attributed to cognitive processes underlying the generation of appropriate responses rather than to the expectancy of different consequences. For example, if I would have to respond to “Are you a Ph.D. student?”, my honest response would be *yes* and my dishonest response would be *no*. If I was asked “Are you a professor?”, I would honestly respond with *no* and dishonestly with *yes.* In these paradigms, behavioral, electrophysiological and hemodynamical data points to increased cognitive effort when lying than when telling the truth (e.g., Bhatt et al., 2009; Christ, Van Essen, Watson, Brubaker, & McDermott, 2009; Debey, Verschuere, & Crombez, 2012; Foerster, Wirth, Kunde, & Pfister, 2017; Johnson, Barnhardt, & Zhu, 2003, 2004; Pfister, Foerster, & Kunde, 2014; Spence et al., 2001; Suchotzki et al., 2017; Suchotzki, Crombez, Smulders, Meijer, & Verschuere, 2015; Walczyk, Roper, Seemann, & Humphrey, 2003). These results strongly suggest that the truth does not only predominate decision-making but that it also interferes with the production of a dishonest response even after a decision has been made.

2.2 From truth to lie

The persisting influence of truthful responses during the generation of a lie is a central aspect of the *activation-decision-construction-action theory* (Walczyk, Harris, Duck, & Mulay, 2014; for a former version of the theory, see Walczyk et al., 2003), which offers a comprehensive theoretical assessment of the cognitive architecture of dishonest responding. The *activation* component assumes that most social interactions feature implicit or explicit triggers to represent relevant truthful content in working memory. This process is assumed to be mostly automatic, but it can also be effortful if access to the relevant information in long-term memory is difficult or is even not possible. The *activation* component is the starting point of the theory whereas the three remaining components

explain how agents proceed from their initial tendency to respond truthfully and arrive at a particular dishonest response. The *decision* component proposes that especially the expectancy of negative consequences of being honest gives rise to a comparison of these consequences to the potential consequences of being dishonest. Lower expected values of the outcome of honest actions or higher expected values of the consequences of being dishonest go along with more (hypothetical) dishonest actions (Cassidy, Wyman, Talwar, & Akehurst, 2019; Masip, Blandón-Gitlin, La Riva, & Herrero, 2016; Walczyk, Tcholakian, Newman, & Duck, 2016). The theory assumes that the difference between the value of honest and dishonest actions determines the strength of the motivation to lie. The *construction* component sets the generation of a specific dishonest response into the context of the honest response, the goals of the liar and the social context. The last component concerns the *action* itself and incorporates the actual execution of a dishonest response assuming a parallel inhibition of the activated honest response. It also states that liars monitor and control their demeanor while they also watch the reaction of the target of their lie closely.

This is not an exhaustive presentation of all the aspects of the theory (which can be studied in more detail in Walczyk et al., 2014); this brief introduction nevertheless reveals a crucial aspect of the cognitive underpinnings of lying, namely that honest response activation allegedly takes a prevailing role in this process up to response execution. *Instructed intention paradigms* arguably tap into the process of overcoming the truth during dishonest responding because cues to respond (dis)honestly and simple *yes*/*no* responding render elaborate decision-making or construction of a response obsolete. Furthermore, participants typically respond to questions on a computer with neutral consequences for both, honest and dishonest responses, which does not require monitoring a human target or controlling their own demeanor. Two studies deliver probably the most direct evidence for truth activation during lying by introducing small methodological changes to the *instructed intention paradigm*. For one, Duran, Dale, and McNamara (2010) implemented continuous instead of discrete responding to questions by asking participants to move the cursor of a game console between a starting area and two response areas (*yes* vs. *no*). This allowed them to assess the effects of (dis)honesty from the onset of a question to the end of response execution via a broad range of dependent variables. Crucially, the prevailing role of honest response activation became most obvious for the trajectory of the movement from the start to the end area. For

dishonest responses, trajectories were biased toward the honest response option whereas, for honest responses, trajectories followed a more direct path (for a similar approach to rule violations see, for example, Pfister, 2013). Second, Debey, De Houwer, and Verschuere (2014) added honest and dishonest distractors to an *instructed intention paradigm*. As such, participants were not only confronted with questions and cues to respond honestly or dishonestly, but they also saw either *yes* or *no* as irrelevant distractors above and below each question. They replicated the usual finding of slower and more inaccurate dishonest than honest responses (Suchotzki et al., 2017). Importantly, distractors that corresponded with the honest response helped honest responding compared to distractors that corresponded with the dishonest response. Crucially, distractors corresponding with honest responding also helped dishonest responding, which supports the assumption that honest distractors facilitated initial honest response activation in dishonest responding.

Both studies deliver strong support for the assumption of the *activation-decision-construction-action theory* that the honest response takes a prevailing role when responding dishonestly. The studies presented in Chapter II of this thesis built on this theoretical and empirical foundation and aimed at a fine-grained dissociation of information processing involved in honest versus dishonest responding. By combining two prominent methodologies from cognitive psychology, the *instructed intention paradigm* and the *psychological refractory paradigm*[3], Experiments 1 to 4 collected a comprehensive overview of the impact of the tug-of-war between honest and dishonest response tendencies on information processing. These examinations did not only look at the role of the conflict during the preparation of the response but also targeted downstream consequences on monitoring processes that operate after a dishonest response has already been delivered. Thereby, the presented experiments strongly zoom in on dishonest actions, dissecting their basic cognitive architecture.

2.3 Under control

The experiments of Chapter II make a strong case that honest response activation is a key element of dishonest responding, following prominent theories such as the *activation-decision-construction-action theory* (e.g., Walczyk et al., 2014). Chapter III

[3] Chapter II provides a detailed description of this method.

accommodates experiments that demonstrate that truth activation is, however, not necessarily a by-product of lying but that agents can flexibly gear cognitive processing toward being dishonest. Assuming that people for one, have a strong preference for being honest and second, are mostly tuned to retrieve an appropriate honest response immediately, dishonest responding finds itself in conflict with this overarching goal and its associated representations of appropriate actions. This parallel activation of competing response tendencies qualifies as a behavioral conflict, and behavioral conflicts can be detected and can then trigger the adaptation of cognitive control settings to promote successful goal-based responding (e.g., Botvinick, Braver, Barch, Carter, & Cohen, 2001; Braem et al., 2019).

A considerable amount of empirical work agrees with the notion that agents can contain or even abolish behavioral conflict during dishonest responding by adapting cognitive control settings after experiencing recent or frequent instances of dishonest responding, by implementing response strategies through instructions or by learning particular dishonest responses through practice (e.g., Debey, Liefooghe, De Houwer, & Verschuere, 2015; Foerster, Wirth, Kunde et al., 2017; Güldenpenning, Alaboud, Kunde, & Weigelt, 2018; Hu, Chen, & Fu, 2012; Van Bockstaele et al., 2012; Van Bockstaele, Wilhelm, Meijer, Debey, & Verschuere, 2015; Verschuere et al., 2011). In the framework of the *activation-decision-construction-action theory* (Walczyk et al., 2014), this might translate to more efficient processing or even an entire skipping of one or more of the components after increasing cognitive control or after rehearsing dishonest responses. The experimental synopsis of Chapter III targets the adaptiveness of dishonest responding from multiple angles. It disentangles the impact of control adaptation through recent and frequent dishonest responses (Experiment 5), explores common currencies of adaptation to conflict in dishonest responding and other behavioral conflicts (Experiment 6 and 7) and scrutinizes the role of false alibis in this process (Experiments 8 to 11).[4] The fourth chapter of this thesis integrates these lines of research, delivering a concise overview of how the findings of this thesis advance our current understanding of the cognitive architecture of lying and providing promising avenues to proceed with this endeavor in the future.

[4] Note that I chose a consecutive numbering of Experiments 5 to 11 in this thesis, deviating from the numbering in the published articles.

Empirical synopsis: From truth to lie

3 Capacity limitations of dishonesty

Cognitive theories of dishonesty revolve around an automatic activation of honest response tendencies, which is assumed to impair response selection for the intended dishonest response. Clear-cut evidence for the claim is still limited, however. We, therefore, present a novel approach to dishonest responding that takes advantage of psychological refractory period methodology. Four experiments yielded evidence supporting the assumption of prolonged response selection during dishonest responding. Moreover, they also showed differences in early response activation and they revealed additional downstream consequences of this behavior that are currently not sufficiently covered by common theoretical models. Notably, these downstream consequences included increased monitoring relative to honest behavior. Our results thus provide extensive coverage of the cognitive architecture of dishonest responses, informing current theorizing while simultaneously grounding the assumed processes in the framework of sensorimotor stage models of information processing.

3.1 Introduction

Behaving dishonestly requires complex cognitive and emotional processing of agents before, during, and even after delivering a lie (e.g., Walczyk et al., 2014). On the cognitive level, the generation of dishonest responses has often been suggested to require a sequence of an initial activation and subsequent inhibition of the appropriate honest response (e.g., Debey et al., 2014; Foerster, Wirth, Herbort, Kunde, & Pfister, 2017). When lying affords such an inhibition of a dominant response, it is considerably more difficult than honest responding, which is reflected in behavioral, electrophysiological, and hemodynamical measures (e.g., Bhatt et al., 2009; Debey, Liefooghe et al., 2015; Johnson et al., 2003; Pfister et al., 2014; Spence et al., 2001; Suchotzki et al., 2017). The presence of this two-step process in dishonest responding is well documented in the literature, however, an exact characterization in regard to the stages of information processing that are prolonged during lying still awaits examination. In the present study, we approached the cognitive consequences of dishonesty before, during, and also after delivering a lie systematically from the perspective of the *psychological refractory period* (PRP) *paradigm* (Pashler & Johnston, 1989; Welford, 1952). In what follows, we will first review the current theoretical frameworks of how lies are processed, and we will then move on by discussing how these processes can be mapped to processing stages via PRP methodology.

3.1.1. Honest response activation in dishonest responding

The *activation-decision-construction-action theory* (Walczyk et al., 2003, 2014) brings together cognitive and emotional processes underlying dishonest processing. In particular, the theory states that in many cases, an honest response is automatically activated and that agents decide whether to lie in the face of this response activation. Decisions whether or not to lie are based on factors such as the present social context, expected consequences, and the agent's experiences. In case of a decision to lie, agents then need to inhibit the representation of the honest response to replace it with a suitable dishonest response. Finally, the theory also assumes that agents monitor and control their demeanor and that they monitor the behavior of the receiver of the deceptive message.

The assumption of an initial activation of an honest response representation and its inhibition is typically examined in *instructed intention paradigms*, where participants are prompted to respond honestly and dishonestly with *yes* or *no* to autobiographical or semantic questions. These paradigms reliably produce strong intention effects, as

participants are slower and less accurate when delivering dishonest compared to honest responses (e.g., Duran et al., 2010; Furedy et al., 1988; Spence et al., 2001). Recent studies began to investigate the cognitive foundations of such intention effects by using a modified version of the *instructed intention paradigm* that featured honest and dishonest distractors (Debey et al., 2014; Foerster, Wirth, Herbort et al., 2017, Experiments 3-4). Distractors (*yes* or *no*) appeared simultaneously with the question. If the honest response to a question is *yes*, the same distractors would constitute honest distractors, whereas *no* distractors would constitute dishonest distractors. The opposite is true for questions with an honest *no* response. Assuming that the honest response is initially activated, the presentation of honest distractors should complement this initial response activation and, thus, facilitate honest responding. Because the initial activation of the honest response is also assumed to occur during dishonesty, honest (rather than dishonest) distractors should also expedite the processing of dishonest responses. This unique prediction of the two-step hypothesis was indeed confirmed, with lower RTs and error rates with honest than with dishonest distractors when responding honestly and, crucially, also when responding dishonestly.

Findings in the instructed intention paradigm thus indicate that selecting, planning, and initiating a dishonest response can occur in the face of the activated truthful response. This describes the processing of dishonest responses as being inherently conflicting, effortful, and resource demanding, and such processes are commonly located within a certain stage of information processing, that is, the central bottleneck of response selection (Ferreira & Pashler, 2002; Paelecke & Kunde, 2007; Wirth, Pfister, Janczyk, & Kunde, 2015). However, additional findings have also suggested a profound impact of dishonest processing on response execution, which becomes evident in continuous movement trajectories when participants respond by moving their hands or a cursor toward a *yes* or *no* response location (Duran et al., 2010; Foerster, Wirth, Herbort et al., 2017). Such movements are slower and more curved toward locations that signal dishonest rather than honest responding, which may be taken to suggest that dishonest processing also affects processes after a response has already been selected (for related findings on rule-breaking, see Pfister, Wirth, Schwarz, Steinhauser, & Kunde, 2016; Wirth, Pfister, Foerster, Huestegge, & Kunde, 2016). Scrutinizing these speculations requires advanced experimental setups as we will describe in the following.

3.1.2. Localizing the two-step process

Sensorimotor approaches mostly assume that information processing can be described as stages of (mainly perceptual) precentral processing, a central judgments that is concerned, among other things, with response selection, and postcentral, motoric processing (e.g., McClelland, 1979; Smith, 1968; Sternberg, 1969). Vast empirical evidence supports the assumption that the central process is capacity-limited and cannot run at all, or not with the same efficiency, in two tasks at a time, whereas processes before and after this central process can mostly run in parallel with all other stages of another task (e.g., Massaro & Cowan, 1993; Meyer & Kieras, 1997; Pashler, 1984, 1994a; Pashler & Johnston, 1989). Most part of the two-step process should draw upon this response selection stage, rendering the inhibition of an honest response and the generation of a dishonest response a central, capacity-limited operation.

However, as outlined above, there is evidence suggesting a unique signature of dishonest responding also after a response has been selected (Duran et al., 2010; Foerster, Wirth, Herbort et al., 2017). Speculatively, these findings might indicate the operation of a late capacity-limited process that monitors responses and their consequences (Jentzsch, Leuthold, & Ulrich, 2007; Welford, 1952). Such a monitoring process seems to be especially engaged when response selection or execution creates conflicts. This happens when producing errors (i.e., conflict between erroneous and correct response; Jentzsch & Dudschig, 2009; Steinhauser, Ernst, & Ibald, 2017) or incompatible response effects (i.e., when a left response had produced a right stimulus; Wirth et al., 2015; Wirth, Janczyk, & Kunde, 2018; Wirth, Steinhauser, Janczyk, Steinhauser, & Kunde, 2018). To the extent that dishonest responding comes with a conflict between honest and dishonest representations, it might invoke such monitoring as well (Foerster, Pfister et al., 2018).

The PRP paradigm provides an established tool to disentangle the involvement of dishonest processes in the stages of information processing (e.g., Pashler, 1984, 1994a; Pashler & Johnston, 1989). In this paradigm, participants work on two tasks in close temporal succession (see Figure 1). The temporal proximity of the two tasks varies via the manipulation of the stimulus onset asynchrony (SOA) of the task stimuli. According to the model, the temporal overlap between the two tasks should not affect the performance of the first task. The performance of the second task, however, should worsen with an

increasing temporal overlap of the tasks; this impact of the SOA on RTs and error rates is referred to as PRP effect. From the introduction of an experimental manipulation of interest, separately in Task 1 or Task 2, and from its impact on RTs of both tasks, experimenters draw inferences about the localization of these effects in sensorimotor stages.

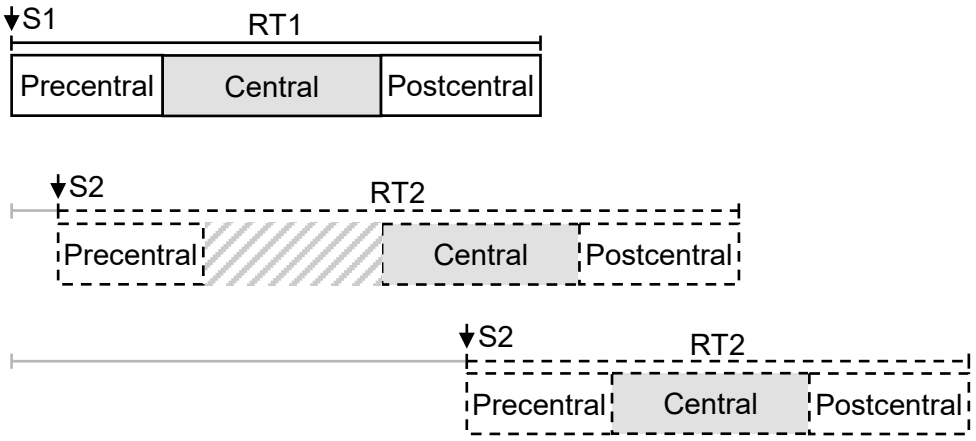

Figure 1 | Illustration of the processing stages of a Task 1 (solid black lines) and a Task 2 (dashed black lines) in the *psychological refractory period paradigm*. The stimulus of Task 1 (S1) and the stimulus of Task 2 (S2) appear with short (middle row) or long (bottom row) stimulus onset asynchrony (SOA; gray lines). Central stages of response selection are assumed to be capacity-limited and therefore unable to operate in parallel, leading to a cognitive slack (shaded gray area) when response selection of Task 2 has to wait for response selection of Task 1 to finish. This inevitably prolongs response times of Task 2 (RT2; dashed lines) to a larger degree with short SOAs than with long SOAs, whereas RTs of Task 1 (RT1; solid lines) are mostly unaffected by manipulations of SOA. As such, experimental manipulations of precentral and central stages of Task 1 should also become visible in the performance of Task 2 for relatively short SOAs (*effect propagation*). Manipulations in the precentral stage of Task 2 should affect RT2 to a larger degree at long SOAs than at short SOAs, as longer precentral processing can stretch into the cognitive slack in the latter case, whereas manipulations of central and postcentral stages in Task 2 should affect RT2 to the same degree at all SOA levels (*locus-of-slack logic*).

3.1.3. The present experiments

The current experiments offer a comprehensive inspection of cognitive effects of dishonesty in the stages of information processing by combining established methods from theories on dishonesty and from sensorimotor approaches to information processing. Therefore, the current experiments featured a (dis)honest task in combination with a tone classification task. The order of the two tasks varied between experiments, with the *locus-of-slack logic* (Experiment 1 and 2) employing the tone task first and the (dis)honest task

second, and with the *effect propagation logic* (Experiment 3 and 4) employing the reversed order of tasks (for detailed descriptions of both methodological approaches, see Jentzsch & Dudschig, 2009; Kunde, Pfister, & Janczyk, 2012; Miller & Reynolds, 2003).

3.2 Experiment 1

The first experiment used the *locus-of slack logic* to elaborate on whether dishonest responding relates to precentral or later stages of information processing. Previous work on the cognitive basis of effects of responding dishonestly suggests that these effects should mostly draw upon the later stages, that is, response selection, motor execution, and/or monitoring rather than on the precentral stage (e.g., Debey et al., 2014; Duran et al., 2010; Walczyk et al., 2014). From the background of sensorimotor theories, however, there is evidence that suggests a contribution of precentral response activation processes (e.g., Hommel, 1998a; Miller, 2006) that may also be affected in dishonest responding.

In particular, these studies suggested the existence of early response activation processes by showing that response characteristics of a Task 2 can facilitate or hamper responding of a Task 1. These results are plausible under the assumption that a stimulus already heightens activation of its associated response, despite ongoing response selection of another task, and only the final selection of a response is subject to capacity-limitations. Following this logic, the presentation and processing of a question in a (dis)honest task could activate its associated honest response. At the same time, the dishonest cue could already boost activation of the very opposite response as the activated response of the question is not the appropriate one to be delivered. As such, part of the difference between honest and dishonest responding could be the result of precentral processing.

These considerations lead to specific predictions for the data pattern of Experiment 1 (e.g., Pashler, 1994a). First, capacity-limited response selection processes should be mirrored in increased RTs and error rates for the short SOA compared to the long SOA of the (Dis)honest Task 2 but not of the Tone Task 1. Second, dishonest responding should be more difficult than honest responding, producing longer RTs and higher error rates in the (Dis)honest Task 2 (i.e., intention effects). Finally, if precentral processes contribute to delays of dishonest responding, these delays should be smaller for the short SOA than for the long SOA (see Figure 2A). Delays due to dishonest as compared to honest responding in later processing stages would affect Task 2 performance independently of

the SOA (see Figure 2B). As such, similar intention effects for both SOA conditions would contradict the assumption of precentral processing as a source of the intention effect.

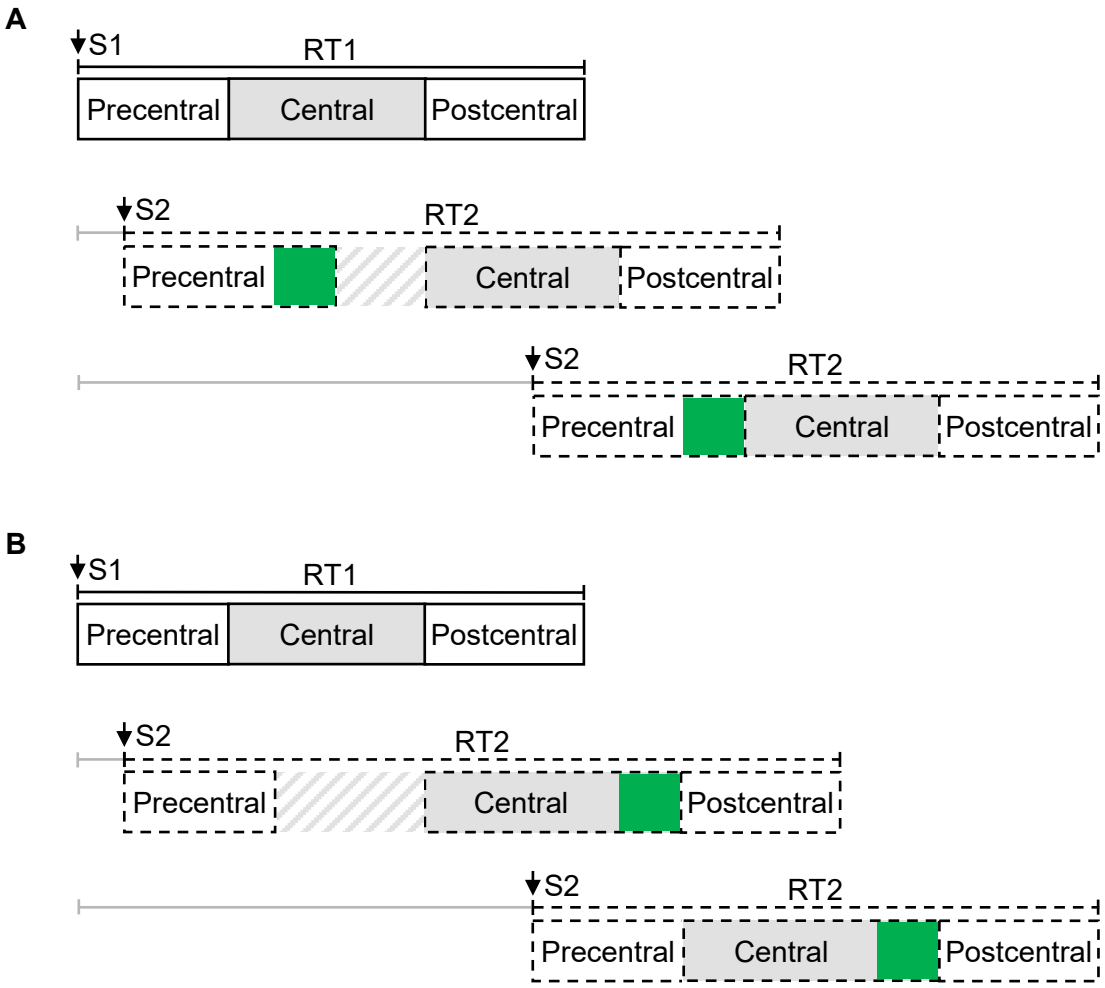

Figure 2 | Illustration of the idea that increased RTs of dishonest responses (dark green [dark gray] areas) occur in precentral response activation (A) or central, capacity-limited response selection (B) and its impact on response times of the Tone Task 1 (solid black lines; RT1) and of the (Dis)honest Task 2 (dashed black lines; RT2) for the short and the long stimulus onset asynchrony (SOA; gray lines) of Experiment 1. Note that the effect of dishonest responding could also be involved in both, the precentral and the central stage and that the current paradigm cannot differentiate between effects in central and postcentral stages (see Experiment 3).

3.2.1. Method

Participants | Thirty-two participants took part in this experiment. This sample size ensures a high power to observe performance differences between honest and dishonest responses as these differences are usually large (for a recent meta-analysis, see Suchotzki et al., 2017). All participants gave informed consent and received monetary compensation or course credits.

Apparatus and stimuli | Participants sat in front of 17" TFT monitors with a display resolution of 1680 × 1050 and a refresh rate of 60 Hz. For the tone task, participants had to classify a 300 Hz and an 800 Hz tone of 100 ms duration as low and high by pressing *K* and *L* with the index and middle fingers of their right hand. For the (dis)honest task, questions were chosen randomly out of a set of 72 questions about daily activities (see Table 1 in Appendix 1). We adapted these questions from previous work (Van Bockstaele et al., 2012), translated them and modified them slightly to make them accessible for German participants (see also Foerster, Wirth, Kunde et al., 2017). Participants pressed *S* and *D* on a standard German QWERTZ keyboard with the middle and index fingers of their left hand to respond to questions with *yes* and *no*. The font color of the questions indicated whether to respond honestly or dishonestly in each trial. The assignment of the font color to intention was counterbalanced across participants as was the key assignment within each task.

Procedure | Participants started the experiment by responding honestly to a random set of questions from the prepared question pool. Participants had to indicate whether they performed these actions during the same day until 10 questions had been negated and 10 questions affirmed. If participants affirmed (negated) more than 10 questions before negating (affirming) 10 other questions, these surplus questions were discarded. Participants were strongly encouraged to consult the experimenter if they were uncertain how to respond or if they had delivered a false response.

Afterward, they learned that they would execute the tone task and the (dis)honest task in close temporal succession during the experiment. They received instructions about both tasks and went through three practice blocks, practicing only the Tone Task 1, only the (Dis)honest Task 2 and then both tasks. Crucially, participants received the instruction to always respond to the tone first without waiting for the question and to deliver a response to the question afterward, and we also stressed both speed and accuracy (cf. Pashler, 1994b; Pashler & Johnston, 1989).

A trial started with the presentation of a central fixation cross for 500 ms. Afterward, the tone played and a question appeared on screen 150 ms (short SOA) or 1500 ms (long SOA) after tone onset. Participants had to deliver both responses within 4000 ms from tone onset. The next trial started after 500 ms. If participants gave an early response before stimulus onset, delivered a false response (commission errors) or failed to deliver

a response in any of the two tasks, or responded to the question before providing a response in the Tone Task 1, error-specific feedback appeared for 1500 ms. The combination of 20 questions × 2 intentions (honest vs. dishonest) × 2 SOA (short vs. long) × 2 tones (low vs. high) resulted in 160 individual trial combinations in a block. Participants went through three of these blocks with self-paced breaks after each 40th trial.

3.2.2. Results

Data and the commented analysis scripts of all experiments are publicly available on the *Open Science Framework* (osf.io/dfgx4).

Data treatment │ The practice blocks and each trial following a self-paced break were excluded from statistical analyses. Error rates were computed as the proportion of commission errors to commission errors plus correct responses. One participant had to be excluded because of delivering false responses during the selection of questions at the beginning of the experiment. The participant informed the experimenter after completing the experiment. All remaining participants committed less than 50% commission errors in all experimental cells and could be considered for all statistical analyses.

Trials following an incorrect trial were excluded (17.5%). Other errors than commission errors in the Tone Task 1 were excluded before analyzing error rates of the Tone Task 1 (0.5%). Error rate analysis of the (Dis)honest Task 2 was restricted to trials with correct tone responses and we then excluded other errors than commission errors of the (Dis)honest Task 2 (1.3%). Only correct trials with inter-response intervals above 100 ms (0.1% excluded) and RTs within 2.5 *SD*s of the corresponding cell mean (4.3% excluded) entered RT analyses of both tasks.

Data analyses │ RTs and error rates of both tasks were analyzed in separate analyses of variance (ANOVAs) with the within-subjects factors SOA (150 ms vs. 1500 ms) and intention (honest vs. dishonest). Significant two-way interactions were scrutinized in planned two-tailed paired-samples *t*-tests. Descriptive statistics of the error rates are presented in Table 4 of Appendix 2 and of the RTs in Table 5 of Appendix 2 and in Figure 3.

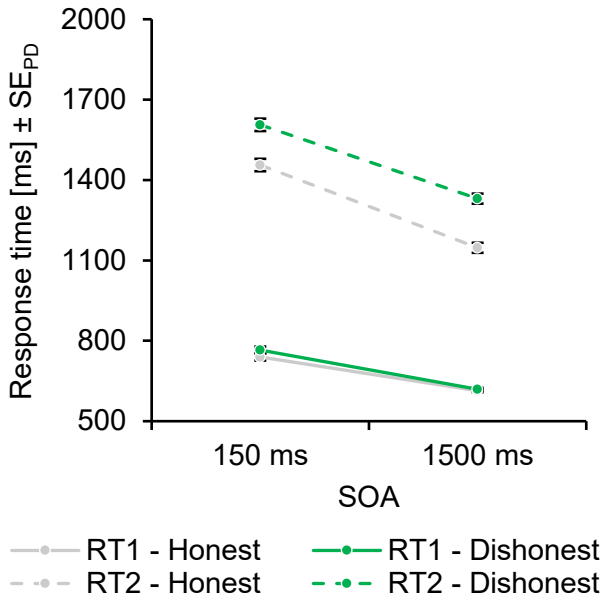

Figure 3 | Mean RTs of the Tone Task 1 (RT1; solid lines) and of the (Dis)honest Task 2 (RT2; dashed lines) of Experiment 1. Light gray lines constitute honest trials and dark green (dark gray) lines dishonest trials. Error bars represent standard errors of paired differences (SE_{PD}; Pfister & Janczyk, 2013), computed separately for each stimulus onset asynchrony (SOA) and task.

Tone Task 1 | Tone RTs were slower with subsequent dishonest than honest responses (Δ = 16 ms), $F(1, 30) = 4.74$, $p = .037$, $\eta_p^2 = .14$, and with short compared to long SOAs (Δ = 136 ms), $F(1, 30) = 25.68$, $p < .001$, $\eta_p^2 = .46$. The interaction of both factors was not significant, $F(1, 30) = 2.05$, $p = .163$, $\eta_p^2 = .06$. The main effects and the interaction were not significant in error rates, Fs < 1.

(Dis)honest Task 2 | Responses were slower when they were dishonest than honest (Δ = 167 ms), $F(1, 30) = 78.68$, $p < .001$, $\eta_p^2 = .72$, and with short SOAs than with long SOAs (Δ = 292 ms), $F(1, 30) = 108.49$, $p < .001$, $\eta_p^2 = .78$. There was a nonsignificant trend toward an interaction of both factors, $F(1, 30) = 3.13$, $p = .087$, $\eta_p^2 = .09$, pointing toward descriptively smaller intention effects with a short (Δ = 150 ms) than with a long SOA (Δ = 183 ms). Responses were less accurate for dishonest than for honest responses (Δ = 5.1%), $F(1, 30) = 23.33$, $p < .001$, $\eta_p^2 = .44$. The main effect of SOA and the interaction were not significant in error rates, Fs < 1.

3.2.3. Discussion

Experiment 1 used the *locus-of-slack logic* to disentangle the involvement of precentral processes from later processes in dishonest responding. As expected, performance measures of the (dis)honest task were worse with short than with long SOAs

and participants had more difficulties with responding dishonestly than honestly. With the (dis)honest task following the tone task, the intention effect was evident for all SOA levels, but there was a nonsignificant trend toward larger effects with the long SOA.[5] The results of the current experiment point toward the recruitment of central or postcentral stages in dishonest responding, which will be disentangled in Experiment 3.

The pattern of results hints toward an impact of response grouping, as RT1 was slower with the short than with the long SOA and in dishonest compared to honest trials, even though we took countermeasures to response grouping. We instructed participants to respond to Task 1 without waiting for Task 2 and to respond as fast and accurate as possible (cf., Pashler, 1994b; Pashler & Johnston, 1989). We further excluded temporally close responses. Note, that such a main effect of SOA in Task 1 was not evident with the reversed task order in Experiment 3, hinting toward a crucial role of task difficulty or task dominance in these phenomena. Whereas the tone task seems to be easily queued up, the (dis)honest task might be too imposing to be completely ignored while processing the tone task.

The current results demonstrate that the contribution of precentral processes to dishonest responding could be at best small while later, possibly capacity-limited processes predominantly account for intention effects. We corroborated these findings in the following experiments; before using a suitable methodology for inferring capacity limitations, we first addressed a potential limitation of the employed stimulus material in the following experiment.

3.3 Experiment 2

In Experiment 1, we had provided our participants with a set of questions at the beginning of the session in order to learn about activities they had or had not performed on the day of the experiment. This procedure is used routinely in research on the cognitive architecture of dishonesty, because it is easily applicable and does not require the participants to engage in mock activities before the actual experiment (Foerster, Wirth, Berghoefer, Kunde, & Pfister, 2018; Foerster, Wirth, Kunde et al., 2017; Spence et al.,

[5] In Experiment 2, this interaction was significant, and we will discuss its implications in the corresponding discussion section.

2001; Van Bockstaele et al., 2012). Despite its widespread use, this procedure may come with several limitations as past instances of the inquired actions (e.g., during the preceding days) may affect responding during the inquiry. To address this limitation, Experiment 2 closely replicated the former *locus-of-slack logic* but introduced a set of activities in the laboratory to apply the same set of questions about these activities in the (Dis)honest Task 2 for all participants (for a similar procedure, see Foerster, Wirth, Herbort et al., 2017). In particular, participants performed one set of activities but did not perform another set of activities and responded to questions about both sets honestly and dishonestly. The same theoretical assumptions as in Experiment 1 hold for this conceptual replication but with higher control of the item set, thus, a similar pattern of results as in Experiment 1 should emerge.[6]

3.3.1. Method

We preregistered this experiment (osf.io/367xw) and invited a new sample of 32 participants. Apparatus, design, and procedure were as for Experiment 1 with the following modifications. For the (dis)honest task, we prepared two sets of 10 activities and corresponding questions (see Table 2 in Appendix 1). We counterbalanced across participants which set of activities had to be performed and provided participants with a box that contained the relevant objects for these activities (each object appeared only for one but not the other set of activities). For example, one half of the participants took apart two bricks that were stuck together with hook-and-loop fasteners. We asked participants to perform each of the actions carefully and presented them consecutively in random order on the computer screen. Participants proceeded through the instructions by keypress (with a forced pause of 5 s between actions). After the completion of all actions, the experimenter checked their accuracy and continued the experiment if all actions were performed correctly or presented action instructions again if an action had not been executed properly. In the (dis)honest task, participants responded to 10 questions about performed actions honestly with *yes* and dishonestly with *no* whereas the opposite was true for the 10 questions about not performed actions.

[6] Note that we conducted this experiment after Experiments 1, 3 and 4, following suggestions of the editor. We thank Nelson Cowan for pointing out this possible limitation and for stimulating a suitable control experiment.

3.3.2. Results

Data treatment and analyses. Data were treated and analyzed as in Experiment 1. We excluded two participants because they committed at least 50% commission errors in one of the design cells and, thus, performed at (or below) chance level. Two other participants had to be excluded because they had an unusually high rate of omissions (one participant never responded in the tone task and the other participant only responded to one of the two tones). We excluded post-error trials (16.9%). Other errors than commission errors in the Tone Task 1 were excluded (0.4%) before analyzing error rates of the Tone Task 1. The error rate analysis of the (Dis)honest Task 2 was restricted to trials with correct tone responses and we then excluded other errors than commission errors of the (Dis)honest Task 2 (1.3%). Only correct trials with inter-response intervals above 100 ms (1.5% excluded) and RTs within 2.5 *SD*s of the corresponding cell mean (4.1% excluded) entered RT analyses of both tasks. Figure 4 shows the main results of the RT analyses. Descriptive statistics of the error rates are presented in Table 6 of Appendix 2 and detailed RT statistics are presented in Table 7 of Appendix 2.

Tone Task 1. Tone RTs were slower with short SOAs compared to long SOAs (Δ = 127 ms), $F(1, 27) = 43.22$, $p < .001$, $\eta_p^2 = .62$. The main effect of intention and the interaction of both factors were not significant, Fs < 1. The main effects and the interaction were not significant in error rates, Fs < 1.46, ps > .237.

(Dis)honest Task 2. Responses were slower when they were dishonest than honest (Δ = 156 ms), $F(1, 27) = 48.49$, $p < .001$, $\eta_p^2 = .64$, and with short SOAs than with long SOAs (Δ = 268 ms), $F(1, 27) = 70.28$, $p < .001$, $\eta_p^2 = .72$. There was also a significant interaction of both factors, $F(1, 27) = 9.49$, $p = .005$, $\eta_p^2 = .26$, as intention effects were smaller with a short (Δ = 132 ms), $t(27) = 5.18$, $p < .001$, $d_z = 0.98$, than with a long SOA (Δ = 181 ms), $t(27) = 8.18$, $p < .001$, $d_z = 1.55$. Responses were less accurate for dishonest than for honest responses (Δ = 8.2%), $F(1, 27) = 58.04$, $p < .001$, $\eta_p^2 = .68$. The main effect of SOA and the interaction were not significant in error rates, Fs < 1.

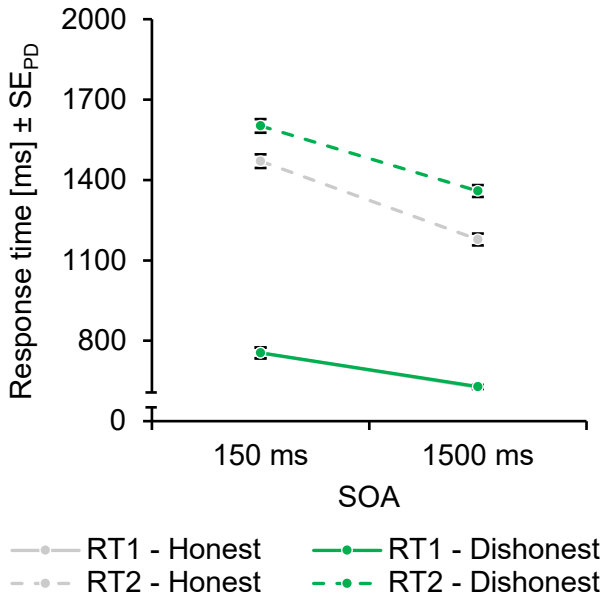

Figure 4 | Mean RTs of the Tone Task 1 (RT1; solid lines) and of the (Dis)honest Task 2 (RT2; dashed lines) of Experiment 2. Light gray lines constitute honest trials and dark green (dark gray) lines dishonest trials. Error bars represent standard errors of paired differences (SE_{PD}), computed separately for each stimulus onset asynchrony (SOA) and task.

3.3.3. Discussion

Experiment 2 conceptually replicated the setup of Experiment 1 to differentiate the contribution of precentral from later processes to dishonest responding. Again, we found a considerable PRP effect and intention effect in the (Dis)honest Task 2. The difference between honest and dishonest responding was pronounced also at the short SOA though it was significantly smaller than with a long SOA. This finding qualifies the descriptive trend observed in Experiment 1 and it points toward somewhat prolonged precentral processing for dishonest compared to honest responses.

One possible explanation for this modulation is that response activation in dishonest responding differed from honest responding (e.g., Hommel, 1998a). Although the question itself should have led to honest response activation in both conditions, the dishonest cue could also have triggered dishonest response activation. Speculatively, these stimuli did not only produce response activation but also honest response inhibition. Previous results from a PRP paradigm with a two-choice Task 1 and a go/no-go Task 2 demonstrated an impact of Task 2 responding on Task 1 responding with slower responses in no-go trials (Miller, 2006). Such an impact of Task 2 on Task 1 processing was, however, not evident in the data of the current experiment.

A second, more speculative explanation relates to a general change in response threshold. The color cue to dishonesty was salient and might have alerted participants toward more cautious processing because of the difficulty of dishonest responding. Accordingly, they could already have been more cautious when they were reading the question. Such an automatic impact of task cues on processing speed has been demonstrated before (Reuss, Kiesel, Kunde, & Hommel, 2011). A more cautious response criterion would result in longer RTs and fewer errors, thus, mirroring the effects of the two-step process in RTs but counteracting them in error rates. Usually, intention effects are indeed smaller in error rates than in RTs, but this could also be the result of higher variance in errors because of fewer observations for errors than for RTs in an experiment. Note that this explanation would predict an intention effect in the Tone Task 1 for the short SOA but not for the long SOA. However, such an interaction did not emerge even though responses again show a pattern of grouping as SOAs affected Task 1 responding as in Experiment 1.

Importantly, the data indicate that precentral processes cannot be the sole source of the difference between dishonest and honest responses. RTs of the (Dis)honest Task 2 differed between the short and long SOA by 268 ms, that is, capacity-limited central processing of Task 2 should have waited for this time period at the short SOA (see Figure 1). Dishonest responding was 181 ms slower than honest responding in trials with a long SOA and assuming that this effect is precentral in nature would predict that it would fall entirely into this waiting period at the short SOA. The same logic applies to the data of Experiment 1 where the intention effect of 183 ms at the short SOA was smaller than the PRP effect of 292 ms.[7]

Taken together, the results so far suggest a contribution of precentral processes to dishonest responding, but they point toward a more dominant role of later processes. The following experiment reversed the order of the tasks to disentangle exactly which of these late processes (capacity-limited response selection vs. postcentral processing) contribute to responding dishonestly.

[7] To strengthen this argument, we conducted another experiment that used the same setup as Experiment 2 but with SOAs of 0 ms and 150 ms. Dishonest responses were slower than honest responses, and this effect did not differ between SOAs. Appendix 3 features a detailed description of this experiment.

3.4 Experiment 3

As stated earlier, theoretical assumptions and empirical evidence suggest that dishonest responding could rely on central, capacity-limited processes but also on postcentral, motoric processes or late, capacity-limited response monitoring processes (e.g., Debey et al., 2014; Duran et al., 2010; Jentzsch et al., 2007). Experiment 3, therefore, assessed the degree of propagation of the intention effect from the (Dis)honest Task 1 to the following Tone Task 2 to differentiate the contribution of these processes to dishonest responding (e.g., Jentzsch et al., 2007; Pashler, 1994a). Response selection cannot proceed for the Tone Task 2 before it has finished for the (Dis)honest Task 1. Prolonged precentral or central processing for a dishonest response compared to an honest response should lead to an even longer idle time in the Tone Task 2 (see Figure 5). As such, the intention effect should propagate to the following Tone Task 2, especially with the short SOA.

Crucially, the degree of propagation with the short SOA is informative to whether motoric or monitoring processes contribute to intention effects. We picked a short SOA of 150 ms, and with this close temporal succession, response selection in the (Dis)honest Task 1 should never be finished before the presentation or start of response selection of the Tone Task 2. As such, the intention effect should fully propagate to the Tone Task 2 if it originates from premotor stages entirely. If postcentral processes contribute to the intention effect, a smaller intention effect should emerge in the Tone Task 2 than in the (Dis)honest Task 1 as this stage is supposed to operate in parallel with all other stages of another task (see Figure 5A). On the other hand, the intention effect of the Tone Task 2 might even be larger than in the (Dis)honest Task 1 if dishonest responding prolongs not only precentral and central stages but also response monitoring (Figure 5B). This monitoring process would not affect responding in the (Dis)honest Task 1 but would delay central processing of the Tone Task 2.

3.4.1. Method

A new sample of 32 participants took part in this experiment. We only list methodological details where this experiment deviated from Experiment 1. Participants went through three practice blocks, practicing only the (Dis)honest Task 1, then only the Tone Task 2 and then both tasks. Participants always had to respond to the question first without waiting for the tone and to deliver a response to the tone afterward. After the

presentation of a fixation cross for 500 ms, the question appeared on screen and a tone played after 150 ms (short SOA) or 1500 ms (long SOA). Participants had to deliver both responses within 4000 ms from tone onset. The next trial started after 500 ms.

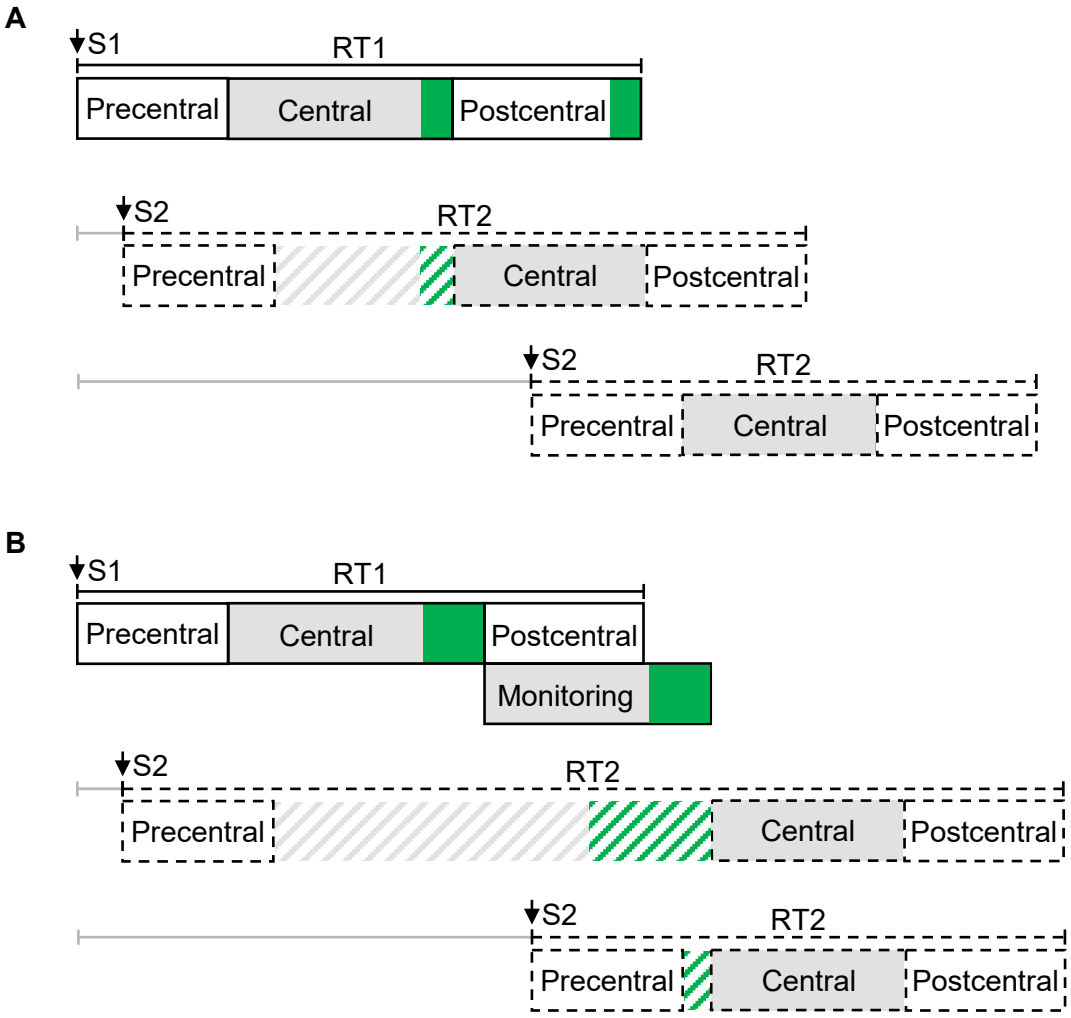

Figure 5 | Illustration of the idea that increased RTs of dishonest responses (dark green [dark gray] areas) occur in central, capacity-limited response selection and postcentral, motor execution (A) or in central, capacity-limited response selection and response monitoring (B) and its impact on response times of the (Dis)honest Task 1 (solid black lines; RT1) and of the Tone Task 2 (dashed black lines; RT2) for the short and the long stimulus onset asynchrony (SOA; gray lines) of Experiment 3.

3.4.2. Results

Data treatment | The same exclusion criteria as in Experiment 1 and 3 were applied. We excluded three participants because they committed at least 50% commission errors in one of the design cells and, thus, performed at (or below) chance level.

Post-error trials were excluded (16.7%) for all statistical analyses. To analyze error rates of the (Dis)honest Task 1, we excluded trials with an erroneous (dis)honest response

that did not constitute a commission error (0.8%). For the error rate analyses of the Tone Task 2, we selected trials with a correct (dis)honest response and a tone response that was correct or constituted a commission error (1.1% other errors excluded). For all RT analyses, we only considered correct trials. We further excluded trials where both responses appeared to be grouped (inter-response interval within 100 ms; 0.6%) and any RT that deviated more than 2.5 *SD*s from the corresponding cell mean (3.2%).

Data analyses | Error rates and RTs of both tasks were analyzed in separate ANOVAs with the within-subjects factors SOA (short vs. long) and intention (honest vs. dishonest). Significant two-way interactions were scrutinized in planned two-tailed paired-samples *t*-tests. In case of significant intention effects in both tasks, these intention effects were compared between both tasks in planned two-tailed paired-samples *t*-tests to assess the extent of propagation from the (Dis)honest Task 1 to the Tone Task 2. These comparisons were made separately for the two SOAs in case of a significant interaction of SOA and intention in one or both of the two tasks. Descriptive statistics of the error rates are presented in Table 8 of Appendix 2 and of the RTs in Table 9 of Appendix 2 and in Figure 6.

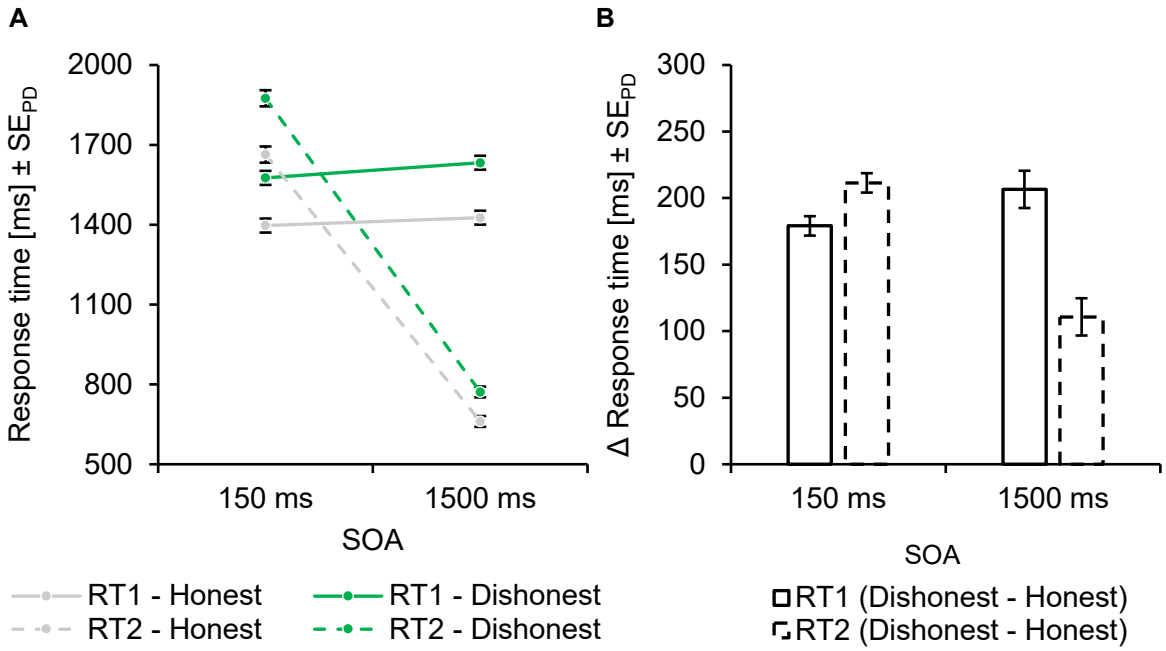

Figure 6 | Mean RTs (A) and mean RT intention effects (B) of the (Dis)honest Task 1 (RT1; solid lines) and of the Tone Task 2 (RT2; dashed lines) of Experiment 3. In the left panel, light gray lines constitute honest trials and dark green (dark gray) lines dishonest trials. In the right panel, RT intention effects were computed as the mean differences between dishonest and honest trials. Error bars represent the standard errors of paired differences (SE_{PD}), computed separately for each stimulus onset asynchrony (SOA) and task in the left panel and for each SOA in the right panel.

(Dis)honest Task 1 | Dishonest response were slower than honest responses (Δ = 193 ms), $F(1, 28) = 63.53$, $p < .001$, $\eta_p^2 = .69$, and also more error-prone (Δ = 4.0%), $F(1, 28) = 21.92$, $p < .001$, $\eta_p^2 = .44$. Neither the main effect of SOA nor the interaction of both factors was significant in RTs, Fs < 1.85, ps > .185, or error rates, Fs < 1.

Tone Task 2 | Tone responses were slower with the short than with the long SOA (Δ = 1053 ms), $F(1, 28) = 1480.05$, $p < .001$, $\eta_p^2 = .98$, and with dishonest than with honest responses in the (Dis)honest Task 1 (Δ = 161 ms), $F(1, 28) = 50.01$, $p < .001$, $\eta_p^2 = .64$. The interaction of both factors was significant in RTs, $F(1, 28) = 17.05$, $p < .001$, $\eta_p^2 = .38$, as the intention effect was larger with the short SOA (Δ = 211 ms), $t(28) = 6.96$, $p < .001$, $d_z = 1.29$, than with the long SOA (Δ = 111 ms), $t(28) = 5.45$, $p < .001$, $d_z = 1.01$. Tone responses were less accurate with the short than with long SOA (Δ = 2.8%), $F(1, 28) = 27.62$, $p < .001$, $\eta_p^2 = .50$. The main effect of intention and the interaction of both factors were not significant in error rates, Fs < 1.23, ps > .277.

Propagation of intention effects | As intention effects were not significant in error rates of the Tone Task 2, we limited our propagation analyses to RTs. Intention effects were smaller in the (Dis)honest Task 1 than in the Tone Task 2 with the short SOA (Δ = -32 ms), $t(28) = -4.43$, $p < .001$, $d_z = -0.82$, but the opposite was true with the long SOA (Δ = 96 ms), $t(28) = 6.84$, $p < .001$, $d_z = 1.27$.

3.4.3. Discussion

Experiment 3 used a PRP paradigm with the *effect propagation logic* to locate processing differences between honest and dishonest responding in postcentral stages of information processing. Participants executed a (dis)honest task shortly before responding to a tone and the temporal distance between question and tone onset was either short or long. In line with the assumption that responses have to be selected consecutively because of capacity limitations (e.g., Pashler, 1994a; Welford, 1952), tone responses were slower and less accurate with the short SOA than with the long SOA, whereas (dis)honest responses were not affected by the SOA manipulation.

More importantly, the current data delivers strong support for the recruitment of precentral and central processes as well as monitoring. The intention effect of the (Dis)honest Task 1 propagated to the Tone Task 2, supporting the assumption that

dishonest responding relies more on precentral and central, capacity-limited processing than honest responding. For the short SOA, the propagated intention effect even exceeded the intention effect of the (Dis)honest Task 1, pointing to stronger recruitment of monitoring processes during dishonest than during honest responding (Jentzsch et al., 2007; Wirth et al., 2015). Preceding descriptions of monitoring processes diverge in their assumptions about the localization of the monitoring process, assuming that it either starts right after response selection (Jentzsch et al., 2007), or at one point during response execution (Kunde, Wirth, & Janczyk, 2018). For dishonest responding, the conflict between the activated honest response and the necessary dishonest response should already be evident during response selection and could initiate response monitoring right after response selection. If these monitoring processes also outlive all motor execution processes, effects in motor execution would be masked and not detectable in the current paradigm (see Figure 5B). Against this background, we assessed the extent of monitoring processes in Experiment 4.

3.5 Experiment 4

The former three experiments used traditional PRP paradigms that originally did not include assumptions about a monitoring process, but evidence for monitoring can be found in effect propagation designs anyway, in terms of propagated effects that exceed their original effects (e.g., Wirth et al., 2015). Monitoring processes can also be studied in designs where the stimulus in one trial appears only after the response in a preceding trial (e.g., Jentzsch & Dudschig, 2009; Wirth, Janczyk et al., 2018). Such a sequential task arrangement allows for an assessment of monitoring processes that outlive all other stages traditionally assumed in stage models. If monitoring after dishonest responding lasts until response selection of the following task, they should hinder these selection processes as both are capacity-limited (Jentzsch et al., 2007; Welford, 1952).

Experiment 4 examined the extent of capacity-limited monitoring processes in dishonest responding and therefore, the experiment again featured the (Dis)honest Task 1 and the Tone Task 2 as in Experiment 3. Crucially, the sequential arrangement of the two tasks did not come with a manipulation of SOAs, that is, the temporal distance between both task stimuli, but employed a variation of the temporal distance between the delivery of the (dis)honest response in Task 1 and the onset of the tone of Task 2 instead (response-stimulus interval; RSI). The tone played either simultaneously with (dis)honest

responding (short RSI of 0 ms) or with a brief temporal delay (long RSI of 1000 ms). Extensive monitoring processes should interfere with response selection of the tone task, leading to a propagation of the intention effect for the short RSI but not for the long RSI.

3.5.1. Method

The experimenter collected data of 33 participants to compensate for the abort of data collection of one participant before the end of the experiment. The experimenter noticed that this participant went through trials that did not feature a question because this participant had responded with *no* only to two of the 72 questions.

Experiment 4 was very similar to the preceding experiments. Accordingly, we only refer to methodological details where this experiment deviates from the former one. Again, the (dis)honest task preceded the tone task, but crucially, the tone always played after the (dis)honest task had been executed. In case of an error in the (Dis)honest Task 1, error-specific feedback immediately appeared for 1500 ms. The Tone Task 2 only followed after a correct response in the preceding (Dis)honest Task 2. The RSI between the (dis)honest response and tone onset amounted to either 0 ms (short RSI) or 1000 ms (long RSI). Participants had to deliver the (dis)honest response within 3000 ms after question onset and the tone response within 1000 ms after tone onset.

3.5.2. Results

Data treatment and analyses | We used the same exclusion criteria and statistical analyses as in the former experiment with two exceptions: the temporal factor was the RSI (short vs. long) instead of SOA, and we did not have to filter grouped responses because grouping was not possible in the current design. One participant committed at least 50% commission errors in one of the experimental cells and could not be considered for any statistical analyses.

All post-error trials were excluded before computing further analyses (20.9%). To analyze error rates of the (Dis)honest Task 1, we excluded errors (0.7%) except commission errors. The Tone Task 2 only followed correct (dis)honest responses, and for the error rate analysis of the Tone Task 2, we excluded other erroneous tone responses than commission errors (3.3% other errors excluded). For RT analyses of both tasks, we excluded all errors and outliers (4.7% outliers excluded). Descriptive statistics of the error

rates appear in Table 10 of Appendix 2 and descriptive statistics of the RTs in Table 11 of Appendix 2 and in Figure 7.

(Dis)honest Task | Dishonest response were slower (Δ = 143 ms), $F(1, 30) = 62.66$, $p < .001$, $\eta_p^2 = .68$, and more error-prone (Δ = 6.7%), $F(1, 30) = 34.54$, $p < .001$, $\eta_p^2 = .54$, than honest responses. Neither the main effect of RSI, nor the interaction of both factors was significant in RTs, Fs < 2.25, ps > .144, or in error rates, Fs < 1.

Tone Task 2 | Tone responses were slower (Δ = 20 ms), $F(1, 30) = 187.98$, $p < .001$, $\eta_p^2 = .86$, and less accurate, (Δ = 1.7%), $F(1, 31) = 10.11$, $p = .003$, $\eta_p^2 = .25$, with the short RSI compared to the long RSI. The main effect of intention and the interaction of both factors were neither significant in RTs, Fs < 1.43, ps > .241, nor in error rates, Fs < 1.

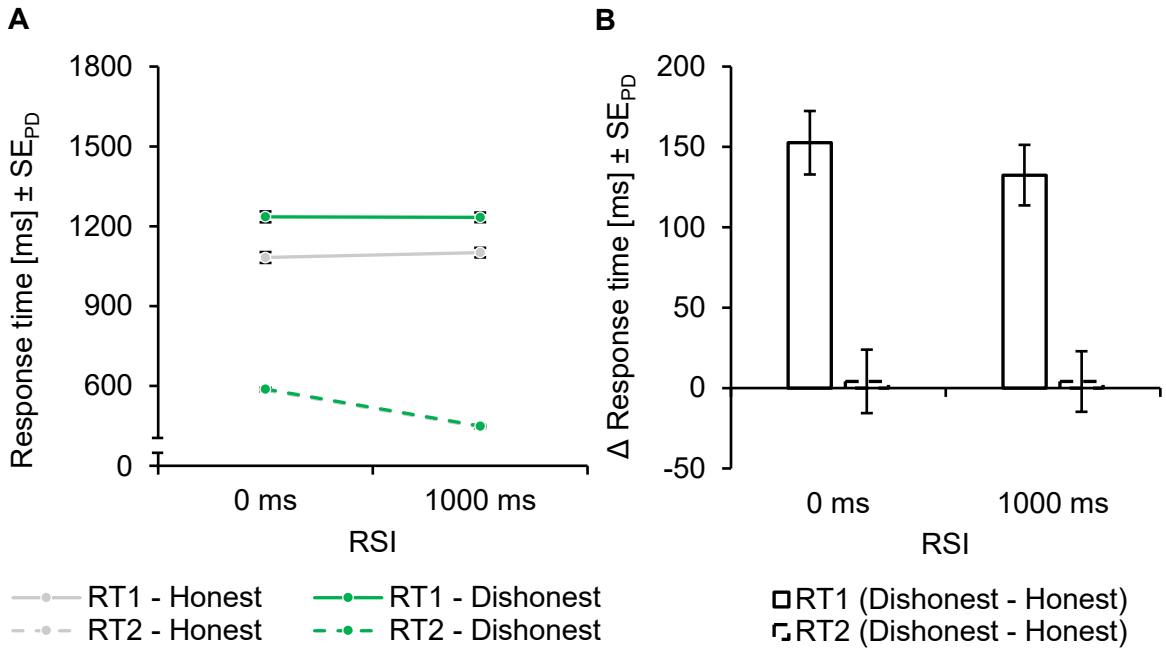

Figure 7 | Mean RTs (A) and mean RT intention effects (B) of the (Dis)honest Task 1 (RT1; solid lines) and of the Tone Task 2 (RT2; dashed lines) of Experiment 4. Note that scaling of the y-axes differs from the former experiments. In the left panel, light gray lines constitute honest trials and dark green (dark gray) lines dishonest trials. In the right panel, RT intention effects were computed as the mean differences between dishonest and honest trials. Error bars represent the standard errors of paired differences (SE_{PD}), computed separately for response-stimulus interval (RSI) and task in the left panel and for RSI in the right panel.

3.5.3. Discussion

The preceding Experiment 3 provided evidence for the notion of a prolonged late, capacity-limited monitoring process for dishonest compared to honest responses but did

not allow for inferences about the extent of this monitoring process. Experiment 4 featured an adapted effect propagation paradigm where the tone task commenced always after the (Dis)honest Task 1 had been finished to assess the extent of dishonest monitoring. The intention effect in the (Dis)honest Task 1 did not propagate to the Tone Task 2. This could mean two things: There was no monitoring process or there was a monitoring process that had been finished before the response selection processes of the tone task began. When taking into account the results of Experiment 3 and 4, the latter explanation appears to be the more plausible one.

In a preceding study, contrast effects in one task were entirely absorbed into the cognitive slack of monitoring processes triggered by an error in the preceding task, even with an RSI of 50 ms (Jentzsch & Dudschig, 2009). If these perceptual processes could fall entirely into the monitoring process, this monitoring process must have been relatively enduring. An important difference between the cited and the current study is the presentation of external feedback. Participants received feedback if they did not provide a correct (dis)honest response and in this case, another (Dis)honest Task 1 instead of a Tone Task 2 followed. The presentation of the tone, thus, served as an explicit signal of a correct response and might have rendered further response monitoring obsolete.

Accordingly, monitoring processes would only be beneficial if there is uncertainty about the appropriateness of the monitored response. Whether monitoring indeed operates this flexibly could be assessed by presenting error feedback either immediately after each response as in the current design or only after the delivery of both tasks' responses. A propagation of monitoring effects should be more probable in the former than in the latter feedback condition.

3.6 General discussion

Four experiments aimed at uncovering the stages of information processing at which inhibition of honest responding and the generation of dishonest responding takes place. Therefore, we combined two powerful and established experimental tools, the instructed intention paradigm from the lying literature and the PRP paradigm with its *effect propagation* and *locus-of slack logic* from sensorimotor stage theories (e.g., Debey et al., 2014; Pashler, 1994a). The resulting data pattern is in line with strong involvement of capacity-limited processes of response selection and a relatively weaker contribution of

precentral processes of response activation (Experiments 1 to 3) as well as capacity-limited processes of monitoring (Experiment 3) in dishonest responding. These monitoring processes are either short-lived and targeted for the intended response or they could be adaptive in length depending on feedback (Experiment 4).

3.6.1. Revisiting the cognitive basis of lies

Empirical research pinpointed the two-step process of truth-inhibition and lie-activation as the basis of dishonest responding when this particular response could not have been rehearsed or prepared in form of a false alibi (e.g., Debey et al., 2014; Foerster, Wirth, Herbort et al., 2017). The current study characterized most of this process as capacity-limited but also as precentral processing. Together, the cue and the question might have already triggered an automatic, activation and/or inhibition of the honest and dishonest response (e.g., Hommel, 1998a; Miller, 2006). More speculatively dishonest cues might signal adaptations in response criterion (Reuss et al., 2011). Examining the role of such speed-accuracy trade-offs in dishonest responding should be the aim of future research. Observing effects of response criteria would call for implementation of these mechanisms in established theories of dishonest processing as the *activation-decision-construction-action theory* (e.g., Walczyk et al., 2014).

The *activation-decision-construction-action theory* already accounts for prolonged monitoring of own behavior when lying, assuming that liars strive to appear convincing and thus increase monitoring and control of their demeanor. The current results suggest that prolonged monitoring in dishonesty can occur on a basic cognitive level of response selection, either specifically because of the presence of response conflict, or more generally because of the difficulty that arises from such conflict. In other words, monitoring own behavior might increase when response selection is difficult, and decrease when it is easy. This perspective on response selection and monitoring suggests that whenever dishonest responding becomes easier as, for example, when rehearsing specific dishonest responses (Hu, Chen et al., 2012; Hu, Rosenfeld, & Bodenhausen, 2012), monitoring should also diminish, allowing subsequent tasks to run more smoothly.

3.6.2. Uncovering hidden postcentral processes

Experiment 4 of the current study did not show any residual monitoring effects when the Tone Task 2 did no longer temporally overlap with the (dis)honest task but rather

followed the (dis)honest response in time. If there were monitoring processes at work, they might either have been finished before interfering with the processing of the tone task, or they might have been stopped by the tone as a signal for correct responding. In the interim discussion of Experiment 3, we already mentioned that monitoring effects could overshadow any intention effects in motor execution. As such, monitoring effects would need to be eliminated to get a grasp on potential motor effects.

Potentially overlapping effects are not only an issue for the examination of dishonest processes but they pertain to basic mechanisms within the PRP paradigm and its assumptions in general (Kunde et al., 2018). A challenge for future studies in this paradigm will thus be to control for monitoring processes when aiming to localize an effect clearly within the motor execution stage.

3.6.3. Open challenges

The PRP paradigm as used in the present studies proved fruitful to map the cognitive architecture of dishonesty to different stages of information processing. In order to employ such methodology, however, our setup intentionally boiled down dishonest responding to the cognitive aspect of truth activation and inhibition. In this setting, participants did not have to make up own lies or practice particular responses. They also did not have to fear any positive or negative consequences of lying. Motivational and emotional aspects as for example, the expectancy of loss or gain (Schindler & Pfattheicher, 2017), the accessibility of justifications for being dishonest (Shalvi et al., 2012), or the extent of reward (Hilbig & Thielmann, 2017) can affect the prevalence of lies and could alter the way lies are processed. Experimental rigor often calls for the instruction of dishonesty as in the current experiments whereas such commands certainly are rare and a special instance of lying when it comes to real-world communication. We would argue, however, that whenever truth activation and inhibition accompany dishonest responding, these processes should prolong mostly response selection but also response activation and monitoring processes. Whenever truth activation and inhibition take a smaller or no role in dishonest responding, these processes should also be recruited to a lesser extent. To scrutinize such assumptions, research should not only confine to the identification of multiple moderators of lying but should also strive to pinpoint their impact on cognitive processing in clear-cut experiments. PRP paradigms deliver a tried-and-tested method to pinpoint the

contribution of experimentally manipulated variables to information processing. Implementing different instances of lying in such systematic investigations will be a challenge and we hope that the current study can be a stepping stone for such approaches.

3.6.4. Conclusion

The current study set out to isolate and localize the activation and inhibition of the truthful response in dishonest responding within specific stages of information processing. First and foremost, the results suggested prolonged response selection when responding dishonestly. Furthermore, our studies pinpointed additional processes to precentral response activation, as well as late capacity-limited response monitoring. Together, the current results demonstrate a pervasive adaptation of information processing in order to produce dishonest responses. To get a full picture of the cognitive underpinnings of dishonest responding, potential contributions of motor execution need to be disentangled from monitoring and different instances of lying need to be taken into account in future studies.

Empirical synopsis: Under control

4 Focused cognitive control in dishonesty

Giving a dishonest response to a question entails cognitive conflict due to an initial activation of the truthful response. Following conflict-monitoring theory, dishonest responding could, therefore, elicit transient and sustained control adaptation processes to mitigate such conflict, and the current experiments take on the scope and specificity of such conflict adaptation in dishonesty. Transient adaptation reduces differences between honest and dishonest responding following a recent dishonest response. Sustained adaptation has a similar behavioral signature but is driven by the overall frequency of dishonest responding. Both types of adaptation to recent and frequent dishonest responses have been separately documented, leaving open whether control processes in dishonest responding can flexibly adapt to transient and sustained conflict signals of dishonest and other actions. This was the goal of the present experiments, which studied (dis)honest responding to autobiographical *yes*/*no* questions. Experiment 5 showed robust transient adaptation to recent dishonest responses whereas sustained control adaptation failed to exert an influence on behavior. It further revealed that transient effects may create a spurious impression of sustained adaptation in typical experimental settings. Experiment 6 and 7 examined whether dishonest responding can profit from transient and sustained adaption processes triggered by other behavioral conflicts. This was clearly not the case: Dishonest responding adapted markedly to recent (dis)honest responses but not to any context of other conflicts. These findings indicate that control adaptation in dishonest responding is strong but surprisingly focused and they point to a potential trade-off between transient and sustained adaptation.

4.1 Introduction

Most people lie regularly, and many do so on a daily basis (e.g., Debey, De Schryver et al., 2015; DePaulo et al., 1996; Halevy et al., 2014; Hilbig & Hessler, 2013). This renders lying an integral part of human communication and, not surprisingly, a considerable amount of research seeks to elucidate such deceptive behavior.

General theoretical frameworks highlight that deception can come in different forms, comprising not only outright lying but also deliberate acts of withholding relevant information or strategically using other conversational norms to one's advantage (see, e.g., recent formulations of information manipulation theory; McCornack, 2015; McCornack, Morrison, Paik, Wisner, & Zhu, 2014). These different kinds of deception may differ in the motivational and cognitive processes that are involved in producing the deceptive response and may require individual empirical approaches.

For the present argument, we focus on outright lying, that is, delivering a factually wrong response. This type of behavior has traditionally been studied either from a motivational perspective or from a cognitive perspective. Motivational approaches to lying typically investigate situational factors, justifications, and moral considerations that will cause a given individual to lie or cheat, and they often employ economic games to incentivize dishonest responding (e.g., Fischbacher & Föllmi-Heusi, 2013; Gneezy, 2005; T. R. Levine, Kim, & Hamel, 2010). Cognitive approaches to lying, by contrast, target the cognitive processes that are assumed to mediate dishonest responding (Debey et al., 2014). Rather than incentivizing dishonest responding, cognitive approaches are based on controlled laboratory tasks that isolate individual processes and therefore allow testing specific predictions from cognitive theories on deception (e.g., *activation-decision-construction-action theory*; Walczyk et al., 2014). Furthermore, the profound interest in the cognitive signature of lying also has potential practical applications: discovering a reliable signature of dishonesty – a cognitive counterpart of Pinocchio's long nose – would be invaluable for the development of lie detection methods.

4.1.1. Two cognitive steps to dishonest responding

Even though previous research did not yet uncover any index that would be as telling as Pinocchio's nose, dishonest responding has been found to recruit a series of cognitive processes that are not as involved during honest responding in a comparable way (Debey

et al., 2014; Walczyk et al., 2003). Because of this difference, honest and dishonest behavior may be understood as being controlled by qualitatively different task sets (Debey, Liefooghe et al., 2015; Foerster, Wirth, Kunde et al., 2017). Lying is set apart from honest responding because it necessarily involves an initial activation of the truthful response as stated in the *activation* component of the *activation-decision-construction-action theory* (Walczyk et al., 2014; for a corresponding theoretical notion, see Truth-Default Theory; T. R. Levine, 2014). This initial honest action tendency has to be inhibited in order to generate a dishonest response, which is more effortful than giving in to an initial action tendency as can be done for honest responding (for a recent review and meta-analysis, see Suchotzki et al., 2017).

Whereas other aspects, such as the construction of a plausible lie, the source of motivation or the intensity of a temptation to lie or tell the truth, can also play an important role in determining the occurrence and difficulty of dishonest responding (e.g., Hilbig & Thielmann, 2017; T. R. Levine et al., 2010; Schindler & Pfattheicher, 2017; Shalvi et al., 2012; Walczyk et al., 2003), the current experiments specifically targeted the described two-step process of activation and inhibition and associated control adaptation.

To isolate this two-step process in controlled experimental tasks, participants are usually asked to respond to simple *yes*/*no* questions about autobiographical or semantic content on a PC with keypresses. They further do not have to fear any negative consequences of their lies. In this setting, dishonest responses have been shown to be slower and less accurate than honest ones, to come with electrophysiological patterns that indicate a more difficult response retrieval and to lead to a stronger activation of brain areas that are associated with executive functions (e.g., Bhatt et al., 2009; Debey et al., 2012; Johnson et al., 2003, 2004; Pfister et al., 2014; Spence et al., 2001; Suchotzki et al., 2015; Walczyk et al., 2003).

4.1.2. Cognitive conflict in dishonesty and means to adaptation

These findings document that the two processing steps required for dishonest responding cause cognitive conflict as they mirror the behavioral and neurophysiological effects that have been observed in a range of cognitive conflict tasks (Botvinick et al., 2001). This two-step process seems to pose a considerable challenge for agents as performance differences between honest and dishonest responses are impressively large, often fueling arguments in favor of using such effects as a basis for lie detection (e.g., Suchotzki et al.,

2017). However, the efficiency of the execution of these dishonest processes is not definite but a function of cognitive control settings and the current experiments thoroughly examine such adaptation of cognitive control in dishonest responding. Viewing dishonest responding from the perspective of cognitive conflict suggests that overcoming conflict – that is, successful dishonest responding – should leave a noticeable fingerprint on the following behavior. In particular, the *conflict-monitoring theory* assumes that cognitive conflicts can be detected and this detection leads to enhanced cognitive control (Botvinick et al., 2001), resulting in smaller conflict effects immediately after another conflict and when conflict is frequent (e.g., Gratton, Coles, & Donchin, 1992; Logan, 1988).

The current study observes dishonest responding in (dis)honest and other conflicting contexts of varying scopes and, thus, provides insight into how cognitive processing of dishonest responses adapts to a wide variety of behavioral contexts. Such a close examination of the scope and specificity of control adaptation in lying contributes to a deeper theoretical understanding of cognitive processing of dishonest responses and provides relevant insights for the development of lie detection methods. The study also puts great emphasis on methodological details of the examination of conflict contexts, which might prompt a reinterpretation of previous research on frequency effects of lying and provide the groundwork for future studies on context effects of lies and other conflicts.

Such context effects have recently been reported in a range of studies that investigated how performance during lying and honest responding is affected by the recency and relative frequency of (dis)honest responding (Debey, Liefooghe et al., 2015; Foerster, Wirth, Kunde et al., 2017; Van Bockstaele et al., 2012; Van Bockstaele et al., 2015; Verschuere et al., 2011). This broader perspective provides an elaborate approach to studying the role of cognitive control for dishonest responding by addressing dynamic changes in cognitive control (i.e., *control adaptation*). Control adaptation becomes visible in improved lying performance if dishonest responses are generated frequently or have been generated immediately before. That is, whenever an agent has lied very recently or frequently, lying becomes easier and possibly even easier than telling the truth. This is a crucial finding for lie detection efforts that seek to classify truth-tellers and liars on the basis of behavioral differences originating from the mentioned effortful cognitive processing of dishonest responses. In a nutshell, a thorough understanding of the different forms of control adaptation

in dishonesty and their appropriate triggers is not only motivated by basic cognitive research efforts but also warranted for the development of cognitive lie detection methods.

4.1.3. The present experiments

So as a first goal, the present experiments targeted the scope of control processes in lying, namely whether transient adaptations to recent dishonest responses and sustained adaptations to frequent dishonest responses operate independently or whether they interact with each other (Foerster, Wirth, Kunde et al., 2017; Van Bockstaele et al., 2012). The current study approached both control mechanisms in dishonesty in concert, to evaluate whether both types of adaptation can operate simultaneously whereas previous studies are limited by studying the impact of either recent dishonest responses or of frequent dishonest responses in separation. The *conflict-monitoring theory* predicts simultaneous adaptation to recent and frequent conflict as both are the result of the same mechanism, namely the detection of conflict in terms of competing response activations (e.g., Botvinick et al., 2001).

Furthermore, the *conflict-monitoring theory* also predicts that control adaptation operates globally for all types of conflict, allowing the transfer of control adaptation between types of conflicts (e.g., Botvinick et al., 2001; Kunde & Wühr, 2006; but see also Braem, Abrahamse, Duthoo, & Notebaert, 2014). As a second goal, the present experiments, therefore, examined the specificity of control processes in lying – that is, how dishonest conflict adapts to transient and sustained contexts of other, unrelated cognitive conflicts as induced via typical conflict tasks (Simon & Rudell, 1967; Stroop, 1935).

4.2 Experiment 5

Adaptation to cognitive conflict in terms of decreased congruency effects can occur either transiently, in response to recent conflict (Gratton et al., 1992), or in a sustained fashion when conflict is frequent (Logan & Zbrodoff, 1979). Because the automatically activated true answer to a question and the actual response of the participant are congruent for honest responses and incongruent for dishonest responses, we will similarly refer to the difference between dishonest and honest responses as a congruency effect. Although descriptions in terms of congruent and incongruent responses are not a common choice in the literature on dishonesty, this terminology emphasizes the potential link to control processes in other domains and it facilitates the description of the statistical

analyses in the following experiments that targeted other sources of conflict besides dishonesty.

Applied to dishonesty, transient conflict adaptation becomes evident in a reduced performance difference between honest and dishonest responses immediately following a dishonest relative to an honest response. Sustained adaptation, by contrast, becomes evident in a reduced performance difference between honest and dishonest responses when dishonest responses are frequent as compared to frequent honest responding. A common method to study sustained adaptation relies on inducer stimuli and probe stimuli. Inducer stimuli are used to manipulate the overall frequency of conflict and we used this method in the current experiments by employing inducer questions that always required either an honest response or always a dishonest response. This manipulation yields frequent honest/dishonest responding, but it also comes with a consistent stimulus-response pairing for each question stimulus because each inducer question always requires the same response and hence stimulus-response regularities can be learned over the course of the experiment. As such, delivering a dishonest response in a dishonest context would be easy because the dishonest response can be directly retrieved from the question and the same is true for honest responses in an honest context. In this case, question-specific learning mechanisms, as well as control adaptation, could be the source of adaptation effects. That is why probe questions (intermixed with inducer questions) have to be answered honestly and dishonestly with an equal frequency to separate control adaptation from question-specific learning. Answering a question with an honest response in half of the trials and with a dishonest response in the other half of trials precludes learning of a particular response to a question (see Foerster, Wirth, Herbort et al., 2017). Indeed, increasing the frequency of dishonest responses to the inducer questions reduced the difference between honest and dishonest performance in probe questions (Van Bockstaele et al., 2012, 2015; Verschuere et al., 2011).

Even though question-specific learning mechanisms cannot drive this effect, it is not clear which top-down control mechanism is responsible for the modulation, that is, transient or sustained processes. Both could be in charge, as changing the frequency of (dis)honest responses also leads to an unbalanced set of transitions between honest and dishonest responding (Foerster, Wirth, Kunde et al., 2017; Van Bockstaele et al., 2012, 2015). When both intentions are instructed with the same frequency and in a random

sequence, “honest → honest”, “honest → dishonest”, “dishonest → honest” and “dishonest → dishonest” sequences appear about equally often. When dishonest responding is more frequent than honest responding, “dishonest → honest” sequences are more likely than “honest → honest” sequences as are “dishonest → dishonest” sequences compared to “honest → dishonest” sequences. In regard to the impact of transient adaptation (e.g., Debey, Liefooghe et al., 2015; Foerster, Wirth, Kunde et al., 2017), those frequent sequences render honest responding relatively difficult (“dishonest → honest” > “honest → honest”) and dishonest responding relatively easy (“dishonest → dishonest” > “honest → dishonest”). For the opposite ratio with more honest than dishonest responding, in contrast, frequent sequences render honest responding relatively easy (“dishonest → honest” < “honest → honest”) and dishonest responding relatively difficult (“dishonest → dishonest” < “honest → dishonest”). Thus, adaptation effects for different proportions of dishonesty could not just stem from sustained control adaptation processes, but it is also plausible to assume that transient control adaptation is the true source of this effect. As such, there would be no general change in attentional processing to favor the frequent task, but only flexible transient adaptation to the recent task (which also happens to be frequent).

Methods to disentangle the influence of transient and sustained adaptation processes have been suggested for standard conflict tasks like the Simon and the spatial Stroop task (Torres-Quesada, Funes, & Lupiáñez, 2013). In both tasks, participants were to press a left and a right key to upward and downward pointing arrows. In the Simon task, the arrows appeared either on the left or on the right side of the display, causing stimulus-response (S–R) incongruency. In the spatial Stroop task, the arrows appeared either on the upper or lower half of the display, causing stimulus-stimulus (S–S) incongruency (e.g., Kornblum, Hasbroucq, & Osman, 1990). In a training block, participants only worked on the Simon task, one group of them with a high proportion of congruent trials, and another group with a low proportion of congruent trials. In the following blocks, participants worked on a random sequence of both tasks with an equal frequency of congruent and incongruent trials. Crucially, the proportion manipulation of the Simon task in the training block transferred to the spatial Stroop task. The congruency effect was smaller for Stroop responses for participants who had responded frequently to incongruent Simon stimuli in

the training block than for those who had frequently responded to congruent Simon trials. This modulation can only be attributed to sustained but not to transient control adaptation.

In a similar vein, the transfer of sustained effects to a situation where transient adaptation is controlled for was examined for honest and dishonest responses (Van Bockstaele et al., 2012). In a design in which the proportion of dishonest trials was manipulated via inducer questions, sustained effects also emerged for balanced probe questions. However, in a subsequent test block, participants gave equally frequent honest and dishonest responses to both, inducer and probe questions. In this condition, and in contrast with the results obtained with standard cognitive conflicts (Torres-Quesada et al., 2013), sustained effects only emerged for inducer questions, but not for probe questions. The continued effect on inducer questions is likely driven by question-specific learning mechanisms. The absent effect on probe questions in this situation gives a further hint that proportion manipulations of dishonesty do not induce sustained but transient adaptation processes by means of changing the frequency of transitions between honesty and dishonesty (Foerster, Wirth, Kunde et al., 2017; Van Bockstaele et al., 2012).

In a recent study, however, transient influences were controlled for with a slightly different design: Inducer and probe trials were arranged in a fixed sequence to hold transient influences constant while examining the impact of sustained influences (Experiment 1 in Van Bockstaele et al., 2015). For example, a sequence of 10 dishonest inducer trials was followed by a sequence of 10 probe trials with honest and dishonest responses in alternation, which were again followed by a sequence of 10 dishonest inducer trials. In this setting, smaller differences between honest and dishonest responding still emerged in error rates but were not evident in RTs with a high frequency compared with a low frequency of dishonest responses. This modulation must stem from sustained adaptation processes as the influence of transient adaptation was held constant.

Taken together, sustained adaptation effects can emerge when transient processes cannot come into action (Experiment 1 in Van Bockstaele et al., 2015) but there are also strong hints that allegedly sustained effects could, in fact, stem from transient adaptation processes to dishonest conflict (Foerster, Wirth, Kunde et al., 2017; Van Bockstaele et al., 2012). A missing puzzle piece is the role of sustained processes when transient processes can operate as well. Do agents adapt to both, recent and frequent dishonest responding

at the same time? And when they do, does adaptation to recent and frequent dishonest responding happen independently or interactively?

Conflict-monitoring theory predicts the presence of both adaptation mechanisms as they merely rely on the detection of conflict, but empirical work suggests that control adaptation does not seem to be an inevitable consequence of recent or frequent conflict experience (Botvinick et al., 2001). Studies on standard cognitive conflicts showed that sustained mechanisms seem to operate independently from transient mechanisms as they did not interact within the same task (e.g., Funes, Lupiáñez, & Humphreys, 2010). For cognitive conflict, sustained control adaptation further transferred between two tasks, while at the same time such transfer was not observed for transient control adaptation in most studies (e.g., Funes et al., 2010; Torres-Quesada et al., 2013; Torres-Quesada, Lupiáñez, Milliken, & Funes, 2014; Wühr, Duthoo, & Notebaert, 2015). These studies suggest that if adaptation to recent and frequent dishonest responding takes place, independent operations of both mechanisms but no interaction between them should be observed. However, this is not necessarily the case for dishonest responding, as the conflict that is triggered by dishonest responding differs from standard conflict tasks, like in the Simon, Eriksen or Stroop task. Whereas the conflicting information is necessary for response selection when giving unrehearsed dishonest responses (e.g., Debey et al., 2014; Walczyk et al., 2014), it can be completely ignored in the standard conflict tasks as it is not necessary to select a response (e.g., Hommel, 2011; Kornblum et al., 1990).

Experiment 5 of the present study tackled the scope of cognitive control in dishonest processing by examining whether transient and sustained adaptation emerge simultaneously and whether those two adaptation processes operate independently or in interaction. Our procedure featured simple *yes/no* questions about daily events and participants were cued to respond honestly or dishonestly in each trial to isolate the dishonest conflict from other processes that are involved in dishonest processing (see, e.g., Foerster, Wirth, Kunde et al., 2017). The proportion of dishonest trials varied between experimental blocks while honest and dishonest responses changed randomly from trial to trial. The manipulation of dishonest proportion was implemented via inducer questions whereas it was always 50/50 for probe questions to control for question-specific learning mechanism (cf. Van Bockstaele et al., 2012, 2015; Verschuere et al., 2011). Accordingly, results on inducer questions provided a manipulation check, whereas the results on probe

questions were of central interest here. For these probe questions, transient and sustained effects were assessed separately. Sustained adaptation should become evident by means of a larger congruency effect in mostly honest than in mostly dishonest contexts, and this interaction effect should still be present when including sequential (transient) factors. That is, congruency effects should be smaller, both after a dishonest than after an honest response in the preceding trial, and when dishonest responses are frequent compared to frequent honest responses. However, preceding studies suggest that transient effects might play a larger role than sustained effects (Foerster, Wirth, Kunde et al., 2017; Van Bockstaele et al., 2015). Accordingly, transient adaptation effects should be stronger than sustained adaptation effects and both adaptation processes are expected to operate independently (e.g., Funes et al., 2010; Torres-Quesada et al., 2013; Wühr et al., 2015).

4.2.1. Method

Participants │ Thirty-two participants (age: *M* = 25.9, *SD* = 8.97; 24 female; 28 right-handed) were recruited. They gave written informed consent and received either monetary compensation or course credit. This sample size ensures a power of 80% to detect a medium effect size d_z of about 0.5 in a two-tailed test (with a. = 5%; calculated with the power.t.test function in R version 3.1.1). Medium effect sizes are a conservative estimate for effects of dishonesty on RTs and error rates and their transient and sustained modulation, because studies on all these effects observed large effects (d_z > 0.80; e.g., Foerster, Wirth, Kunde et al., 2017; Suchotzki et al., 2017; Van Bockstaele et al., 2012). One participant of this sample was excluded from statistical analyses as the number of trials left for RT analyses was more than 2.5 *SD*s below the mean of all participants in at least one experimental cell.

Apparatus and stimuli │ Participants sat in front of a 22-in. TFT monitor. They responded to questions about daily activities from a set of 72 questions (see Table 1 in Appendix 1). These questions were adapted from previous work (Van Bockstaele et al., 2012), translated to German and modified slightly. Participants responded with *yes* and *no* by pressing the keys *D* and *K* on and a standard German QWERTZ keyboard with their index fingers. The assignment of the responses to keys was counterbalanced across participants.

Procedure │ To use an equal amount of questions about already performed and not performed activities in the experiment, participants responded to a random selection of the question pool beforehand. If participants had performed the probed action on the same day, they answered *yes,* whereas they responded *no* if they had not performed it. Participants were to respond at leisure and were strongly encouraged to contact the experimenter if they were uncertain about a response or gave a false one. The procedure stopped when participants had given 10 affirmative and 10 negative questions, respectively. The program discarded any surplus questions if more than 10 affirmative (or negative) answers had been provided before the tenth negative (or affirmative) answer.

Each trial started with a white fixation cross, centrally presented on black background for 250 ms. Then the question (font: Arial, font size: 18 pt.) appeared centrally on black background. The font color of the question was either yellow or blue and indicated whether participants were to respond honestly or dishonestly in the current trial. The assignment of congruency to color was counterbalanced across participants. Furthermore, the response labels *yes* and *no* (font: Arial, 15 pt.) were presented in the lower left and right corner of the display (centered around 25% and 75% in the horizontal and 70% in the vertical of the display) in accordance with response-key assignment. When participants responded too early (during fixation), did not respond within 3000 ms, or provided a false response to the question, they received an appropriate error message for 1500 ms. The next trial started after 250 ms.

In an initial practice block, participants responded to four additional questions (“Are you at a beach?”, “Are you in a room?”, “Are you lying down?”, “Are you sitting in front of the PC?”) eight times honestly and eight times dishonestly in a random order without any response deadline. In the experimental blocks, five affirmative and five negative questions were inducer questions and the remaining 10 questions were probe questions. Inducer questions afforded an unequal frequency of honest and dishonest responses whereas probe questions had balanced frequencies. In one block of the experiment, each inducer question came with an honest instruction in 80% of the trials and a dishonest instruction in 20% of the trials (low dishonest proportion). In the other block of the experiment, the relation was reversed with 20% honest trials (high dishonest proportion). Instructions of the probe questions were 50% honest and 50% dishonest in each of the two blocks. Inducer and probe questions appeared equally often within a block. Accordingly, overall

65% of the trials in the low dishonest proportion block were honest whereas 35% were in the high dishonest proportion block. A block featured 200 trials. The sequence of the dishonest proportion conditions was counterbalanced across participants. All manipulated conditions within a block followed a random sequence. Participants were offered a self-paced break after every 50th trials and between blocks.

4.2.2. Results

Analyses and data treatment | The data and the commented syntaxes with our statistical analyses of all three experiments are publicly available on the *Open Science Framework* (osf.io/gqv8p/). We ran two separate ANOVAs for both, the RT and the error rate data. The first ANOVA was conducted with the within-subjects factors item (inducer vs. probe), dishonest proportion (low vs. high), and current congruency (honest vs. dishonest) to assess whether the dishonest proportion manipulation for the inducer questions transferred to probe questions. This analysis corresponds to previous assessments of sustained conflict adaptation, which does not account for potential transient effects. The second ANOVA targeted sustained and transient effects simultaneously only in probe items by employing the within-subjects factors dishonest proportion (low vs. high), current congruency (honest vs. dishonest), and preceding congruency (honest vs. dishonest). Preceding congruency refers to the congruency in the preceding trial. We scrutinized significant three-way interactions in separate 2 × 2 ANOVAs and two-way interactions in paired-samples *t*-tests and report BFs for these tests in the text. BFs > 3 indicate evidence in favor of the null hypothesis of no effect, whereas BFs < 0.3 indicate evidence in favor of the alternative hypothesis.

For both analyses, we excluded the first trial of each block and those following a break as well as trials that featured the same question as the preceding one to eliminate potential repetition effects (4.5% of trials). We selected trials that were correct or entailed a commission error (i.e., honest response instructed, dishonest response delivered; dishonest response instructed, honest response delivered) and followed a correct trial (12.1% trials excluded) for error analyses. For RT analyses, we also excluded all erroneous trials and those trials that followed them (20.2%). For the second ANOVA, we also excluded inducer trials. Error rates were calculated as the rate of incorrect responses in relation to the remaining correct trials. Outliers were defined as RTs that deviated more

than 2.5 *SD*s from their respective cell mean. Note that the number of observations for each cell differed between the two ANOVAs reported below. Accordingly, two distinct outlier identification and exclusion procedures were conducted (outlier exclusion rate was at 2.5% in the first procedure and at 2.1% in the second procedure).

Inducer versus probe | Table 14 in Appendix 2 depicts the mean error rates and RTs, computed separately for each combination of the factors item (inducer vs. probe), dishonest proportion (low vs. high) and current congruency (honest vs. dishonest). On average each cell included 37 observations.

Dishonest responses were significantly more error-prone than honest responses, $F(1, 30) = 34.74$, $p < .001$, $\eta_p^2 = .54$. Furthermore, a high proportion of dishonest trials increased error rates in comparison to a low proportion of dishonest trials, $F(1, 30) = 9.26$, $p = .005$, $\eta_p^2 = .24$. None of the remaining effects approached significance (Fs < 2.87, ps > .101).

Dishonest responses showed increased RTs in comparison to honest responses, $F(1, 30) = 78.50$, $p < .001$, $\eta_p^2 = .72$. This main effect was qualified by a significant interaction of current congruency and dishonest proportion, $F(1, 30) = 39.79$, $p < .001$, $\eta_p^2 = .57$, as the difference between dishonest and honest responding was evident for both proportions, but considerably larger with a low proportion of dishonest trials, $t(30) = 9.15$, $p < .001$, $d_z = 1.64$, BF < 0.01, than with a high proportion of dishonest trials, $t(30) = 5.35$, $p < .001$, $d_z = 0.96$, BF < 0.01. There was only a nonsignificant trend toward a three-way interaction of all factors, $F(1, 30) = 3.51$, $p = .071$, $\eta_p^2 = .11$, and none of the remaining effects were significant (Fs < 1.50, ps > .230).

Sustained and transient effects combined | Figure 8 shows the mean error rates (upper panels A and B) and RTs (lower panels C and D) for probe items for each combination of current and preceding congruency for low (left panels A and C) and high dishonest proportion trials (right panels B and D). On average each cell included 18 observations.

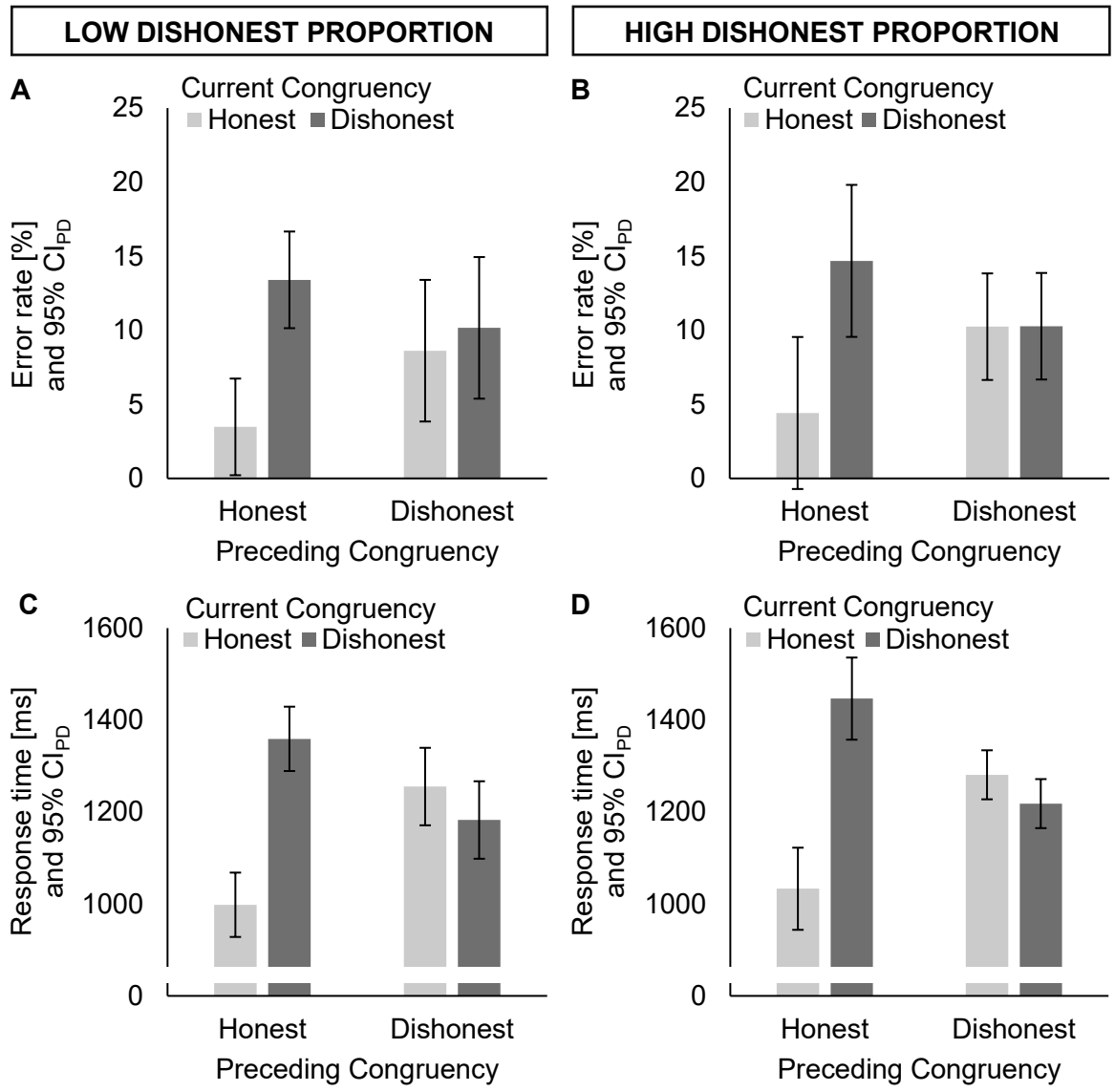

Figure 8 | Mean error rates (upper panels, A and B) and RTs (lower panels, C and D) for probe items in Experiment 5, plotted as function of current congruency and preceding congruency for the low dishonest proportion (left panels, A and C) and the high dishonest proportion (right panels, B and D). Dishonest responses were more error-prone and slower than honest responses after honest responding. A reversed congruency effect was evident after dishonest responding in RTs but not in error rates. Error bars represent the 95% confidence interval of paired differences ($CI_{PD;}$ Pfister & Janczyk, 2013), computed separately for preceding honest and dishonest trials in each dishonest proportion condition.

Mirroring the results reported for the first ANOVA, error rates showed a main effect of current congruency, $F(1, 30) = 22.89$, $p < .001$, $\eta_p^2 = .43$. The main effect of dishonest proportion, $F(1, 30) = 1.05$, $p = .313$, $\eta_p^2 = .03$, and the interaction between current congruency and dishonest proportion, $F < 1$, were not significant. The two-way interaction between current and preceding congruency, $F(1, 30) = 21.28$, $p < .001$, $\eta_p^2 = .42$, was significant because of a considerable congruency effect after honest responses, $t(30) = 6.77$, $p < .001$, $d_z = 1.22$, BF < 0.01, but no such effect after dishonest responses, $t(30) =$

0.51, p = .616, d_z = 0.09, BF = 4.63. None of the remaining effects were significant (Fs < 1).

Again, RTs were higher for dishonest than for honest responses, $F(1, 30) = 76.91$, $p < .001$, $\eta_p^2 = .72$. The main effect of dishonest proportion, $F(1, 30) = 1.63$, $p = .211$, $\eta_p^2 = .05$, and the interaction between current congruency and dishonest proportion, $F(1, 30) = 1.71$, $p = .201$, $\eta_p^2 = .05$, were not significant. The interaction between current and preceding congruency was significant, $F(1, 30) = 78.50$, $p < .001$, $\eta_p^2 = .72$. Dishonest responding was slower than honest responding after honest responses, $t(30) = 10.88$, $p < .001$, $d_z = 1.95$, BF < 0.01, but a smaller reversed effect was evident after dishonest responses, $t(30) = 2.53$, $p = .017$, $d_z = 0.45$, BF = 0.35. None of the remaining effects were significant (Fs < 2.65, ps > .114).

4.2.3. Discussion

The aim of Experiment 5 was to evaluate the scope of conflict adaptation in lying by examining whether sustained adaptation to (dis)honest contexts emerges when agents can also adapt transiently to (dis)honest contexts and whether these adaptation mechanisms work independently or interactively. That is, we investigated how dishonest responding is affected by recent and/or frequent dishonest responding. Participants showed the typical pattern of impaired performance when responding dishonestly in our experiment (e.g., Debey, Liefooghe et al., 2015; Pfister et al., 2014; Spence et al., 2001; Suchotzki et al., 2017). At first sight, the current results also seem to corroborate previous findings on sustained adaptation, as a high dishonest proportion leads to more errors and, more importantly, diminished the congruency effect on RTs in inducer and probe questions (Van Bockstaele et al., 2012; Verschuere et al., 2011). A joint observation of transient and sustained influences for probe items showed, however, a considerable modulation of the congruency effect only through the previous congruency but not through dishonest proportion.

These results demonstrate that transient modulations of the difference between honest and dishonest responding can, in principle, completely account for assumed sustained modulations (Foerster, Wirth, Kunde et al., 2017; Van Bockstaele et al., 2012; but see Van Bockstaele et al., 2015). As noted earlier, manipulating the frequency of dishonest and honest responding also renders certain trial sequences more frequent,

whereas it decreases the frequency of other trial sequences. In low dishonest proportion blocks, an honest trial succeeds mostly another honest trial but it rarely succeeds a dishonest trial. A dishonest trial, however, mostly follows after an honest trial but rarely after another dishonest trial. So the described frequent sequences largely account for honest and dishonest means (the left pair of bars in Figure 8A and Figure 8C), rendering honest responses relatively easy but dishonest responses relatively difficult in this condition (cf. Debey, Liefooghe et al., 2015; Foerster, Wirth, Kunde et al., 2017). The very opposite is true for high dishonest proportion blocks. Now, "dishonest → honest" and "dishonest → dishonest" sequences made up for the majority of trials (the right pair of bars in Figure 8B and Figure 8D), rendering honest responses relatively difficult but dishonest responses relatively easy. The examination of performance in regard to sustained and transient influences, thus, suggests that recent dishonesty changes honest and dishonest processing and that this transient adaptation just happens to also appear frequently.

Although the current study manipulated the proportion of dishonest responding within-subjects, preceding studies relied on a between-subjects comparison (Van Bockstaele et al., 2012, 2015; Verschuere et al., 2011). As such, the absence of sustained effects in the current study could be the result of implementing two different proportion conditions for each participant. Following the suggestion of an anonymous reviewer, we conducted two explanatory analyses on the combined sustained and transient effects in probe trials. First, we selected only the first introduced proportion condition for each participant, thus, having a between-subjects comparison of proportion dishonest.[8] This analysis replicated the presence of transient adaptation effects and the absence of sustained adaptation effects. Second, we introduced the order of dishonest proportion as

[8] ANOVAs with the within-subjects factors current congruency (honest vs. dishonest), and preceding congruency (honest vs. dishonest) and the between-subjects factor dishonest proportion (low vs. high) were conducted only for the first block of each participant and, for the sake of brevity, we only report interactions relating to adaptation effects. The interaction of current and preceding congruency was significant for error rates, $F(1, 29) = 8.07$, $p = .008$, $\eta_p^2 = .22$, and RTs, $F(1, 29) = 62.55$, $p < .001$, $\eta_p^2 = .68$. The interaction of current congruency and dishonest proportion and the three-way interaction were not significant for error rates or RTs (Fs < 1.96, ps > .172).

a between-subjects factor in the original analysis.[9] This analysis indicated that a high proportion of dishonest responses enhanced transient adaptation effects compared to a low proportion of dishonest responses but only for participants who started with the high dishonest proportion condition. Independent sustained adaptation effects were, however, still not evident.

Transient adaptation effects to dishonesty corroborate findings on adaptation after conflicts and, thus, point toward similar underlying control processes of dishonest responding and other conflicts (e.g., Gratton et al., 1992). It is unclear from Experiment 5, however, whether slowed dishonest responding and responses to conflicting stimuli in standard conflict tasks do indeed rely on the same control processes. One possible corollary of such a common mechanism account would predict that control adaptations in one task (e.g., responding in a dishonest trial) generalize to another task (e.g., responding in a conflict trial). Experiment 6 and 7 were designed to test whether transient and sustained control settings generalize from other conflict tasks to dishonest responding and vice versa.

---

[9] ANOVAs with the within-subjects factors dishonest proportion (low vs. high), current congruency (honest vs. dishonest), and preceding congruency (honest vs. dishonest) and the between-subjects factor proportion order (low first vs. high first) were conducted and, for the sake of brevity, we only report interactions relating to adaptation effects and their modulation by proportion order. Error rates showed a significant interaction of current and preceding congruency, $F(1, 29) = 20.75$, $p < .001$, $\eta_p^2 = .42$, which was not further modulated by proportion order ($F < 1$). The interaction of current congruency and proportion dishonesty was not significant and was also not qualified by proportion order (Fs < 1). The interaction of current, preceding congruency and dishonest proportion as well as the four-way interaction were not significant (Fs < 1). In RTs, the interaction of current and preceding congruency was significant, $F(1, 29) = 80.14$, $p < .001$, $\eta_p^2 = .73$, but was not further modulated by proportion order, $F(1, 29) = 1.97$, $p = .171$, $\eta_p^2 = .06$. The interaction of current congruency and proportion dishonesty was not significant and was also not qualified by proportion order (Fs < 1.66, ps > .208). The interaction of current, preceding congruency and dishonest proportion was not significant ($F < 1$), but the four-way interaction was significant, $F(1, 29) = 8.14$, $p = .008$, $\eta_p^2 = .22$. The interaction of current, preceding congruency and dishonest proportion was not significant in the low first condition, $F(1, 14) = 2.15$, $p = .165$, $\eta_p^2 = .13$, but in the high first order condition, $F(1, 15) = 6.86$, $p = .019$, $\eta_p^2 = .31$, as transient adaptation effects were larger in the high than in the low proportion dishonest condition, $t(15) = 2.62$, $p = .019$, $d = 0.65$.

4.3 Experiment 6

Experiment 5 focused on adaption processes within the domain of conflict that is triggered by dishonest responding. However, there is currently no data to assess whether control adaptation processes can transfer between dishonesty and other behavioral conflicts. That is, whether dishonest responding triggers conflict adaptation for behavioral conflicts that are unrelated to lying and vice versa.

While *conflict-monitoring theory* assumes that transfer of adaptation should emerge between different types of conflict (e.g., Botvinick et al., 2001), research on standard cognitive conflicts has identified conditions that render the transfer of control adaptation between different tasks and/or conflicts more or less likely. Relevant moderators for the transfer of transient and sustained control adaptation include the similarity of relevant stimulus dimensions, conflict dimensions and context as well as task boundaries (Hazeltine, Lightman, Schwarb, & Schumacher, 2011; Notebaert & Verguts, 2008; Spapé & Hommel, 2008; Wühr et al., 2015; for a review on transient transfer effects, see Braem et al., 2014). The transfer of transient and sustained processes do not necessarily go hand in hand as the transfer for one of the adaptation mechanisms can emerge while at the same time the other adaptation mechanism operates task-specifically (e.g., Torres-Quesada et al., 2013; Wühr et al., 2015).

There are also proposals and observations that particularly distinctive tasks can share control settings, presumably, because interference between the two tasks is low (Braem et al., 2014) or agents are especially motivated to use high levels of control in a task (Kleiman, Hassin, & Trope, 2014). In the latter study, control adaptation transferred from a standard letter Flanker task to a task that measures stereotypical biases by using a prime and a target (Kleiman et al., 2014, Experiment 2). The prime showed either a white or a black face, the target a weapon or a tool. Participants were to classify the targets as tools or weapons. Participants showed stereotypical biases with a faster weapon- and a slower tool-identification after the presentation of black compared to white faces, critically, only after congruent but not after incongruent standard Flanker trials. The authors argued that control settings might have passed from one task to the other despite their distinctiveness as people do not want to appear biased and, thus, benefit from the transfer.

The transfer of control adaptation between standard conflict tasks was the target of many empirical studies and researchers made first steps toward the definition of clear boundary conditions for the emergence or failure of this transfer. Still, there is plenty of work to do to understand how (un)specific control processes operate and which conditions set the parameters of the scope of transfer (e.g., Braem et al., 2014). In the same vein, previous studies on conflict adaptation in dishonesty have addressed very specific processes. Especially, as the dishonest conflict differs from the usually observed conflicts (i.e., the conflicting information is relevant for task execution), it is difficult to make a prediction about whether and how strongly control settings can transfer from the dishonest to another conflict or vice versa. While control adjustments to standard conflicts would ideally lead to complete inhibition of the irrelevant stimulus or stimulus dimension, this is not useful for dishonest responding. For dishonest responding, it would be plausible to assume that experience of dishonesty improves dishonest processing by means of facilitating the switch from the dominant honest response to the appropriate dishonest response. However, examinations about how control adaptation affects dishonest responding in particular are not available, yet. Existing evidence on stereotypical biases (Kleiman et al., 2014) suggests that transfer could take place from standard conflict tasks to responding dishonestly, as a successful liar should normally be inclined to hide dishonesty (like stereotypical biases). To examine control transfer between conflicts, Experiment 6 combined the setup of Experiment 5 with a Stroop task. In the Stroop task, participants have to respond to the font color of a color word while ignoring the semantic meaning of the word (e.g., RED printed in blue; Stroop, 1935). So relevant and irrelevant stimulus dimensions overlap in this task and cause response conflict when these stimuli are mapped to different responses (e.g., Kornblum et al., 1990). In dishonest responding, conflict emerges from the automatic activation of the dominant truthful response and the required response that has to be derived from the dominant one (e.g., Debey, Liefooghe et al., 2015). Even though the two tasks and their sources of conflict differ considerably, control adaptation settings could transfer between both tasks (e.g., Braem et al., 2014; Kleiman et al., 2014). The current experiment targets whether sustained control adaptation from the Stroop task can generalize to the (dis)honest task and whether transient control adaptation transfers from one of the tasks to the other in whatever

direction. Therefore, both tasks appeared in a random sequence while the proportion of congruency within the Stroop task was manipulated between experimental blocks.

As previous findings suggest that transient and sustained effects should operate independently with a greater chance of transfer in the sustained domain (e.g., Funes et al., 2010; Torres-Quesada et al., 2013; Wirth, Pfister, & Kunde, 2016; Wühr et al., 2015), we expected a transfer of sustained Stroop conflict adaptation to (dis)honest responding. This should lead to smaller differences between honest and dishonest responding in the high conflict context compared to the low conflict context. In addition, if agents adapt similarly to recent dishonest and Stroop conflict, we should observe unspecific effects of transient control adaptation. Congruency effects should be smaller after dishonest and incongruent trials.

4.3.1. Method

Participants | A new sample of 32 participants (age: *M* = 30.6, *SD* = 9.34; 22 female; 30 right-handed) took part in the experiment for either monetary compensation or course credit. All participants gave written informed consent. Based on the same criteria as in Experiment 5, 3 participants could not be considered in the following statistical analyses.

Apparatus and stimuli | Experiment 6 was similar to Experiment 5 except for the following changes. Participants sat in front of a 17-in. monitor and responded on and a standard German QWERTZ keyboard. In this experiment, participants responded with *yes* and *no* by pressing the keys *A* and *S* with their left middle and index finger. The assignment of the responses to keys was counterbalanced across participants. The Stroop task featured four color words and font colors: blue, brown, yellow and purple. Participants were to respond according to the font color of the color word with their right index (blue), middle (brown), ring (yellow) and little finger (purple) which rested on the adjacent keys *K*, *L, Ö,* and *Ä*. The keys were marked with appropriately colored labels. We used four font colors, color words, and responses in the Stroop task so we could select trial sequences for statistical analyses with complete stimulus alternations to control for feature integration within the Stroop task. This stimulus constellation is still confounded with the proportion manipulation as the color word is highly predictive for the response when the proportion of congruent trials is high. However, this problem only relates to effects within the Stroop task but not to the transfer effects between tasks.

Procedure | Again, participants started with a pre-experimental procedure to select an equal amount of questions that asked about activities that had been performed and activities that had not been performed on the same day. In contrast to Experiment 5, a selection of 15 affirmative and 15 negative questions was taken from the question pool.

The trial procedure was the same as in Experiment 5 except that in half of the trials, a Stroop stimulus instead of a question was presented. The position of the Stroop stimulus (font: Arial, 15 pt.) was in the center of the display on black background. As in (dis)honest trials, participants had to respond within 3000 ms in the Stroop task. Stroop and (dis)honest trials appeared in a random sequence. The font style of the question (i.e., bold or italic) indicated whether participants were to respond honestly or dishonestly in the current trial to disentangle the congruency manipulations of both tasks. The assignment of intention to font style was counterbalanced across participants. After the first practice block with questions, a new practice block introduced participants to the Stroop task. Each possible color word/font color combination appeared once, resulting in 16 practice trials without a response deadline.

Participants responded to each question equally often honestly and dishonestly within a block, whereas the proportion congruency of the Stroop task was varied between blocks. In one half of the experiment (i.e., four blocks), there was a low conflict proportion with 80% congruent Stroop trials (i.e., color word same as font color) and 20% incongruent Stroop trials (i.e., color word different than font color). In the other half of the experiment, the relation was reversed, with 20% congruent trials (high conflict proportion). A block featured 120 trials, that is, 60 (dis)honest and 60 Stroop trials. The sequence of the conflict proportion conditions was counterbalanced across participants. Between blocks and after each 40th trial, there was a self-paced break.

4.3.2. Results

Data treatment | Data exclusion followed the same rules as in Experiment 5. Trials were excluded from further analyses when they featured the same task as the preceding trial with, also, either the same question (0.8%) or the same Stroop color word and/or Stroop font color to control for feature repetition effects (19.8%). Before analyzing error rates, we selected trials that were correct or entailed a commission error and followed a correct trial and excluded all other trials (10.1% of (dis)honest trials, 9.6% of Stroop trials

were excluded). For RT analyses, we excluded all erroneous trials and those trials that followed them (19.7% of (dis)honest trials, 14.0% of Stroop trials). For analyses on sustained effects (see below), 2.3% of (dis)honest trials and 2.7% of Stroop trials were identified as outliers and excluded. For analyses on transient effects, outlier exclusion amounted to 2.3% and 2.9%, respectively.

Sustained effects │ Error rates and RTs were analyzed in a 2 × 2 × 2 ANOVA with the within-subjects factors task ([dis]honest vs. Stroop), Stroop conflict proportion (low vs. high) and current congruency (congruent vs. incongruent). Sustained adaptation effects should produce a significant interaction between conflict proportion and current congruency. When such adaptation processes fully transferred from the Stroop to the (dis)honest task, there should be no three-way interaction between all three factors. We scrutinized significant three-way interactions in separate 2 × 2 ANOVAs and two-way interactions in paired-samples *t*-tests and report BFs for these tests in the text.

Figure 9 shows the mean error rates (upper panels A and B) and RTs (lower panels C and D) for each combination of current congruency and conflict proportion for the (dis)honest task (left panels A and C) and the Stroop task (right panels B and D). On average each cell included 83 observations.

Critically, the analysis of error rates showed that neither the interaction between current congruency and conflict proportion ($F < 1$) nor the interaction between current congruency, conflict proportion and task was significant, $F(1, 28) = 1.44$, $p = .241$, $\eta_p^2 = .05$. Significant main effects of task, $F(1, 28) = 13.50$, $p = .001$, $\eta_p^2 = .33$, and current congruency, $F(1, 28) = 9.25$, $p = .005$, $\eta_p^2 = .25$ emerged. The (dis)honest task was more error-prone than the Stroop task as were dishonest/incongruent trials in comparison to honest/congruent trials. However, the two-way interaction of these factors was significant, $F(1, 28) = 4.43$, $p = .044$, $\eta_p^2 = .14$, as the difference in error rates was evident for the comparison of dishonest and honest trials, $t(28) = 3.24$, $p = .003$, $d_z = 0.60$, BF = 0.08, but not for the comparison of incongruent and congruent Stroop trials, $t(28) = 0.57$, $p = .574$, $d_z = 0.11$, BF = 4.36. None of the remaining effects were significant (Fs < 0.28, ps > .603).

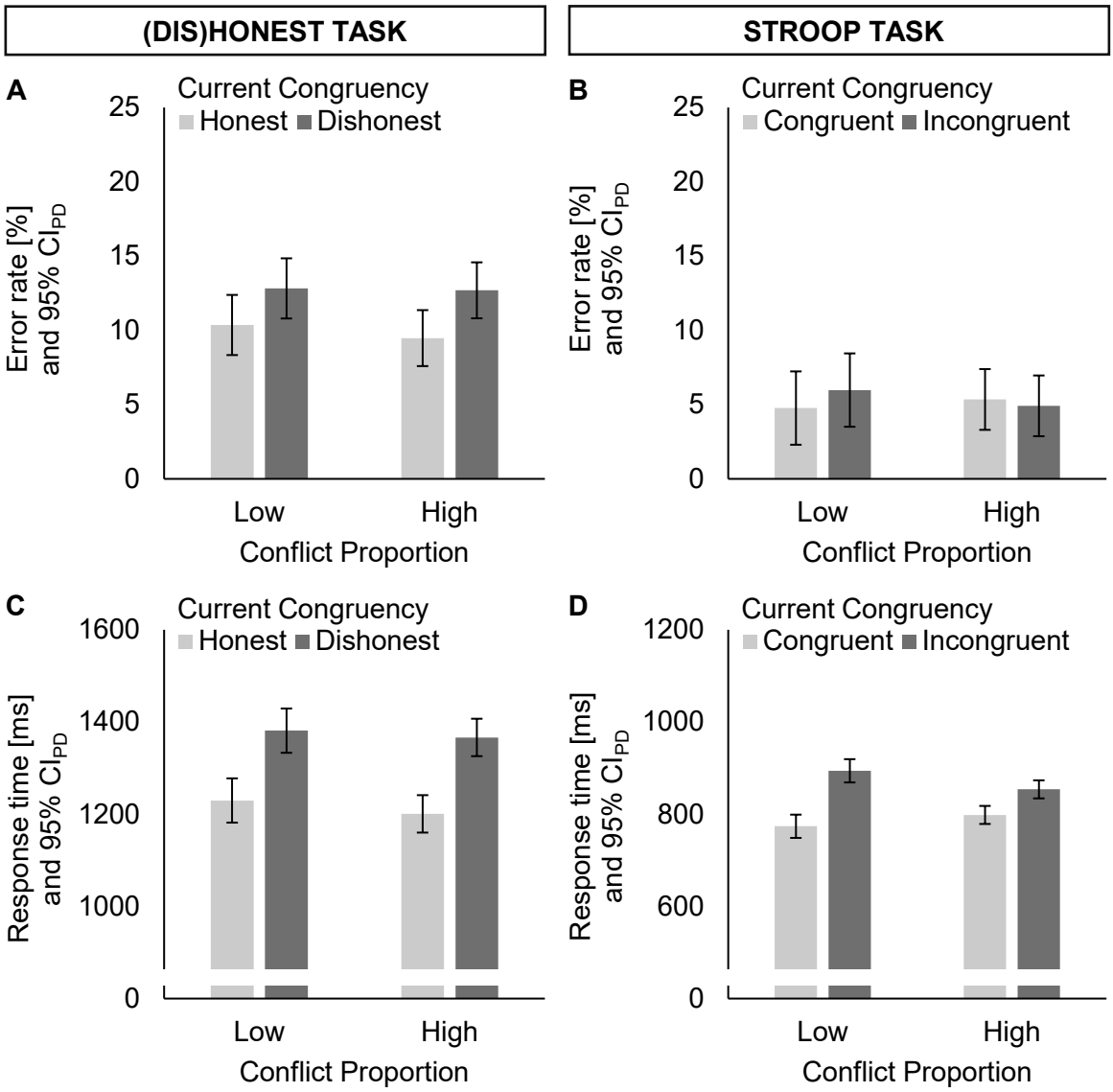

Figure 9 | Sustained conflict adaptation effects on error rates (upper panels, A and B) and RTs (lower panels, C and D) in Experiment 6, plotted as function of current congruency and conflict proportion for the (dis)honest task (left panels, A and C) and the Stroop task (right panels, B and D). Note that the RT plots are scaled differently. Dishonest responses were more prone to error than honest responses but error rates in incongruent and congruent trials of the Stroop task were similar. The congruency effect in RTs was a bit larger in the Stroop than in the (dis)honest task. In the Stroop task, the congruency effect in RTs was smaller in the high than in the low conflict context but this effect did not transfer to the (dis)honest task. Error bars represent the 95% CI_{PD}, computed separately for low and high conflict proportion in each task.

RTs showed a significant two-way interaction between current congruency and conflict proportion, $F(1, 28) = 5.42$, $p = .027$, $\eta_p^2 = .16$, and a significant three-way interaction of all factors, $F(1, 28) = 7.10$, $p = .013$, $\eta_p^2 = .20$. Furthermore, responses in the (dis)honest task were slower than in the Stroop task, $F(1, 28) = 144.25$, $p < .001$, $\eta_p^2 = .84$. Current congruency affected RTs, $F(1, 28) = 169.33$, $p < .001$, $\eta_p^2 = .86$, and was

qualified by a significant interaction between task and current congruency, $F(1, 28) = 10.28$, $p = .003$, $\eta_p^2 = .27$. None of the remaining effects were significant (Fs < 0.51, ps > .479).

Separate ANOVAs for the two tasks clarified the former three-way and two-way interactions in RTs and we only report effects that help to understand these interactions. Conflict proportion and current congruency did not interact in the (dis)honest task ($F < 1$, BF = 4.27), but it did in the Stroop task, $F(1, 28) = 33.88$, $p < .001$, $\eta_p^2 = .55$, BF = 4.27, with a larger congruency effect in low conflict proportion blocks, $t(28) = 9.77$, $p < .001$, $d_z = 1.81$, BF < 0.01, than in high conflict proportion blocks, $t(28) = 5.80$, $p < .001$, $d_z = 1.08$, BF < 0.01. The main effect of congruency was significant in both tasks, but smaller in the (dis)honest task, $F(1, 28) = 75.67$, $p < .001$, $\eta_p^2 = .73$, BF < 0.01, compared to the Stroop task, $F(1, 28) = 85.42$, $p < .001$, $\eta_p^2 = .75$, BF < 0.01.

Transient effects │ Second, a 2 × 2 × 2 × 2 ANOVA with the within-subjects factors task ([dis]honest vs. Stroop), task sequence (repetition vs. switch), current congruency (honest/congruent vs. dishonest/incongruent) and preceding congruency was conducted on error rates and RTs. The factor task sequence describes whether the preceding trial featured the same task as the current trial (repetition) or the other task (switch). Transient adaptation effects should produce a significant interaction between current and preceding congruency.

When such adaptation processes transfer between tasks, this two-way interaction should not be further qualified by task sequence. We scrutinized significant three-way and four-way interactions in separate planned ANOVAs and significant two-way interactions in planned paired-samples t-tests and report BFs for these tests in the text.

Figure 10 shows the mean error rates and Figure 11 depicts the mean RTs for each combination of current and preceding congruency for task repetitions (upper panels A and B) and task alternations (lower panels C and D) in the (dis)honest task (left panels A and C) and the Stroop task (right panels B and D). On average each cell included 42 observations.

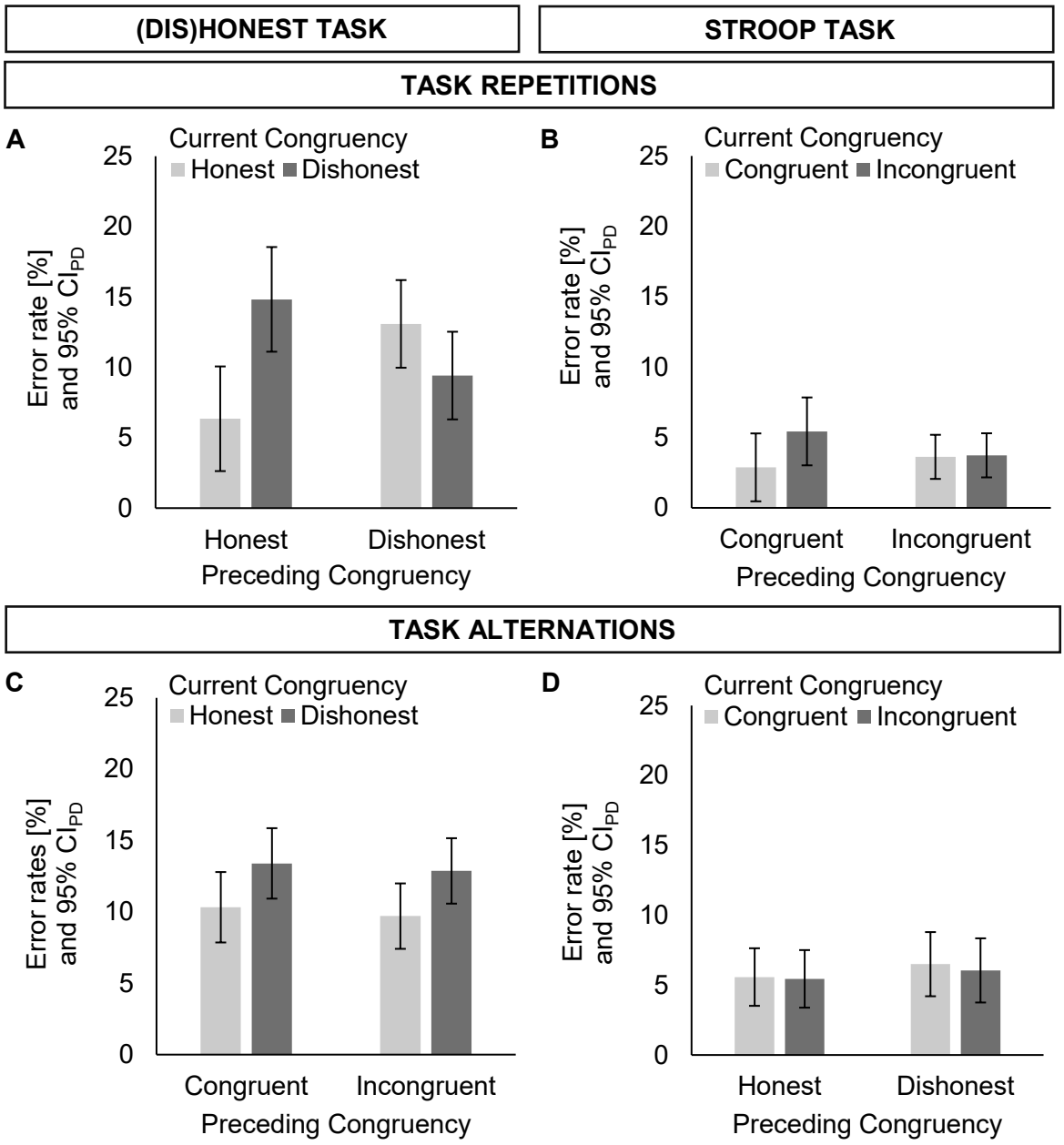

Figure 10 | Transient conflict adaptation effects on error rates in Experiment 6, plotted as function of current and preceding congruency, separately for task repetitions (upper panels, A and B) and task alternations (lower panels, C and D) for the (dis)honest task (left panels, A and C) and the Stroop task (right panels, B and D). Errors only showed a modulation by current and preceding congruency when the (dis)honest task was repeated (A). After honest responses, dishonest responses were more error-prone whereas after dishonest responses honest responses were more error-prone. Error bars represent the 95% CI_{PD}, computed separately for the conditions of preceding congruency and task sequence in each task.

For error rates, the two-way interaction of current and preceding congruency was significant, $F(1, 28) = 12.63$, $p = .001$, $\eta_p^2 = .31$, which was further modulated by task and task sequence, as indicated by the respective three-way interactions (Task × Current Congruency × Preceding Congruency: $F(1, 28) = 9.20$, $p = .005$, $\eta_p^2 = .25$; Task Sequence

× Current Congruency × Preceding Congruency: $F(1, 28) = 16.62$, $p < .001$, $\eta_p^2 = .37$). Finally, the four-way interaction was significant, $F(1, 28) = 12.91$, $p = .001$, $\eta_p^2 = .32$. Furthermore, a switch between tasks resulted in more errors compared to task repetitions, $F(1, 28) = 9.38$, $p = .005$, $\eta_p^2 = .25$. Mirroring the error rate analysis on sustained effects, the main effects of task, $F(1, 28) = 14.31$, $p = .001$, $\eta_p^2 = .34$, and current congruency, $F(1, 28) = 10.22$, $p = .003$, $\eta_p^2 = .27$, and the two-way interaction of both factors were significant, $F(1, 28) = 4.88$, $p = .036$, $\eta_p^2 = .15$. None of the remaining effects was significant (Fs < 2.62, ps > .117).

Separate ANOVAs on error rates for the (dis)honest and the Stroop task were conducted to scrutinize the former interactions and, for the sake of brevity, we only report those effects that are informative to understand the former interactions. Stroop error rates showed no significant interaction of current and preceding congruency, $F(1, 28) = 1.91$, $p = .178$, $\eta_p^2 = .06$, BF = 2.15, or of current congruency, preceding congruency and task sequence, $F(1, 28) = 1.50$, $p = .231$, $\eta_p^2 = .05$, BF = 2.57. However, as the interaction of the initial ANOVA suggested, the two-way interaction between current and preceding congruency was significant for the (dis)honest task, $F(1, 28) = 15.50$, $p < .001$, $\eta_p^2 = .36$, BF = 0.02, and was also further modulated by task sequence, $F(1, 28) = 0.97$, $p < .001$, $\eta_p^2 = .43$, BF < 0.01. Separate ANOVAs on task repetition and switches for (dis)honest trials showed that a significant interaction of current and preceding congruency only emerged for task repetitions, $F(1, 28) = 21.99$, $p < .001$, $\eta_p^2 = .44$, BF < 0.01, but not for switches, $F < 1$, BF = 5.04. When tasks repeated from the preceding to the current trial, dishonest responses were more error-prone than honest responses after honest responding, $t(28) = 4.67$, $p < .001$, $d_z = 0.87$, BF < 0.01, but the pattern of results was reversed after dishonest responding, $t(28) = 2.41$, $p = .023$, $d_z = 0.45$, BF = 0.44.

For RTs, current and preceding congruency interacted significantly, $F(1, 28) = 79.32$, $p < .001$, $\eta_p^2 = .74$, however this interaction was again modulated. The three-way interactions between task, current and preceding congruency, $F(1, 28) = 171.11$, $p < .001$, $\eta_p^2 = .86$, and between task sequence, current and preceding congruency, $F(1, 28) = 119.68$, $p < .001$, $\eta_p^2 = .81$, as well as the four-way interaction of all factors, $F(1, 28) = 104.63$, $p < .001$, $\eta_p^2 = .79$, were significant. Switches between tasks prolonged responses compared to task repetitions, $F(1, 28) = 141.41$, $p < .001$, $\eta_p^2 = .84$. In line with the RT analysis on sustained effects, there were significant main effects of task, $F(1, 28) = 146.03$,

$p < .001$, $\eta_p^2 = .84$, and current congruency, $F(1, 28) = 153.91$, $p < .001$, $\eta_p^2 = .85$, as well as a significant two-way interaction of these factors, $F(1, 28) = 13.91$, $p = .001$, $\eta_p^2 = .33$. None of the remaining effects was significant (Fs < 2.06, ps > .162).

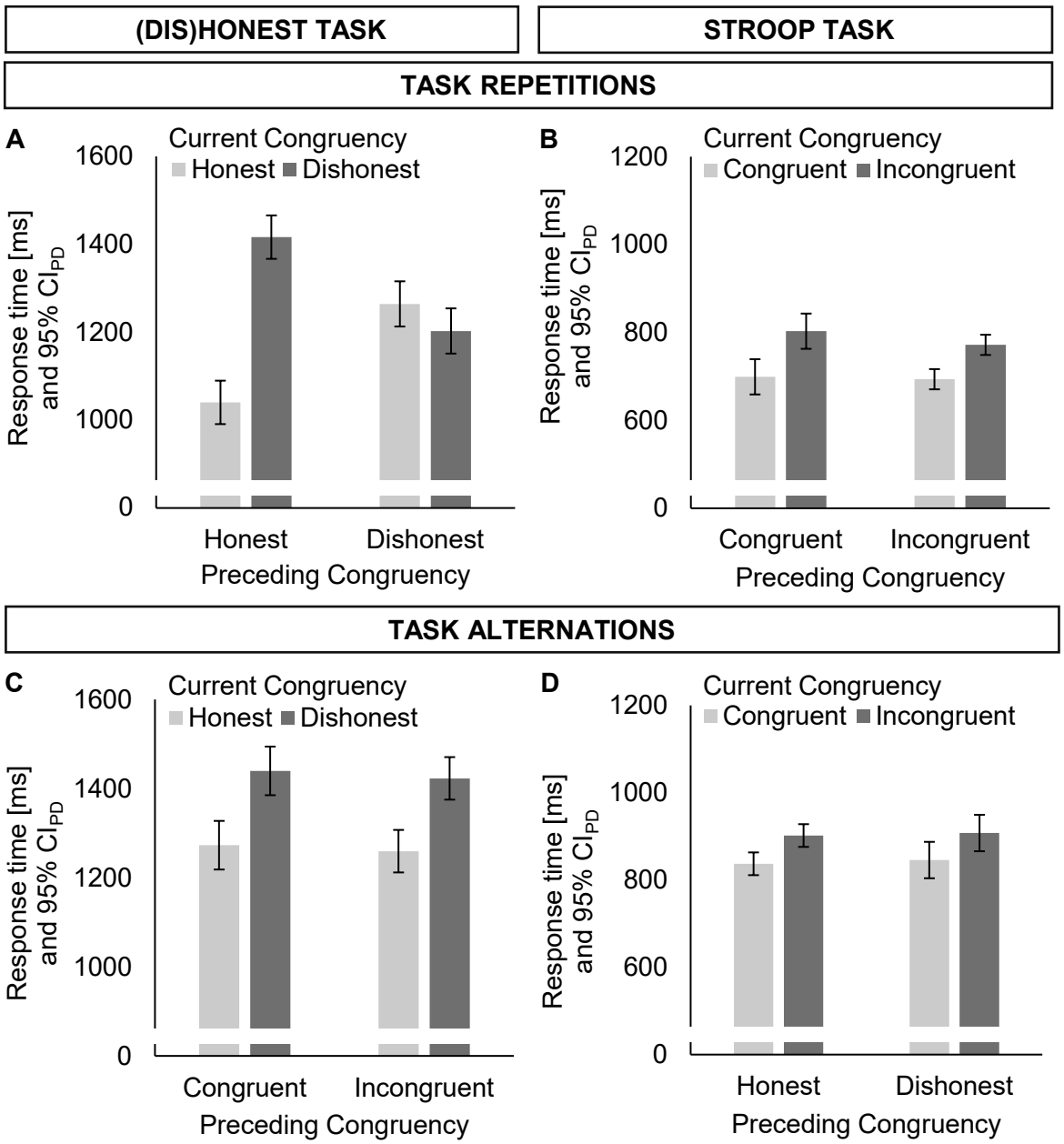

Figure 11 | Transient conflict adaptation effects on RTs in Experiment 6, plotted as function of current and preceding congruency, separately for task repetitions (upper panels, A and B) and task alternations (lower panels, C and D) for the (dis)honest task (left panels, A and C) and the Stroop task (right panels, B and D). Note that the RT plots are scaled differently. RTs also only showed a modulation by current and preceding congruency when the (dis)honest task was repeated (A). After honest responses, dishonest responses were slower whereas after dishonest responses honest responses were slower. Error bars represent the 95% CI_{PD}, computed separately for the conditions of preceding congruency and task sequence in each task.

A separate statistical assessment of the RTs of the two tasks scrutinized the former interactions and again, we only report those effects here that are informative to understand the former interactions. Current and preceding congruency did not interact in Stroop trials, $F(1, 28) = 1.37$, $p = .251$, $\eta_p^2 = .05$, BF = 2.73, and this interaction was also not qualified by task sequence, $F < 1$, BF = 3.29. In (dis)honest trials, however, this two-way interaction, $F(1, 28) = 149.85$, $p < .001$, $\eta_p^2 = .84$, BF < 0.01, and the three-way interaction of task sequence, current and preceding congruency, $F(1, 28) = 167.04$, $p < .001$, $\eta_p^2 = .86$, BF < 0.01, were significant. Separate ANOVAs for task repetitions and task switches in dishonest trials showed that preceding congruency did not affect current congruency when tasks switched from the preceding to the current trial, $F < 1$, BF = 5.02. The two-way interaction was significant for task repetitions, $F(1, 28) = 295.20$, $p < .001$, $\eta_p^2 = .91$, BF < 0.01, as dishonest responses were slower than honest responses after honest responding, $t(28) = 15.59$, $p < .001$, $d_z = 2.89$, BF < 0.01, but an opposite effect was evident after dishonest responding, $t(28) = 2.46$, $p = .021$, $d_z = 0.46$, BF = 0.40.

4.3.3. Discussion

Experiment 6 combined dishonest and Stroop conflict to explore the potential transfer of transient and sustained adaptation processes and, thus, the specificity of conflict adaptation in lying. That is, the experiment was set up to show whether the difference between honest and dishonest responding would be smaller in a highly incongruent Stroop environment, or directly after an incongruent Stroop trial, compared to frequent or recent congruent Stroop conditions. Similarly, the experiment examined whether the congruency effect in the Stroop task would be modulated by recent dishonesty. Even though transfer of transient and sustained control adaptation between such different tasks is possible (e.g., Hazeltine et al., 2011; Kleiman et al., 2014; Wirth, Pfister, & Kunde, 2016; Wühr et al., 2015), there was no transfer of transient control adaptation between both tasks in either direction and also no transfer of sustained effects from the Stroop task to the (dis)honest task.

It is important to note here, that Experiment 6 used separate response keys for the (dis)honest and the Stroop tasks. In contrast, studies that found a transfer of transient control adaptation between different tasks with distinct conflict sources used same response keys for those tasks (Hazeltine et al., 2011; Kleiman et al., 2014). Using same

response keys could lead to less salient task boundaries and thus to an enhanced probability of transferring control adaptation (Hazeltine et al., 2011). A transfer of sustained effects, however, was also observed when tasks, conflict sources, and responses differed like in the present experiment (Experiment 3 of Wühr et al., 2015).

Like in Experiment 5, transient adaptation was again found within the (dis)honest task, that is, dishonest responding was slower and less accurate than honest responding after an honest response but this effect was absent or even reversed after dishonest responses (cf. Debey et al., 2014; Foerster, Wirth, Kunde et al., 2017). In contrast, transient adaptation processes did not emerge in the Stroop task. It is difficult to interpret that null effect. Increased error rates and RTs in the (dis)honest compared to the Stroop task indicate that the former task execution was more difficult than the latter. This might have increased saliency of the (dis)honest task with prioritization of control adaptation in this task. Note also, that sustained effects within the Stroop task do not necessarily derive from control adaptation but could stem from learning frequent S–R pairings in the low conflict proportion condition.

In a nutshell, there was no evidence of a transient or sustained transfer of control adaptation from the Stroop task to the (dis)honest task or of a transient transfer in the opposite direction in the present experiment. However, this does not preclude the existence of general adaptation processes. The tasks of Experiment 6 featured different relevant stimulus dimensions, task rules, sources of conflict and responses. In Experiment 7, we, therefore, replaced the Stroop task with an S–R compatibility task that is more similar to dishonest responding (e.g., Davidson, Amso, Anderson, & Diamond, 2006) to render transfer of adaptation settings more likely. Most importantly, as in dishonest responding, the conflicting information in the conflict task was now relevant for task execution.

4.4 Experiment 7

Experiment 7 combined the (dis)honest task with a location task in which participants responded to the left and right location of a square with a response on the same or on the opposite side. Four color cues indicated same or opposite responses in each trial and the proportion of conflict in the location task varied between blocks. Here, the conflict source in the location task was very similar to the one in dishonest responding. In both cases,

conflict emerges from the automatic activation of the dominant response (truthful response vs. location-congruent response) and the required response that has to be derived from the dominant one. However, the dominant information in the location task is exogenously triggered by stimulus position, whereas it is endogenously triggered for dishonest responding, as the honest response is derived from question content and memory. In both cases, the conflict then emerges because an endogenous rule urges an opposing response.

A recent study suggests that both conflicts could share common adaptation mechanisms (Experiment 3 and 4 of Foerster, Wirth, Kunde et al., 2017). One group of participants in this study received honest and dishonest instructions to respond to simple *yes/no* questions while another group of participants received instructions that they should respond to questions from the perspective of two different agents. One of the agents supposedly shared the same experiences as the participants, whereas the other agent had opposite experiences. Notable, only the instructions but not the S–R rules of the experiment changed and we assumed that both, dishonest and opposite responses to follow a two-step process, where a dominant response has to be inhibited to allow for the opposite required response. In line with that assumption, transient effects were similar under both instructions. These results suggest that the (dis)honest and the location task could share adaptation mechanisms. In that case, preceding congruency should modulate the congruency effect when tasks switch, with a larger congruency effect in either task when the preceding other task was congruent relative to when it was incongruent. With a high proportion of congruent trials in the location task, the difference between honest and dishonest responding should be larger than with a low proportion of congruent trials.

4.4.1. Method

Participants │ A new sample of 32 participants (age: $M = 28.4$, $SD = 8.21$; 22 female; 31 right-handed) took part in the experiment and received either monetary compensation or course credit. All participants gave written informed consent. The statistical analyses are based on 29 participants because three participants did not provide enough data according to the same criteria as used in Experiment 5 and 6.

Apparatus and stimuli | For the sake of brevity, we only report details in which Experiment 7 deviated from Experiment 6. Instead of the Stroop task, Experiment 7 featured an S-R compatibility task where participants responded to the location of a square (location task). The edges of the square were 30 pixels long. The square appeared in one of four colors (blue, brown, yellow and purple) to control for feature integration effects. Participants were to respond according to the location and font color of the square with their right index and middle finger by pressing the keys *K* and *L.* Half of the participants were to respond in accordance with the square position (i.e., square left, left keypress; square right, right keypress) when the color of the square was blue or yellow (congruent). When the square was brown or purple, these participants responded with a left keypress to right squares and with a right keypress to left squares (incongruent). For the other half of participants, blue and yellow squares implied incongruent responding and brown and purple squares indicated congruent responding.

Procedure | Participants again started with a pre-experimental procedure to select an equal amount of questions that asked about activities that had been performed and activities that had not been performed on the same day. Like in Experiment 5, 10 questions with affirmative and 10 questions with negative answers were selected from the question pool.

The square appeared either on the left side or the right side of the display. As in (dis)honest trials, participants had to respond within 3000 ms. Both tasks followed a random sequence. Again, responding honestly and dishonestly to questions was practiced in the first block of 16 trials. Then, participants practiced each combination of square color and position twice in the second block of 16 trials.

Like in Experiment 6, each question appeared equally often with honest and dishonest responses within a block whereas the proportion of conflict in the location task varied between blocks. Each of the 20 blocks featured 80 trials, that is, 40 (dis)honest and 40 location task trials. The sequence of the conflict proportion conditions was counterbalanced across participants. Participants were allowed self-paced breaks between blocks.

4.4.2. Results

Data treatment | As in Experiment 5 and 6, trials were excluded from further analyses when they featured the same task as the preceding trials with also either same question (1.3%) or same square position and square color (6.5%). We only included correct trials or trials with commission errors, which also followed a correct trial for error rate analyses. This led to the exclusion of 7.8% of (dis)honest trials and 7.6% of location trials. For RT analyses, we excluded all error trials and those trials that followed them (15.8% of (dis)honest trials, 11.2% of location trials). For analyses on sustained and transient effects, 2.7% and 2.6% of (dis)honest trials and 2.8% and 3.0% of location trials, respectively, were identified as outliers and excluded.

Sustained effects | Error rates and RTs were analyzed in a 2 × 2 × 2 ANOVA with the within-subjects factors task ([dis]honest vs. location), location conflict proportion (low vs. high) and current congruency (congruent vs. incongruent). Sustained adaptation effects should produce a significant interaction between conflict proportion and current congruency. When such adaptation processes fully transferred from the location to the (dis)honest task, there should be no three-way interaction between all three factors. We scrutinized significant three-way interactions in separate 2 × 2 ANOVAs and two-way interactions in paired-samples *t*-tests and report BFs for these tests in the text.

Figure 12 shows the mean error rates (upper panels A and B) and RTs (lower panels C and D) for each combination of current congruency and conflict proportion for the (dis)honest task (left panels A and C) and the location task (right panels B and D). On average each cell included 160 observations.

The two-way interaction of current congruency and conflict proportion was significant, $F(1, 28) = 26.33$, $p < .001$, $\eta_p^2 = .49$, as was the three-way interaction of all factors, $F(1, 28) = 30.94$, $p < .001$, $\eta_p^2 = .53$. Furthermore, errors were more frequent in the (dis)honest than in the location task, $F(1, 28) = 15.15$, $p = .001$, $\eta_p^2 = .35$, and in incongruent/dishonest compared to congruent/ honest trials, $F(1, 28) = 4.80$, $p = .037$, $\eta_p^2 = .15$. There was also a significant two-way interaction of task and current congruency, $F(1, 28) = 15.24$, $p = .001$, $\eta_p^2 = .35$. None of the remaining effects were significant (Fs < 1.49, ps > .232).

Separate analyses for the two tasks clarified the interactions. The main effect of current congruency, $F(1, 28) = 14.15$, $p = .001$, $\eta_p^2 = .34$, BF = 0.02, but not the two-way

interaction of current congruency and conflict proportion, $F(1, 28) = 2.67$, $p = .114$, $\eta_p^2 = .09$, BF = 1.55, was significant for the (dis)honest task. For the location task, the main effect of current congruency was not significant ($F < 1$, BF = 3.47), but the interaction of current congruency and conflict proportion was, $F(1, 28) = 37.40$, $p < .001$, $\eta_p^2 = .57$, BF < 0.01, as participants committed more errors in incongruent than in congruent trials in low conflict proportion blocks, $t(28) = 3.50$, $p = .002$, $d_z = 0.65$, BF = 0.04, but showed the opposite pattern of results in high conflict proportion blocks, $t(28) = 5.55$, $p < .001$, $d_z = 1.03$, BF < 0.01.

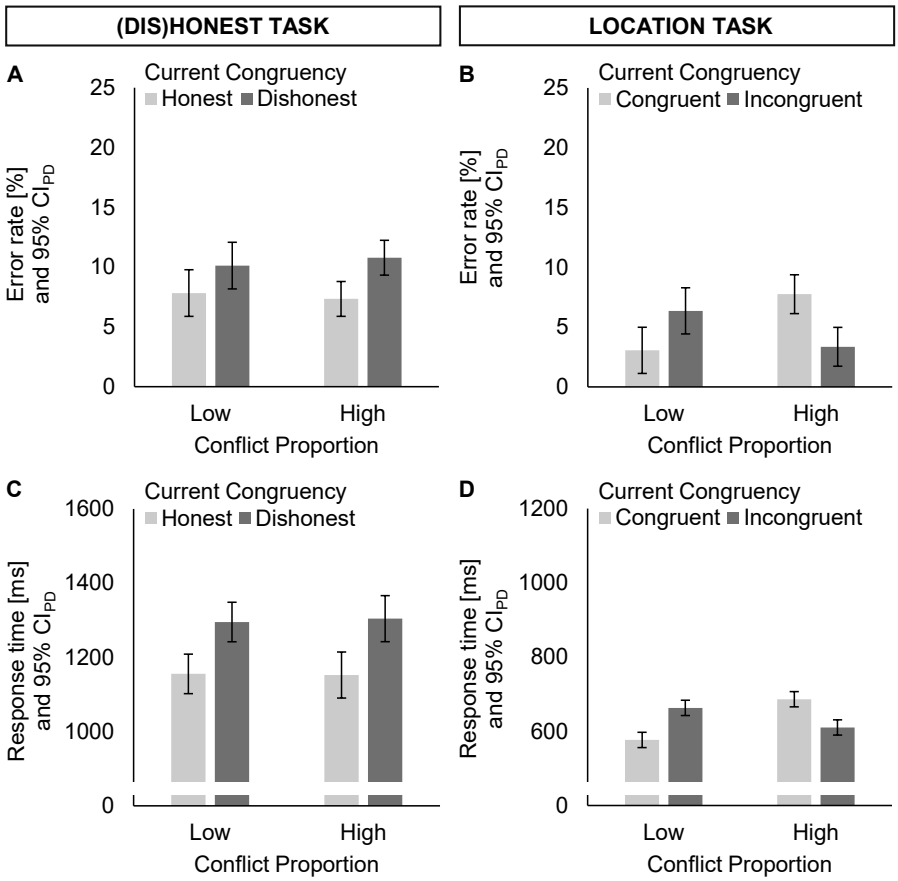

Figure 12 | Sustained conflict adaptation effects on error rates (upper panels, A and B) and RTs (lower panels, C and D) in Experiment 7, plotted as function of current congruency and conflict proportion for the (dis)honest task (left panels, A and C) and the location task (right panels, B and D). Note that the RT plots are scaled differently. Dishonest responses were slower and less accurate than honest responses. The proportion manipulation of the location task did not affect performance in the (dis)honest task. Incongruent responses in the location task were slower and less accurate than congruent responses in low conflict proportion contexts but a reversed effect emerged for high conflict proportion blocks. Location responses were also slower in the high conflict proportion condition. Error bars represent the 95% CI_{PD}, computed separately for low and high conflict proportion in each task.

The RT results were in line with the error rate results. Again, the two-way interaction between current congruency and conflict proportion, $F(1, 28) = 70.94$, $p < .001$, $\eta_p^2 = .72$, and the three-way interaction, $F(1, 28) = 78.07$, $p < .001$, $\eta_p^2 = .74$, were significant. Furthermore, responses were considerably slower in the (dis)honest task than in the location task, $F(1, 28) = 740.30$, $p < .001$, $\eta_p^2 = .96$, and in incongruent/dishonest trials compared to congruent/honest trials, $F(1, 28) = 30.65$, $p < .001$, $\eta_p^2 = .52$. The two-way interaction between task and current congruency, $F(1, 28) = 22.62$, $p < .001$, $\eta_p^2 = .45$, was also significant. None of the remaining effects were significant (Fs < 1.45, ps > .239).

Separate ANOVAs for the two tasks showed that dishonest responses were slower than honest responses, $F(1, 28) = 28.65$, $p < .001$, $\eta_p^2 = .51$, BF < 0.01, but the main effect of conflict proportion ($F < 1$, BF = 5.04), and the interaction of current congruency and conflict proportion were not significant in the (dis)honest task ($F < 1$, BF = 3.75). Responses in the location task did not differ with current congruency ($F < 1$, BF = 4.22), but with proportion, $F(1, 28) = 5.86$, $p = .022$, $\eta_p^2 = .17$, BF = 0.43, as responding took longer in high compared to low conflict proportion blocks. There was also a significant interaction between both factors, $F(1, 28) = 194.41$, $p < .001$, $\eta_p^2 = .87$, BF < 0.01, indicating that incongruent responses were slower than congruent responses with low conflict proportion, $t(28) = 8.58$, $p < .001$, $d_z = 1.59$, BF < 0.01, while a reversed congruency effect was evident with high conflict proportion, $t(28) = 7.58$, $p < .001$, $d_z = 1.41$, BF < 0.01.

Transient effects │ A 2 × 2 × 2 × 2 ANOVA with the within-subjects factors task ([dis]honest vs. location), task sequence (repetition vs. switch), current congruency (honest/congruent vs. dishonest/ incongruent) and preceding congruency was conducted on error rates and RTs. Transient adaptation effects should produce a significant interaction between current and preceding congruency. When such adaptation processes transferred between tasks, this two-way interaction should not be further qualified by task sequence. We scrutinized significant three-way and four-way interactions in separate planned ANOVAs and significant two-way interactions in planned paired-samples *t*-tests and report BFs for these tests in the text.

Figure 13 shows the mean error rates and Figure 14 the mean RTs for each combination of current and preceding congruency for task repetitions (upper panels A and B) and task alternations (lower panels C and D) in the (dis)honest task (left panels A and C) and the location task (right panels B and D). On average each cell included 80 observations.

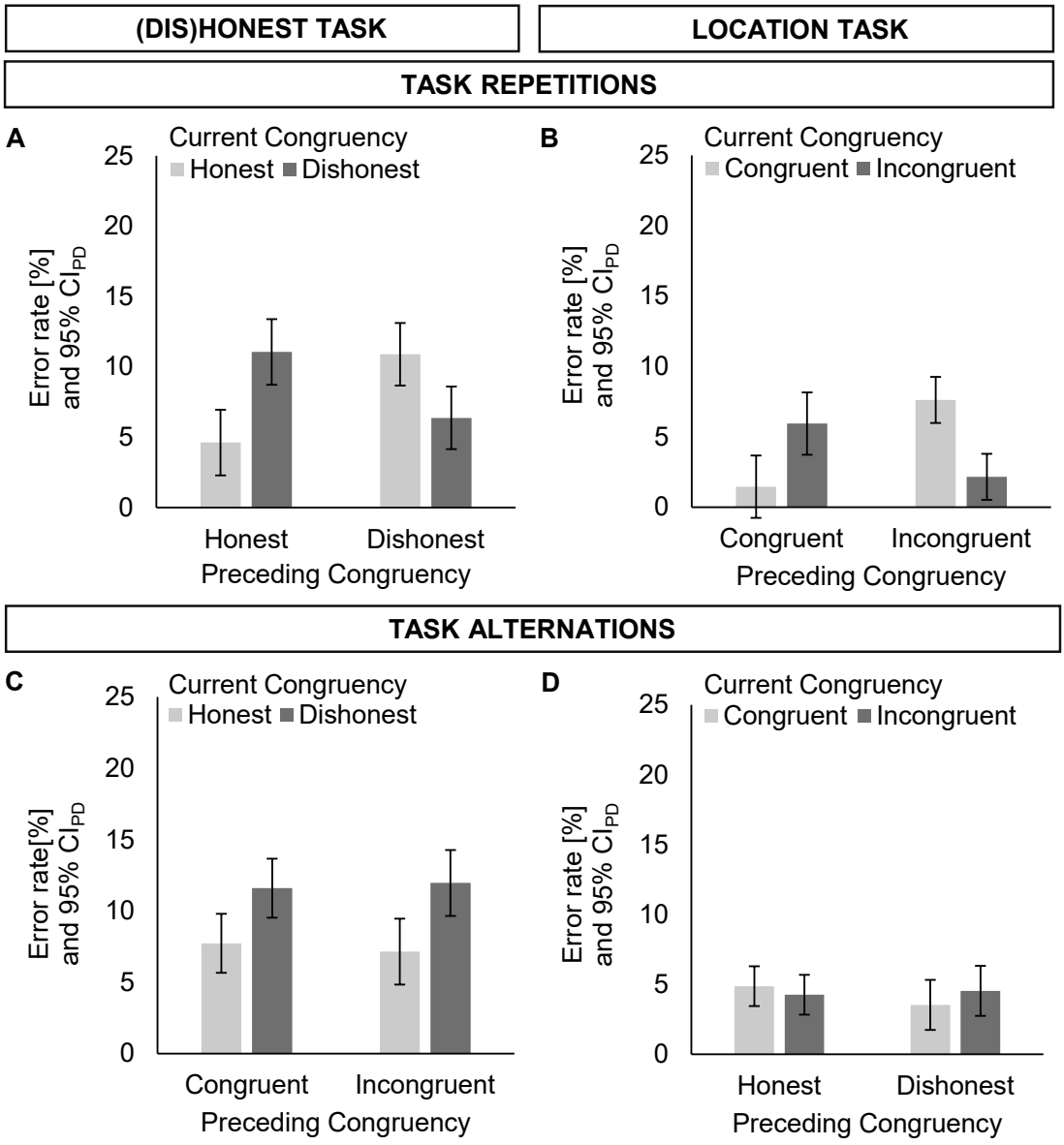

Figure 13 | Transient conflict adaptation effects on error rates in Experiment 7, plotted as function of current and preceding congruency, separately for task repetitions (upper panels, A and B) and task alternations (lower panels, C and D) for the (dis)honest task (left panels, A and C) and the location task (right panels, B and D). Participants committed more errors in the (dis)honest task when tasks switched than when tasks repeated. In the case of task repetitions, there were transient adaptation effects as congruency effects were reversed after dishonest (A) and incongruent trials (B). Responses were also less accurate after (dis)honest/incongruent trials when tasks repeated. A congruency effect was evident in task alternations (C and D). Error bars represent the 95% CI_{PD}, computed separately for the conditions of preceding congruency and task sequence in each task.

Error rate analyses returned a bundle of significant interactions that we will cluster into two convenient groups with the first bundle capturing interactions without the factor preceding congruency and the second bundle comprising all interactions including this factor. First, the interactions between task and task sequence, $F(1, 28) = 5.70$, $p = .024$, $\eta_p^2 = .17$, between task and current congruency, $F(1, 28) = 9.83$, $p = .004$, $\eta_p^2 = .26$, between task sequence and current congruency, $F(1, 28) = 14.68$, $p = .001$, $\eta_p^2 = .34$, and between task, task sequence and current congruency, $F(1, 28) = 5.88$, $p = .022$, $\eta_p^2 = .17$, were significant. Second, the interactions between task sequence and preceding congruency, $F(1, 28) = 4.09$, $p = .053$, $\eta_p^2 = .13$, between current congruency and preceding congruency, $F(1, 28) = 50.97$, $p < .001$, $\eta_p^2 = .65$, and between task sequence, current congruency and preceding congruency, $F(1, 28) = 60.29$, $p < .001$, $\eta_p^2 = .68$, were also significant. In accordance with the analysis of sustained effects, the main effects of task, $F(1, 28) = 23.57$, $p < .001$, $\eta_p^2 = .46$, and current congruency, $F(1, 28) = 8.40$, $p = .007$, $\eta_p^2 = .23$, were significant. Furthermore, there was a nonsignificant trend toward more errors when tasks switched than when tasks repeated, $F(1, 28) = 3.77$, $p = .062$, $\eta_p^2 = .12$. None of the remaining effects were significant (Fs < 1.41, ps > .244).

For the first bundle of significant interactions, we decided to average data over preceding congruency and computed separate 2 × 2 ANOVAs for the two tasks with the factors task sequence and current congruency. Neither the main effects, nor the interaction were significant in the location task (Fs < 1.07, ps > .309, BFs >3.11), whereas all of them were significant in the (dis)honest task. In the (dis)honest task, participants committed more errors when tasks switched than when tasks repeated, $F(1, 28) = 7.56$, $p = .010$, $\eta_p^2 = .21$, BF = 0.23, and when they gave dishonest compared to honest responses, $F(1, 28) = 12.96$, $p = .001$, $\eta_p^2 = .32$, BF = 0.04. These main effects were qualified by their significant two-way interaction, $F(1, 28) = 15.45$, $p = .001$, $\eta_p^2 = .36$, BF = 0.02, as the difference between honest and dishonest responding was only evident when tasks switched, $t(28) = 4.50$, $p < .001$, $d_z = 0.84$, BF < 0.01, but not when tasks repeated, $t(28) = 1.33$, $p = .194$, $d_z = 0.25$, BF = 2.29 (note the sequential modulation of the congruency effect in the following analyses though). For the second bundle of significant interactions, we decided to average data over the two tasks and then computed separate 2 × 2 ANOVAs for task repetitions and switches with the factors current and preceding congruency. Task repetitions did not show a main effect of current congruency,

$F < 1$, BF = 4.43, but of previous congruency, $F(1, 28) = 4.44$, $p = .044$, $\eta_p^2 = .14$, BF = 0.75, as responses were more error-prone after incongruent/dishonest than after congruent/honest trials. However, the interaction of current and preceding congruency was also significant, $F(1, 28) = 72.53$, $p < .001$, $\eta_p^2 = .72$, BF < 0.01. A typical congruency effect emerged after congruent/honest responding, $t(28) = 6.21$, $p < .001$, $d_z = 1.15$, BF < 0.01, but a reversed effect was evident after incongruent/dishonest responding $t(28) = 8.24$, $p < .001$, $d_z = 1.53$, BF < 0.01. When tasks switched, a congruency effect was evident, $F(1, 28) = 16.09$, $p < .001$, $\eta_p^2 = .37$, BF = 0.01, whereas preceding congruency did not affect error rates, $F < 1$, BF = 3.69. There was a nonsignificant trend toward a two-way interaction of both factors, $F(1, 28) = 3.57$, $p = .069$, $\eta_p^2 = .11$, BF = 1.07, pointing toward a smaller congruency effect after congruent/honest than after incongruent/dishonest responding.

In RTs, there were significant interactions between task and task sequence, $F(1, 28) = 42.32$, $p < .001$, $\eta_p^2 = .60$, task sequence and current congruency, $F(1, 28) = 8.62$, $p = .007$, $\eta_p^2 = .24$, current and preceding congruency, $F(1, 28) = 205.72$, $p < .001$, $\eta_p^2 = .88$, task, current and preceding congruency, $F(1, 28) = 11.12$, $p = .002$, $\eta_p^2 = .28$, and task sequence, current and preceding congruency, $F(1, 28) = 258.08$, $p < .001$, $\eta_p^2 = .90$. There was also a nonsignificant trend toward a three-way interaction of task, task sequence, and current congruency, $F(1, 28) = 3.56$, $p = .070$, $\eta_p^2 = .11$. Finally, the analysis yielded a significant four-way interaction of all factors, $F(1, 28) = 27.27$, $p < .001$, $\eta_p^2 = .49$. Mirroring the analysis on sustained effects, the main effect of task, $F(1, 28) = 807.22$, $p < .001$, $\eta_p^2 = .97$, current congruency, $F(1, 28) = 35.18$, $p < .001$, $\eta_p^2 = .56$, and the two-way interaction of both factors were significant, $F(1, 28) = 23.20$, $p < .001$, $\eta_p^2 = .45$. Furthermore, task switches took longer than task repetitions, $F(1, 28) = 82.52$, $p < .001$, $\eta_p^2 = .75$, and there was a nonsignificant trend toward a main effect of preceding congruency, $F(1, 28) = 3.77$, $p = .062$, $\eta_p^2 = .12$. None of the remaining effects were significant (Fs < 1.55, ps > .224).

For a better understanding of the data, separate 2 × 2 × 2 ANOVAs for the two tasks with the factors task sequence, current and preceding congruency were computed. As both ANOVAs returned significant three-way interactions of all factors ([dis]honest task: $F(1, 28) = 139.63$, $p < .001$, $\eta_p^2 = .83$, BF < 0.01; location task: $F(1, 28) = 275.21$, $p < .001$, $\eta_p^2 = .91$, BF < 0.01), separate 2 × 2 ANOVAs with the factors current and preceding congruency were computed for task repetitions and switches for each task, respectively.

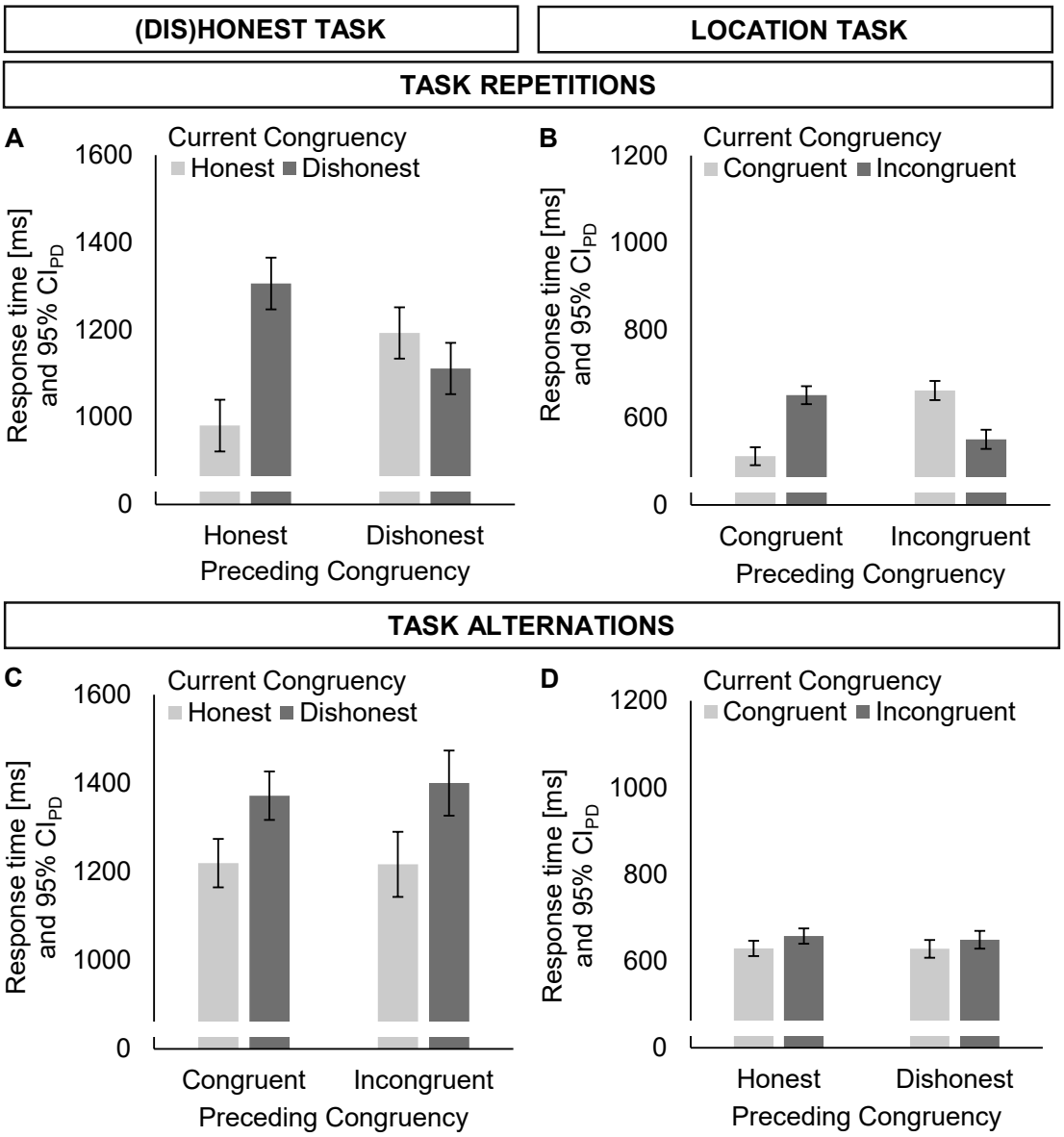

Figure 14 | Transient conflict adaptation effects on RTs in Experiment 7, plotted as function of current and preceding congruency, separately for task repetitions (upper panels, A and B) and task alternations (lower panels, C and D) for the (dis)honest task (left panels, A and C) and the location task (right panels, B and D). Note that the RT plots are scaled differently. Current and preceding congruency interacted when tasks repeated as the congruency effect was in the expected direction after honest/congruent trials but reversed after dishonest/incongruent trials (A and B). This modulation was stronger in the location task. Furthermore, increased RTs emerged in repetition trials of the location task after incongruent than after congruent responses. When tasks alternated, preceding congruency did not affect congruency effects of both tasks. Error bars represent the 95% CI_{PD}, computed separately for the conditions of preceding congruency and task sequence in each task.

Responding in task repetitions of the (dis)honest task was slower for dishonest than for honest trials, $F(1, 28) = 25.11$, $p < .001$, $\eta_p^2 = 47$, BF < 0.01. Preceding congruency did not affect RTs, $F < 1$, BF = 3.3, but the interaction of both factors was significant, $F(1,$

28) = 174.38, $p < .001$, $\eta_p^2 = .86$, BF < 0.01, because of a typical congruency effect after honest trials, $t(28) = 11.25$, $p < .001$, $d_z = 2.09$, BF < 0.01, but a reversed effect after dishonest trials, $t(28) = 2.84$, $p = .008$, $d_z = 0.53$, BF = 0.19.

Task switches of the (dis)honest task showed prolonged dishonest responses in comparison to honest responses, $F(1, 28) = 31.09$, $p < .001$, $\eta_p^2 = .53$, BF < 0.01. The main effect of preceding congruency, $F < 1$, BF = 4.12, and the two-way interaction were not significant, $F(1, 28) = 2.60$, $p = .118$, $\eta_p^2 = .09$, BF = 1.60. Task repetitions of the location task showed a nonsignificant trend toward longer incongruent than congruent responses, $F(1, 28) = 3.96$, $p = .056$, $\eta_p^2 = .12$, BF = 0.91, and responses took longer after incongruent than after congruent responding, $F(1, 28) = 15.32$, $p = .001$, $\eta_p^2 = .35$, BF = 0.02. These main effects were further qualified by a significant interaction, $F(1, 28) = 271.90$, $p < .001$, $\eta_p^2 = .91$, BF < 0.01, as incongruent responses were slower than congruent responses after congruent trials, $t(28) = 13.92$, $p < .001$, $d_z = 2.59$, BF < 0.01, but the opposite was evident after incongruent trials, $t(28) = 10.44$, $p < .001$, $d_z = 1.94$, BF < 0.01.

Task alternations of the location task showed longer RTs in incongruent than in congruent trials, $F(1, 28) = 7.78$, $p = .009$, $\eta_p^2 = .22$, BF = 0.21. The main effect of preceding congruency, $F(1, 28) = 2.39$, $p = .133$, $\eta_p^2 = .08$, BF = 1.75, and the two-way interaction of both factors were not significant, $F(1, 28) = 1.61$, $p = .215$, $\eta_p^2 = .05$, BF = 2.45.

4.4.3. Discussion

Experiment 7 examined transfer effects between dishonest responding on the one hand, and cognitive conflict that likely is due to similar mechanisms as in dishonesty on the other hand. Both conflicts in this experiment are based on the parallel activation of a dominant response and its appropriate counterpart. Still, the proportion of conflict trials in the location task affected only the location task but not the (dis)honest task. Accordingly, there was no transfer of sustained control adaptation. Transient adaptation effects emerged within both tasks but not between tasks. There was a surprising trend in error rates toward a modulation of the congruency effect by preceding congruency when tasks switched with smaller congruency effects after honest/congruent responses. This modulation was not significant though and did not replicate in RTs. As in Experiment 6,

participants responded with different response keys in both tasks which could render transfer of transient control adaptation less likely (Hazeltine et al., 2011; Kleiman et al., 2014) but does not seem to prevent transfer of sustained control adaptation (Experiment 3 of Wühr et al., 2015). Together, Experiment 6 and Experiment 7 observed specific control adaptation in lying that is only triggered by (dis)honest responses but not by other conflicts.

4.5 General discussion

Three experiments examined how dishonest responding changes with different conflicting contexts. Experiment 5 provided an integrated analysis of how dishonest responding adapts to transient and sustained dishonest contexts, that is, how dishonest responding changes when it succeeds an honest or a dishonest response and how it changes with the overall proportion of dishonest responses. Experiment 6 and Experiment 7 introduced two other types of conflict to establish whether control demands induced by another conflict task can generalize to dishonest responding or vice versa. Overall, the present results suggest that transient and sustained adaptation operate independently within dishonest responding and that transient adaptation effects could be mistaken for sustained adaptation effects (Experiment 5). Furthermore, we replicated strong transient effects in all experiments while neither transient nor sustained adaptation effects transferred across conflicts (Experiment 6 and Experiment 7). Together, the experiments paint a consistent picture of a surprisingly focused scope of the context that elicits control adaptation in dishonest responding.

4.5.1. Trading off sustained and transient adaptation

Only transient adaptation to recent (dis)honesty but not sustained adaptation to the proportion of dishonesty emerged in the present analyses. In particular, dishonest responding was more difficult than honest responding after recent honest responses. However, after recent dishonest responses, honest responding was more difficult than dishonest responding even though this reversed effect did not have the same magnitude as the former effect, rendering dishonest responding still more difficult than honest responding overall. As such, the current findings corroborate the robust finding that dishonest responding is more difficult than honest responding (e.g., Suchotzki et al., 2017)

but that the processing of preceding responses can shift that difficulty markedly (Debey, Liefooghe et al., 2015; Foerster, Wirth, Kunde et al., 2017).

Preceding studies did not allow for disentangling whether effects of proportion dishonest manipulations reflected truly sustained adaptation processes or whether they instead reflected transient adaptation processes because information about preceding (dis)honesty was not included in statistical analyses (Van Bockstaele et al., 2012; Verschuere et al., 2011). The present experiment suggests that transient influences drive control adaptation even though, in theory, transient and sustained processes could operate at the same time (e.g., Botvinick et al., 2001).

Of course, these findings should not be taken as evidence against sustained adaptation to dishonesty in general, because such effects emerged when transient adaptation was held constant (Van Bockstaele et al., 2015). In this setting, smaller differences between honest and dishonest responding emerged in error rates but not in RTs when lying was frequent relative to when it was rare. In the present experiment, by contrast, the proportion manipulation did not even modulate the congruency effect in error rates when transient influences were not included in the analysis. The most important difference between both experiments is the fixed versus random arrangement of trials and the resulting constant versus varying nature of transient influences. The predictable trial sequences in Van Bockstaele et al. (2015) accentuated the proportion manipulation as they featured sequences of 10 honest/dishonest trials, followed by 10 trials with honesty and dishonesty in alternation. This accentuation could have rendered sustained control adaptation processes more likely. In contrast, the unpredictable sequence of the present experiment could have worked in favor of transient adaptation as the proportion manipulation might have been less salient. Adaptation to recent events also might have been easier than adaptation to a larger context in this situation. The sustained adaptation effect with predictable trial sequences was of medium size and transient adaptations effects here and elsewhere were of large size (Debey, Liefooghe et al., 2015; Foerster, Wirth, Kunde et al., 2017). Possibly, a trade-off favors transient control adaptation over sustained control adaptation when both mechanisms could affect dishonest responding, whereas sustained influences are only considered when transient adaptation cannot come into action. The first hint for complex flexible trade-off mechanisms in the current data comes from the explanatory analyses of Experiment 5, which showed stronger transient

adaptation in mostly dishonest than mostly honest contexts but only when participants started in the mostly dishonest context.

The present finding of absent sustained adaptation effects in the presence of transient adaptation effects taps into an important issue. With standard conflict tasks, it is difficult to simultaneously scrutinize sustained and transient control adaptation processes within a conflict without introducing confounding variables. For one, proportion manipulations should not introduce frequent S-R pairings which could produce a modulation of the congruency effect by bottom-up learning mechanisms (e.g., Wühr et al., 2015). Thus, inducer and probe stimuli are necessary like in the present study. Second, transitions from congruent/incongruent trials to subsequent congruent/incongruent trials should not be confounded with certain transitions of stimuli and responses between both trials (e.g., Hommel, 2004; Notebaert, Gevers, Verbruggen, & Liefooghe, 2006). In particular, it has to be ensured that all transitions of congruency are similar in regard to stimuli repetitions and switches. Implementing inducer and probe stimuli while at the same time controlling for stimuli repetitions/switches is difficult as those standard conflict tasks typically have a small stimulus set while it is easier with large stimulus sets as in typical (dis)honest tasks. Accordingly, it is more convenient to target common mechanisms of transient and sustained adaptation for standard conflicts via transfer between tasks. If transient and sustained effects dissociate when it comes to transfer, underlying mechanisms should differ as well (e.g., Wühr et al., 2015). Although this transfer situation is elegant, it could miss out on a possible trade-off mechanism between transient and sustained adaptation processes and the moderators of such a trade-off (e.g., Bugg, 2014).

As mentioned earlier, the conflict in dishonest responding differs markedly from standard conflicts, which could also set different rules for transient and sustained adaptation processes. While it is theoretically possible to inhibit the irrelevant stimulus dimension completely in standard conflict tasks and still give a correct response, the dominant truthful response seems to be a prerequisite to generate an unrehearsed dishonest response (e.g., Debey et al., 2014). For rehearsed lies, by contrast, dishonest responses seem to be retrieved directly from stimuli (e.g., Hu, Chen et al., 2012; Walczyk, Mahoney, Doverspike, & Griffith-Ross, 2009, 2012). For unrehearsed lies, such as in the present experiments, inhibition during dishonest responding might follow a specific time-course and, thus, only be beneficial if initiated early enough, but not too early. Such a

time-critical process could be so demanding that agents are not able to additionally consider sustained information about dishonest responding. In a nutshell, it is not obvious from the current point of view whether dishonest responding truly differs from other conflicts in regard to sustained adaptation processes or whether the diverging evidence is based on methodological discrepancies.

A hint that a trade-off could exist for standard conflicts comes from a study where Simon and Stroop conflicts with the same relevant stimulus dimension were presented randomly (Experiment 2 of Torres-Quesada et al., 2014). When the proportion of Stroop conflict was manipulated, sustained effects were also found for the Simon effect. However, such a transfer of sustained adaptation was only evident when the conflict source had switched from the preceding to the current trial (Stroop → Simon). This is important as transient effects emerged for task repetitions but not for task switches. So when transient conflict adaptation in the Simon task took place after conflict experience in a preceding Simon trial, sustained adaptation was absent, but when there was no conflict adaptation in the Simon task after a Stroop task, sustained adaptation effects were present. In the same study, the authors wanted to assess whether the level of the proportion manipulation (e.g., 80/20 vs. 60/40) affected proportion effects. It did, but specifically for the conflict that featured the proportion manipulation. As such, the interaction of the level of proportion and the effect of proportion on congruency could stem from bottom-up learning mechanisms of probable associations of the irrelevant stimulus dimension and the response. With inducer and probe questions in a dishonest task, such bottom-up mechanisms can be controlled easily. It is plausible that a more extreme manipulation of the proportion shifts the trade-off toward sustained control adaptation.

To scrutinize a possible trade-off between transient and sustained adaptation, researchers could set up an experiment where a dishonest task (or any other conflict with a sufficiently large stimulus set) comes with a different temporal interval in between two successive responses. Previous evidence shows that despite an intermitting task and relatively long time intervals between (dis)honest responses, transient adaptation effects within dishonest responding are still evident (albeit reduced; Experiment 2 and 3 of Foerster, Wirth, Kunde et al., 2017). In addition, the proportion of dishonest trials should be manipulated across blocks via inducer questions (possibly by also varying the extent of the proportion manipulation, cf. Experiment 2 of Torres-Quesada et al., 2014). Such

experiments could reveal if sustained influences increase with decreasing transient influences and/or with increasing extremity and if both control adaptation processes can also appear in parallel. This is not only relevant for dishonest responding but also for other behavioral conflicts and our general understanding of cognitive control processes. Furthermore, this insight could also prove valuable considering applied efforts to render honest responding as easy as possible and lying as hard as possible to improve lie detection methods (Van Bockstaele et al., 2012, 2015; Verschuere et al., 2011).

4.5.2. Specific adaptation for dishonesty

The present study yielded no signs of transfer of transient control adaptation in either direction between the conflicts in dishonest responding and other tasks, nor did sustained adaptation transfer from other conflict tasks to dishonest responding. Whereas conflict sources highly differed in Experiment 6, they were similar in Experiment 7 and previous evidence showed a transfer of transient and sustained control adaptation between very different conflict sources and with different relevant stimulus dimensions (e.g., Kleiman et al., 2014; Wühr et al., 2015).

Another conflict that appears to be very similar to dishonest responding in its basic processing is a rule violation. When agents break a rule, the dominant rule-based response has to be inhibited and this rule-based response is necessary to derive the rule violation (Jusyte et al., 2017; Pfister, Wirth, Schwarz, Foerster et al., 2016; Pfister, Wirth, Schwarz, Steinhauser et al., 2016). Rule violation instructions in the present Experiment 7 would result in exactly the same S-R rules and correct actions. However, preceding evidence suggests that the explicit instruction of rule violations would produce larger congruency effects but similar adaptation processes than more neutral rule inversion instructions (Experiment 3 of Wirth, Pfister, Foerster et al., 2016). The current study makes the counterintuitive suggestion that despite both tasks being similar, control adaptation should not transfer between dishonest responding and rule violations. However, a possibly shared negative connotation of both behavioral tendencies could promote transfer (Wirth, Foerster et al., 2018). A close examination of common mechanisms underlying dishonest and violation behavior should be the aim of future research.

The present comparison of different cognitive conflicts and dishonest responding also highlights critical methodological issues. For example, as mentioned earlier, the

examination of dishonesty comes with a considerably larger stimulus set than standard conflicts. In the present location task, participants responded to two locations depending on four colors with two responses. In the dishonest task, participants responded to 20 questions depending on two colors with two responses. A large stimulus set is a prerequisite to make sure that participants do not simply learn S-R associations during dishonest responding (e.g., question A in color B affords response C). Of course, simple responding based on S-R associations is an aspect of dishonest responding as specific dishonest responses can be learned when used frequently (e.g., Hu, Chen et al., 2012; Walczyk et al., 2009, 2012), however, as people do not lie as frequently as they tell the truth, dominance of the truthful response should be the default scenario (e.g., Debey, De Schryver et al., 2015; DePaulo et al., 1996; Halevy et al., 2014; Hilbig & Hessler, 2013; Serota et al., 2010). Besides the larger stimulus set, the current question stimuli also differ in complexity from those employed in most common conflict tasks (including the current Stroop and location task). To the best of our knowledge, there are no theoretical assumptions or empirical evidence on how stimulus set size or stimulus complexity could interact with control adaptation or its transfer. The current evidence in the literature does not provide clear rules, which differences between stimuli, conflict sources, and responses or their constellation are negligible and which have the power to eliminate any transfer of control adaptation. In the light of the large transient adaptation effects that are found within dishonest responding, it is safe to conclude that similar adaptation effects between dishonesty and other tasks do not seem to emerge easily. This does not, however, preclude that there could be specific situations that set the right conditions for such a transfer.

4.5.3. Transient effects: The role of conflict adaptation and task switching

Whereas the current study examined control processes in dishonesty from the perspective of conflict adaptation, recent studies focused on the involvement of task switching processes (Debey, Liefooghe et al., 2015; Foerster, Wirth, Kunde et al., 2017). Both of these theoretical perspectives converge on the notion that the difficulty of dishonest responding can vary due to transient factors. This raises the question of whether or not the transient changes explored in the present experiments might be fully explained in terms of task-switching mechanisms. This does not seem to be the case though, as suggested by several observations. First and foremost, task switching accounts would

predict asymmetric switch costs with larger switch costs for the transition from a difficult to an easier task as for the transition from a relatively easy to a more difficult task (e.g., Allport, Styles, & Hsieh, 1994). Executing a less dominant task requires enhanced activation of the relevant task set, and enhanced inhibition of the irrelevant but more dominant task set, rendering a subsequent switch to the dominant task especially effortful (e.g., Koch, Prinz, & Allport, 2005; Leboe, Whittlesea, & Milliken, 2005; Schneider & Anderson, 2010). This view makes direct predictions for the analysis of transient effects for honest and dishonest responses because responding honestly is dominant and easier than dishonest responding (e.g., Debey et al., 2014). However, five out of six recent experiments (Debey, Liefooghe et al., 2015; Foerster, Wirth, Kunde et al., 2017) and all of the current experiments showed symmetrical switch costs.[10] Symmetrical switch costs have previously been attributed to the inherent activation of the honest response in dishonest responding (Debey, Liefooghe et al., 2015), thus emphasizing the role of conflict for control adaptation in dishonest responding. As such, the transient effects explored in the present experiments cannot be fully accounted for by a task switching perspective whereas they are well in line with conflict adaptation theories. To further corroborate this assessment, we reanalyzed the data of a recently published experiment from our lab (Exp. 4 of Foerster, Wirth, Kunde et al., 2017). In this experiment, truth distractors (i.e., distractors that are compatible with an honest response) and lie distractors (i.e., distractors that are incompatible with an honest response) accompanied each question. Truth distractors facilitated honest and dishonest responses in comparison to lie distractors, revealing the initial honest response activation when responding dishonestly (cf., Debey et al., 2014). Control adaptation should diminish the impact of subsequent conflicting responses. In particular, distractor effects should be reduced after dishonest compared to after honest responses and such a finding could be explained by conflict adaptation but not by task switching. Distractor effects were indeed smaller after dishonest than after honest responses in the error rates of currently dishonest trials, $F(1, 42) = 6.27$, $p = .016$, $\eta_p^2 = .13$ (see Experiment 4 of Foerster, Wirth, Herbort et al., 2017). This finding provides

[10] A statistical comparison of switch costs for honest ([dishonest → honest] – [honest → honest]) and dishonest ([honest → dishonest] – [dishonest → dishonest]) responses for each of the current experiments revealed no significant difference in error rates (*p*s > .342) or RTs (*p*s > .114).

additional support to a control adaptation perspective on transient effects in dishonesty as put forward by the current study.

4.5.4. Connecting control adaption to other processes of dishonesty

The current design deliberately limited the experimental design to the cognitive processes involved in the two-step process of an initial activation and inhibition of the truth, and how these processes can be regulated by transient and sustained control adaptation. Based on the current results, future studies could examine how other components of lying affect the control of these cognitive processes. Motivational tendencies suggest themselves as moderators when considering previous evidence of the conflict and lying literature. For example, reward modulated control adaptation depending on the kind and rules of reward (e.g., Braem, Verguts, Roggeman, & Notebaert, 2012; Van Steenbergen, Band, & Hommel, 2009; but see Stürmer, Nigbur, Schacht, & Sommer, 2011). Similarly, gain and loss affected action control (Wirth, Dignath, Pfister, Kunde, & Eder, 2016) and conflicts were more likely and more easily avoided than approached (e.g., Dignath & Eder, 2015; Dignath, Kiesel, & Eder, 2015). In cheating paradigms, lying was more frequent when it averted loss than when it led to gain (Schindler & Pfattheicher, 2017). In this vein, it would be interesting to establish in future research whether loss compared to gain triggers more control over the activation of the truth and its inhibition, rendering lying not only more frequent but also less challenging.

4.5.5. Conclusion

The present experiments examined whether transient and sustained control adaptation elicited by dishonest and standard cognitive conflicts can affect the two-step process of initial honest response activation and its inhibition in dishonest processing. Adaptation processes did not transfer between dishonest responding and other conflicts in any of the experiments. Transient control adaptation to recent experiences of dishonesty, by contrast, improved dishonest responding substantially in all experiments while sustained control adaptation to frequent dishonest responding was absent. On the basis of previous evidence, we, therefore, propose that sustained adaptation to dishonesty only comes into play when transient adaptation is not possible in a given context; in all remaining contexts, sustained adaptation is traded for transient adaptation instead. Because transient adaptation is likely to be possible in a huge variety of settings, and will,

therefore, override more sustained effects, the present experiments document flexible but surprisingly focused control adaptation in dishonest responding.

5 Lying upside-down: Alibis reverse cognitive burdens of dishonesty

The cognitive processes underlying dishonesty, especially the inhibition of automatic honest response tendencies, are reflected in response times and other behavioral measures. Here we suggest that explicit false alibis might have a considerable impact on these cognitive operations. We tested this hypothesis in a controlled experimental setup. Participants first performed several tasks in a pre-experimental mission (akin to common mock crime procedures) and received a false alibi afterward. The false alibi stated alternative actions that the participants had to pretend to have performed instead of the actually performed actions. In a computer-based inquiry, the false alibi did not only reduce, but it even reversed the typical behavioral effects of dishonesty on response initiation (Experiment 8) and response execution (Experiment 9). Follow-up investigations of response activation via distractor stimuli suggest that false alibis automatize either dishonest response retrieval, the inhibition of the honest response, or both (Experiment 10 and 11). This profound impact suggests that false alibis can override actually performed activities entirely and, thus, documents a severe limitation for cognitive approaches to lie detection.

5.1 Introduction

Responding truthfully on each and every occasion can yield negative consequences at times, and being dishonest may come as a convenient alternative in this case. People may thus withhold information that might be harmful if revealed, or they might even present incorrect but plausible information as true facts. Lies can be further told about different topics, and among these topics, autobiographical events are particularly relevant.

Lying about autobiographical events often comes with a *false alibi*, when incorrect information is provided to conceal or deny actual events and actions. Here, we understand false alibis as giving a false impression about which activities were performed and which were not performed. Such false alibis are especially important in criminal contexts, where guilty subjects are likely motivated to present such alibis.

In the current experiments, we examined the impact of false alibis on the behavioral traces of lying in a controlled experimental design. Understanding the effects of false alibis on dishonest processing is essential to assess whether such alibis constrain the potential of behavioral measures for forensic application (i.e., lie detection). In the following, we first review theoretical models and empirical findings on the cognitive basis of dishonesty, followed by recent observations that point toward factors that moderate the behavioral effects of dishonesty. These moderators also pave the way for an empirical approach to the effects of false alibis that motivated our experimental design.

5.1.1. The cognitive basis of dishonesty

An influential approach to describing the cognitive processes underlying dishonest behavior is the *activation-decision-construction-action theory* (Walczyk et al., 2014; for a former version of the theory, see Walczyk et al., 2003). The theory assumes that respondents usually activate a representation of the truth first. However, once the respondent decides to lie, based on the social context and previous decisions, the activated truthful response needs to be inhibited to construct and deliver a plausible lie. An action component also considers that the agent can control and monitor own behavior and monitor the behavior of the receiver of the lie. The *activation-decision-construction-action theory* further holds that the proposed processes can in principle operate simultaneously and automatically (Walczyk et al., 2014).

The assumption that dishonest responding requires the inhibition of the initially activated truthful response is supported by *instructed intention paradigms*. In these paradigms, participants respond, for example, to simple autobiographical questions and are instructed to respond with a particular intention, that is, honestly or dishonestly. Intention effects in terms of differences between honest and dishonest responding were observed in behavioral, electrophysiological and hemodynamical data. In particular, dishonest responding prolongs response times (RTs) and increases error rates, leads to an enhanced recruitment of brain regions that are associated with cognitive control, and alters electrophysiological signatures in a way that points toward less direct response retrieval (e.g., Bhatt et al., 2009; Johnson et al., 2003, 2004; Pfister et al., 2014; Spence et al., 2001; Suchotzki et al., 2015; Walczyk et al., 2003). When true and false responses are collected as continuous movements toward certain spatial target locations, movement trajectories steer toward the honest response option when responding dishonestly, revealing a continued influence of the truthful response option during response execution (Duran et al., 2010). A recent study further yielded direct evidence for the assumed cognitive detour from the honest to the dishonest response (Debey et al., 2014). Participants saw questions together with truth or lie distractors, that is, the honest or dishonest response. Accordingly, the last word of the question appeared in a random position on the screen with *yes* or *no* written above and below the word. Honest and dishonest responding alike were facilitated in the presence of truth distractors compared to lie distractors, even though truth distractors corresponded to the very opposite of a correct response in dishonest trials. In a nutshell, the available evidence reveals that the honest response has to be overcome for each act of dishonest responding.

The observation of longer RTs and higher error rates for lying also encouraged researchers to study the success of these measures in lie detection. In some of these studies, participants were tested as either truth-tellers or liars throughout the experiment, or they lied in specific domains while telling the truth in others (e.g., Walczyk et al., 2009; Walczyk et al., 2005, 2012). Accordingly, each question required either an honest or a dishonest response. Even though these approaches yielded several promising findings, the resulting classification success is currently insufficient to use this method for lie detection outside the laboratory.

In the *instructed intention paradigm*, by contrast, cues inform participants in each trial whether to respond honestly or dishonestly, and each question is answered equally often with both intentions (e.g., Furedy et al., 1988; Spence et al., 2001). Hence, the instructed intention paradigm maximizes effect sizes and, thus, provides a promising basis for lie detection. Using the instructed intention effect for lie detection, however, also requires a firm understanding of potential moderating factors, and we will, therefore, describe some of these factors in the following section.

5.1.2. Just how basic is the basis?

The size of the intention effect in *instructed intention paradigms* is a direct function of differences in cognitive processing between honest and dishonest responding, and several factors modulate this difference. Rehearsing specific lies, for instance, facilitates lying up to a level where lying can become easier than responding honestly, provided that each question was responded to with only one intention throughout an experimental session (Hu, Chen et al., 2012; Walczyk et al., 2009, 2012). Clearly, participants learn stimulus-response associations in this setting, where the automatically activated response seems to be the dishonest one instead of the honest (akin to storing “instances” of stimulus-response episodes; Logan, 1988).

Besides such item-specific learning, general changes in cognitive control settings influence dishonest processing in a sustained as well as transient manner: Having responded honestly or dishonestly changes future honest and dishonest responding. Sustained influences describe how lying is modulated by the frequency of dishonest behavior. To differentiate between stimulus-response learning and changes in control settings, the frequency of both intentions can be manipulated in inducer questions, while the frequency is held constant in test questions (Van Bockstaele et al., 2012; Verschuere et al., 2011). Typically, inducer questions had to be answered only honestly, only dishonestly, or both across different conditions. Test questions afforded an equal number of responses with both intentions irrespective of condition, and both question types appeared in a random sequence. The intention effect became smaller with a larger proportion of dishonest trials for inducer questions and, importantly, also for test questions. When responses were given in an environment with a balanced proportion of honest and

dishonest trials afterward, this modulation only held for the inducer questions, which again indicates acquired stimulus-response associations (Van Bockstaele et al., 2012).

Similarly, control settings can change transiently, that is, from trial to trial. Therefore, honest and dishonest responses were analyzed as a function of the intention in the preceding trial, akin to methods of the literature on task switching (Debey, Liefooghe et al., 2015; Foerster, Wirth, Kunde et al., 2017; for reviews on task switching, see Kiesel et al., 2010; Monsell, 2003). Repeated honest or dishonest responding was easier than switching between both intentions. Performance differences between honest and dishonest responding, however, were mostly unaffected by these switch costs, except when the upcoming intention was announced shortly before question onset (Foerster, Wirth, Kunde et al., 2017). Switch costs between honest and dishonest responding also provide a new perspective on the studies on sustained influences described above. Because these studies manipulated the frequency of honest and dishonest trials, they also introduced varying ratios for repetition and switch trials. For example, a highly dishonest environment featured mostly repetitions of dishonest trials and rarely switches to dishonest responding, whereas honest trials were mostly intention switches and rarely repetitions. The observed effects of frequency manipulations could, thus, stem from transient instead of sustained changes in control settings or from a combination of both (Debey, Liefooghe et al., 2015; Foerster, Pfister et al., 2018; Foerster, Wirth, Kunde et al., 2017; Van Bockstaele et al., 2012).

Another factor that affects dishonest responding in a sustained manner is the instruction of deliberate response strategies (Hu, Chen et al., 2012). Participants in this study first were naïve about the intention effect and responded honestly and dishonestly to information in separate blocks. Then they learned about the usual intention effect in performance and their own mean performance in the two blocks. They were asked to diminish the intention effect by speeding up responding in the dishonest block and indeed, participants were able to do so. Although the intention effect became smaller as dishonest RTs were decreased, it did not vanish entirely. This was only the case when participants went through an additional dishonest training block with the same information in which they conceivably acquired stimulus-response associations.

Overall, these results suggest that a dishonest response is retrieved directly by means of a stimulus when this association has been learned sufficiently well before, which can render dishonest responding as easy as honest responding. Otherwise, an initially activated honest response has to be overcome when responding dishonestly. This is easier when dishonest processing already took place recently or very frequently, diminishing the intention effect, and it is harder when previous or frequent responding was honest, enhancing the intention effect.

Interestingly, the effect of false alibis as a frequent companion of dishonest behavior has not yet been addressed in research on the behavioral consequences of dishonesty and in the evaluation of lie detection methods. The modulating influences described above indeed suggest that false alibis may alter the way dishonest behavior is processed. Following the existing evidence, preparation of false alibis could render the process of inhibiting the truth more efficient or change dishonest responding even more drastically to a process of directly retrieving the appropriate dishonest response. Investigating precisely this effect of false alibis is the goal of the present experiments. If false alibis change the way dishonest responses are processed, this would be relevant for the development of lie detection methods as for example the *autobiographical implicit association test* (Sartori, Agosta, Zogmaister, Ferrara, & Castiello, 2008), the concealed information test (e.g., Ben-Shakhar & Elaad, 2003) or tests that rely on the cognitive load induced by dishonesty (e.g., Walczyk et al., 2005).

5.2 Experiment 8

In Experiment 8, we aimed at examining the effects of a false alibi on (dis)honesty in an instructed intention paradigm. To establish a situation that resembles applied forensic settings, the experiment was divided into two separate parts: a mission and an inquiry. In the mission, participants engaged in allegedly secret activities. After performing these activities, they received a false alibi that detailed a series of alternative actions. The inquiry took place on a computer with discrete *yes/no* responses via a keypress, and participants were to pretend to have had engaged only in the alibi activities and not in the activities they actually had performed. Accordingly, they were to respond dishonestly during the inquiry when asked to respond honestly, and they were to respond honestly when asked to respond oppositely about the mission. Note that the intention instructions in the inquiry

were honest and “opposite” instead of honest and “dishonest”. This change in the instructed intention paradigm was introduced to make the instructions more applicable in forensic settings, as it would seem rather odd to ask actual suspects to respond honestly or dishonestly. As a baseline, to gauge potential alibi effects, we further included routine questions in the inquiry (relating to daily activities and unrelated to the mission) in addition to the mission questions.

In a nutshell, the experimental design established conditions in which participants responded in correspondence with their actually experienced activities, and conditions in which their responses and activities were noncorresponding. We hypothesized that responding would be easier when a response corresponded with the experiences of participants, for example, when participants gave an affirmative response when asked about activities they actually engaged in and negated questions about activities they did not engage in. Accordingly, the manipulation of activity-response correspondence should affect our behavioral measures, that is, error percentages and RTs. Participants should respond slower and less accurately in noncorresponding trials than in corresponding trials. The critical question was whether the false alibi would reduce this correspondence effect. Such an effect would be evident in reduced correspondence effects for mission questions (for which participants had an alibi) relative to routine questions (for which there was no alibi).

5.2.1. Method

Participants and overall procedure | A sample size of 44 participants was determined with a power analysis based on an effect size of $d_z = 0.5$, $\alpha = .05$ and a power of $1 - \beta = .90$ (calculated with the power.t.test function in R version 3.1.1). We used a generic medium effect size as a conservative estimate because effects of dishonesty and their modulation are usually large in RTs and error rates (e.g., Foerster, Wirth, Kunde et al., 2017; Van Bockstaele et al., 2012). Participants gave written informed consent and received either monetary compensation or course credit (age: mean (*M*) = 20.6, standard deviation (*SD*) = 3.07; 39 female; 40 right-handed).

The experiment was divided into two separate parts as shown in Figure 15. In the first part, participants went through a mission in which they performed certain actions (e.g., drawing a triangle and a circle on a sheet of paper). By the end of the mission, they were

informed about an upcoming inquiry regarding their activities and were instructed about plausible activities that they should pretend to have performed in this inquiry. In the second part, participants worked on the computerized inquiry and were asked to respond to a number of *yes/no* questions with button presses. Both parts are described in more detail in the following.

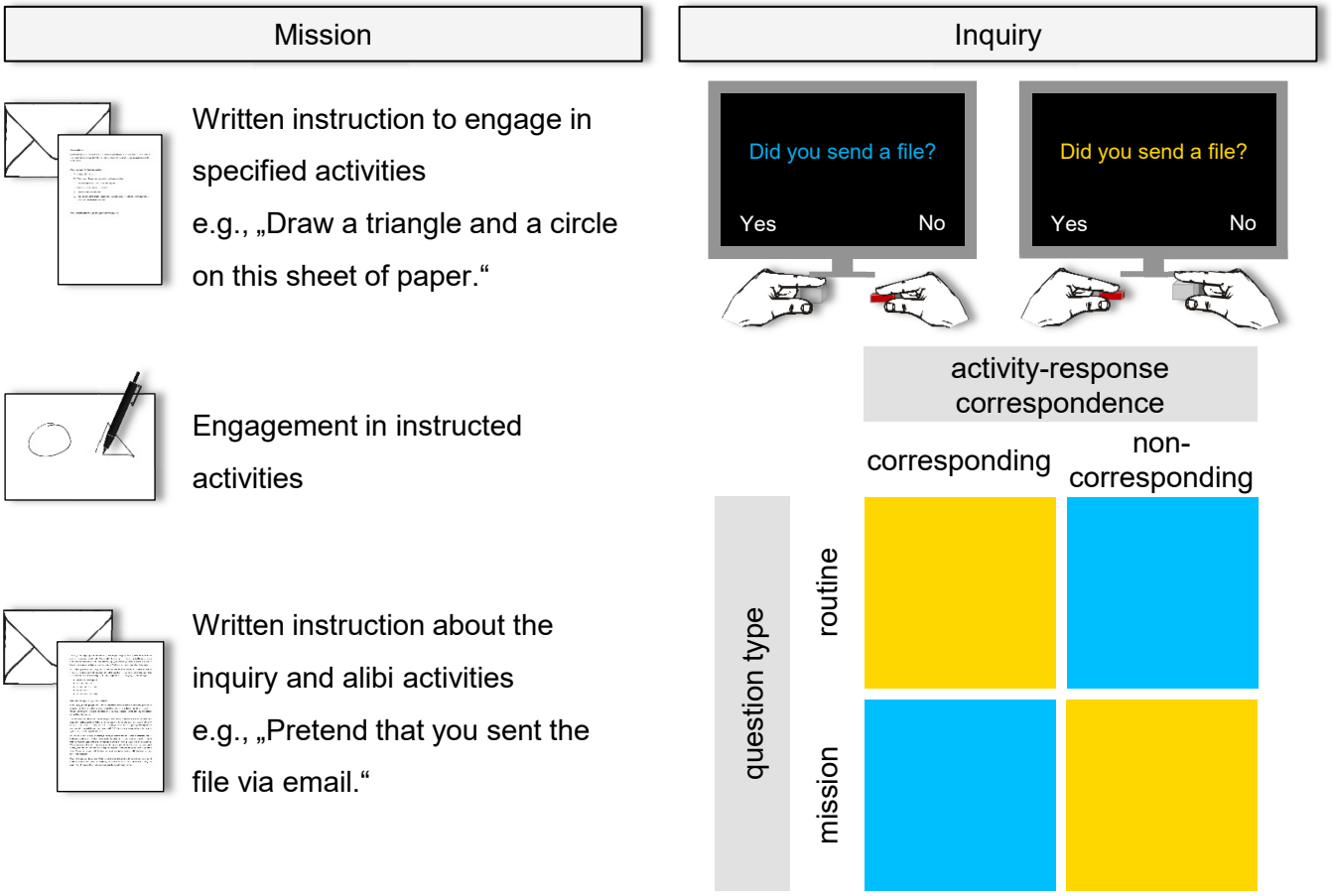

Figure 15 | Design of Experiment 8. Participants responded to routine and mission questions with *yes* or *no*. The font color of the question cued whether participants should give an honest (here: yellow, lighter font) or an opposite to honest response (here: blue, darker font; for legibility, question and response labels are displayed in a larger font than in the actual experiment). Participants had to follow these cues when they responded to routine questions but adapt their responses for mission questions. Participants gave responses that corresponded or did not correspond with their actual activities for routine and mission questions. For mission questions, participants were supplied with an explicit false alibi that stated a series of alternative actions that they were to pretend to have performed instead of their actual activities. For example, participants had to affirm all alibi actions and deny performed actions when the question appeared in yellow font (indicating an honest response), whereas they had to deny the alibi actions and to affirm the performed actions when the question appeared in blue font (indicating an opposite response).

Mission | The experiment began with a mission that required participants to perform activities alone in a room. The following items in the room were important for the mission: a desk with a chair, a box on the desk containing a stack of empty sheets of paper and a pen, a box under the table with a slit, an USB stick on the desk next to the box, a computer with a screen, a keyboard and a mouse.

All participants received the same instructions through a letter (see Appendix 4) in an envelope and were asked to strictly follow these instructions. Participants had to (a) sit down at the desk, (b) take a sheet of paper and the pen from a box on the desk, (c) draw a triangle and a circle on this sheet of paper, (d) put the pen back in the box on the table, (e) tear the sheet of paper in half, and (f) hide one piece under the stack of paper in the box on the desk and the other one in the box below the desk. Participants then learned from the letter that they would be questioned about these activities in a computerized inquiry afterward. Crucially, participants were led to believe that the majority of participants received different instructions. Namely, these other participants allegedly engaged in different activities with the remaining objects, and that the other participants had to be honest about these activities in the following inquiry. In contrast, the letter stated that the current participant was chosen for a special mission in this experiment to learn more about lie detection. This mission required hiding their true activities and pretending to have performed plausible alibi actions. Alibi actions were (a) to switch on the computer, (b) use the USB stick, (c) open a file called “table,” (d) write an e-mail, and (e) send the file via this e-mail. Participants were instructed to not actually engage in any of these activities.

The letter now explained the following inquiry in detail. Participants would not only be asked questions about their activities in the room – referred to as mission questions in the following – but also about activities they could have had experienced or not experienced on the same day – referred to as routine questions in the following. They also learned that, on each trial, they would either be asked to respond honestly or to give the opposite response. How to respond would be indicated by the color of the question (the exact assignment of color to honest or opposite was provided in the inquiry).

When participants had had engaged in a routine activity, they would need to respond with *yes* when the color indicated to give an honest response and with *no* when the color cue indicated to give an opposite response. When participants had not had engaged in a routine activity, however, they would need to respond with *no* when the color indicated to

give an honest response and with *yes* when the color indicated to give an opposite response.

Importantly, participants were to respond differently when confronted with mission questions to give the impression that the false alibi reflected true events. So when confronted with mission questions, participants would always have to lie when the color instructed an honest response, similar to a guilty suspect who would respond to the police. Accordingly, questions about their actually performed activities in the mission (e.g., hiding a piece of paper) would have to be negated when the color instructed an honest response and affirmed when the color instructed an opposite response. In contrast, questions about alibi activities (e.g., sending an e-mail) would need to be affirmed when the color cued an honest response and negated when the color cued an opposite response. Participants were encouraged to remember these instructions well. Afterward, they were asked to insert the letter in the box below the table and to go to another room for the inquiry. Accordingly, responses could be obtained that either corresponded or did not correspond with the actual activities of the participants for routine and mission questions (see Figure 15).

Inquiry | Ten routine and ten mission questions were prepared (see Table 3 in Appendix 1). Routine questions were picked carefully to ask about five activities that were very unlikely experienced and five activities that were very likely or even surely experienced (through the participation in the experiment) on that day. Five mission questions asked about the activities that participants engaged in alone in the room and the other five asked about the activities that participants did not engage in. The questions were matched for length: All questions featured five words, and the average number of characters per question was either 25 or 26 for each condition. Participants sat in front of a 22-inch TFT screen. They saw all questions on the screen, grouped in routine and mission questions, before the inquiry started and were informed that these questions would be presented randomly. Participants responded *yes* and *no* with their index fingers via the keys *D* and *K* of a standard QWERTZ keyboard. The assignment of *yes* and *no* responses to the response keys was counterbalanced across participants, but constant for each participant. The font color of the question (yellow vs. blue) indicated whether participants were to respond honestly or to give an opposite response. To be sure that participants understood the meaning of the cues, they were asked, “Did you understand

the instructions?" and the font color of this question first cued an honest response and then an opposite response. Only a correct response prompted the next screen. The assignment of cue meaning to the colors yellow and blue was counterbalanced across participants. From this point onward, participants were no longer reminded that they had to respond differently to mission questions because participants were led to believe that they were on a special mission and that the experimenter did not know about their true activities in the room. Hence, the presentation of error feedback in case of a false response was not possible. To prevent participants from random responding, however, they were told that the computer would monitor whether they responded inconsistently, that is, gave different answers when the same question was presented repeatedly in the same color. Participants were encouraged to respond as fast and accurately as possible.

Each trial of the inquiry started with a centrally presented white fixation cross on black background for 500 ms. In case of a response during fixation, error feedback was provided for 1500 ms ("Zu früh!" – German for "too early!"). Then a question appeared in yellow or blue font in the center of the screen. The labels for *yes* and *no* (German: "ja" and "nein") were written in white font in the bottom left and right half of the screen as a reminder of the key-response assignment. The question and response labels stayed on the screen until a response was given or a time limit of 3000 ms was exceeded. In the latter case, appropriate error feedback was provided for 1500 ms ("Zu langsam!" – "too slow!"). The next trial started after 500 ms.

The combination of 2 question types (routine vs. mission; with 10 questions each) × 2 activity-response correspondence (corresponding vs. noncorresponding) resulted in 40 trial combinations. Participants went through 11 blocks with 40 trials each, where each combination was presented once. Participants could take self-paced breaks in between blocks.

Data treatment and analyses | The first block served as practice and was thus excluded from all statistical analyses, as was the first trial of each block. Error rates were computed as the number of trials in which participants gave a response that was inappropriate for the given combination of question and cue, relative to the number of trials without any other errors (i.e., commission errors plus correct trials). Accordingly, less than 50% correct trials would mean that participants guessed or wrongly memorized the instructions. Eight participants were excluded because they had error rates of 50% or

above in at least one of the design cells. Thirty-six participants remained for statistical analyses and we did not replace the removed participants because the sample size was computed based on conservative estimates for possible effect sizes. Trials that used the same question as the trial before were excluded from all statistical analyses to avoid confounds due to the retrieval of short-term stimulus-response bindings (2.5%).

Trials in which participants gave an early response during fixation, or did not respond, or responded with any other key than *D* or *K* (6.2%), were excluded prior to computing and analyzing error rates. All erroneous trials were excluded before analyzing RTs. Trials with RTs that deviated more than 2.5 *SD*s from the respective cell mean were eliminated as outliers (1.7%).

Error rates and RTs were examined in separate 2 × 2 ANOVAs with the within-subjects factors question type (routine vs. mission) and activity-response correspondence (corresponding vs. noncorresponding). In case of a significant interaction, we used two-tailed paired *t*-tests to scrutinize the size of the correspondence effect for each question type.

5.2.2. Results

Error rate │ Responses to routine questions were less accurate than responses to mission questions (see Figure 16A), $F(1, 35) = 4.21$, $p = .048$, $\eta_p^2 = .11$. Surprisingly, the main effect of activity-response correspondence was not significant, $F < 1$, whereas the interaction between both factors was significant, $F(1, 35) = 115.08$, $p < .001$, $\eta_p^2 = .77$. Non-corresponding responses to routine questions were more error-prone than corresponding responses to routine questions, $t(35) = 9.10$, $p < .001$, $d_z = 1.52$, reflecting the hypothesized correspondence effect. A reversed correspondence effect with less accurate corresponding than non-corresponding responses, however, emerged for mission questions, $t(35) = -8.40$, $p < .001$, $d_z = -1.40$.

Response time │ Responses to mission questions were slower than responses to routine questions (see Figure 16B), $F(1, 35) = 45.44$, $p < .001$, $\eta_p^2 = .57$. Again, the main effect of activity-response correspondence was not significant, $F < 1$, whereas the interaction between both factors was significant, $F(1, 35) = 172.81$, $p < .001$, $\eta_p^2 = .83$. Responses to routine questions showed the hypothesized correspondence effect, $t(35) =$

13.86, $p < .001$, $d_z = 2.31$, and a reversed correspondence effect was evident for mission questions, $t(35) = -11.24$, $p < .001$, $d_z = -1.87$.

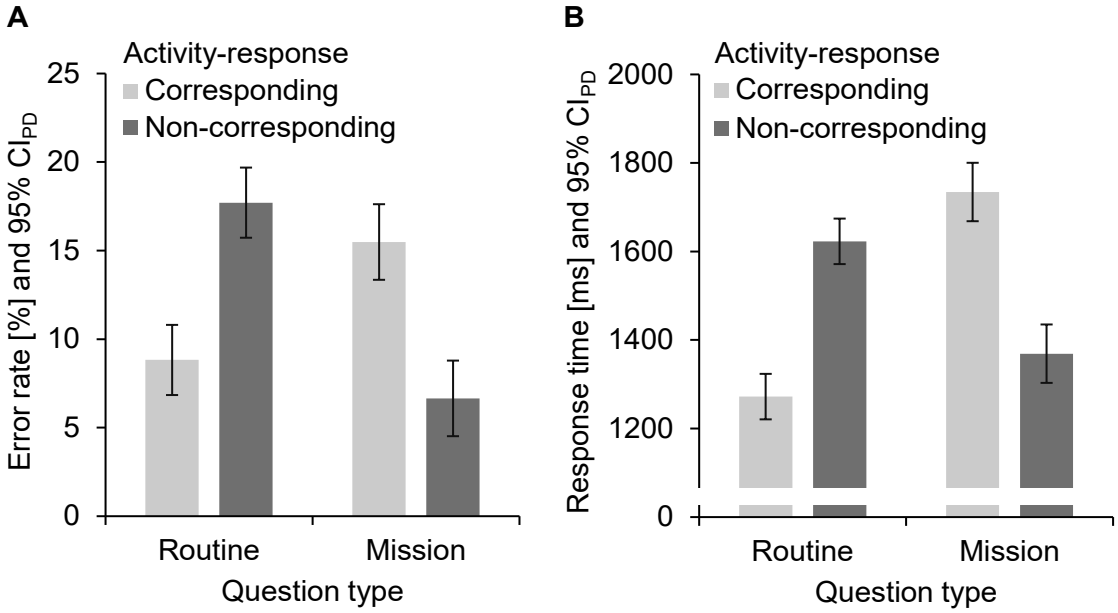

Figure 16 | Mean error rates (A) and response times (B) of Experiment 8, plotted as function of activity-response correspondence and question type. Error bars represent the 95% confidence interval of paired differences (CI_{PD}; Pfister & Janczyk, 2013), computed separately for routine and mission questions.

5.2.3. Discussion

In Experiment 8, we provided participants with a false alibi which entailed to give a false impression about which activities they had performed and which they had not performed in a mission. In a computerized instructed intention paradigm, cues prompted either honest or opposite *yes* and *no* responses to routine and mission questions. As participants were instructed to stick to the false alibi, they had to answer mission questions honestly in the presence of the opposite cue and dishonestly in the presence of the honest cue. By contrast, routine questions had to be answered exactly as the cues instructed.

RTs and error rates were analyzed as a function of whether the required response corresponded to what the participants had actually done. Responses to routine questions replicated common findings on dishonesty, as noncorresponding responses took longer and were more error-prone than corresponding responses (e.g., Debey et al., 2012; Foerster, Wirth, Kunde et al., 2017; Spence et al., 2001). This effect was further moderated by the false alibi instruction in mission questions; the extent of this manipulation came unexpected, however: Responses in accordance with the false alibi were, in fact, faster and more accurate than responses based on the participants' actual activities (i.e., activity-

response corresponding responses). Hence, it was easier for participants to negate activities they had actually experienced in the mission and affirm activities they had not performed. This pattern of results indicates that participants internalized the false alibi to an extent where the noncorresponding, dishonest response became the default. As such, internalizing false alibis seems to change dishonest responding in a similar way as the rehearsal of dishonest responses does (Hu, Chen et al., 2012; Walczyk et al., 2009, 2012), even though responses in accordance with the false alibi and opposite responses are delivered equally often. Participants might have formed explicit intentions for how to respond to the mission questions. Such explicit intentions have been shown to counter automatic retrieval of spontaneous action tendencies (Waszak, Pfister, & Kiesel, 2013), and might, therefore, represent a plausible mechanism to explain the observed effects. To follow up on these findings, we examined the impact of false alibis again in a more fine-grained procedure in Experiment 9 that captures not only response initiation but also response execution.

5.3 Experiment 9

Without false alibis, a reliable signature of dishonest responding has also been reported in a study that measured continuous movements to capture response initiation and execution (Duran et al., 2010). Participants in this study had to move a mouse cursor from a start area to response labels on the top left and top right of a screen. These movements were initiated later, executed more slowly and their trajectory was more strongly contorted toward the alternative response label when responding dishonestly than when responding honestly. Similar observations were made for rule violations, which also show a continued influence of the rule-based response (e.g., Pfister, Wirth, Schwarz, Steinhauser et al., 2016; Wirth, Pfister, Foerster et al., 2016). So for rule violations and lies, there is a conflict between the appropriate response and an automatically activated default response. As such, capturing continuous movements could provide a more detailed picture of the impact of conflicting response tendencies in honest and dishonest responding. In Experiment 9, participants went through the same mission as in the preceding experiment but conducted the inquiry on an iPad to capture continuous finger sweeping movements. This allowed us to study the effects of activity-response correspondence and its modulation by false alibis on response initiation and execution.

5.3.1. Method

Participants and overall procedure | A new sample of 44 participants was recruited for either monetary compensation or course credit (age: *M* = 26.0, *SD* = 7.14; 27 female; 41 right-handed). As effects of dishonesty are similarly large for discrete and continuous performance measures, the same sample size as in Experiment 8 was used (cf. Duran et al., 2010). All participants gave written informed consent. The experiment was again divided into two separate parts. The first part of the experiment, that is, the mission, was the same as in Experiment 8 with minor changes in the procedure. The inquiry was conceptually similar but required finger-sweeping responses on an iPad.

Inquiry | The experiment began in the room where the mission took place. First, participants learned how to respond to questions on an iPad. Participants were asked whether they understood the instruction, whether they were sitting on a chair, whether they were awake and whether they were currently lying on a beach. When participants gave a wrong answer to any of these practice questions, the question was repeated until participants gave the correct answer.

Participants used the index finger of their dominant hand to respond to questions. A question appeared in each trial in the center of the screen (see Figure 17; notice that the font color of the practice questions was black and participants had not learned about the upcoming cues yet). When participants touched the starting area at the bottom of the display, the question disappeared and the response labels for *yes* and *no* appeared randomly on the left and right side at the top of the screen. If participants touched the starting area later than 1500 ms after question onset or left the starting area later than 500 ms after touching it, error feedback was provided in red font in the center of the screen until participants stopped touching the iPad (“Bitte schneller reagieren!” – German for “please respond faster!”). This procedure stressed fast responses to maximize the effects of the independent variables. The next question was presented after 400 ms. After each practice question had been answered correctly, the experimenter left the room with the iPad and took it to the inquiry room. The participant stayed and went through the same mission that was used in Experiment 8. Accordingly, the routine and mission questions were the same as in Experiment 8.

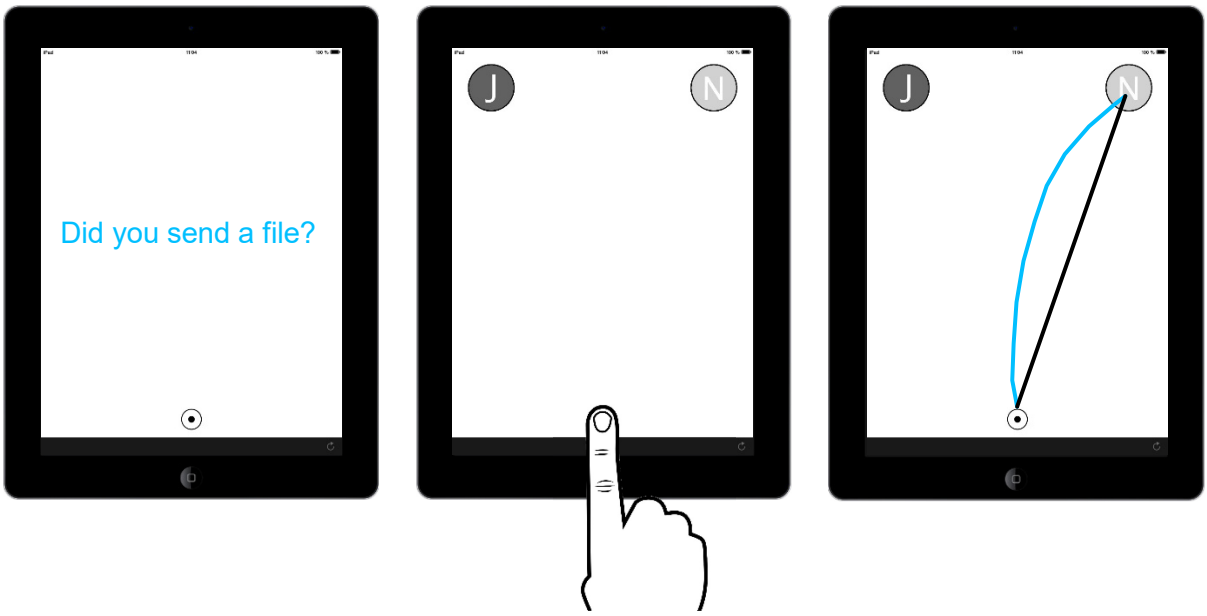

Figure 17 | Setup of the inquiry in Experiment 9 (left and center display), and an exemplary movement trajectory (right display). Trials started with a colored question (not drawn to scale to improve legibility). Touching the starting area in the bottom center of the screen made the question disappear whereas the response labels for *yes* ("J" – German "Ja") and *no* ("N" – German "Nein") appeared. The time that passed between question onset and touching the starting area was the reading duration. Initiation time captures the time that passed from touching the starting area until leaving it. The time that passed from that moment until the finger stopped touching the iPad was the movement time. The area under the curve is the area between the actual trajectory (blue [light] line) and a virtual direct line from the start point to the endpoint of the movement (black [dark] line). Areas under the curve were computed after time-normalizing the individual trajectories to 101 data points via linear interpolation. The larger the area under the curve, the stronger was the attraction of the movement toward the alternative, incorrect response area.

In the inquiry, routine and mission questions appeared in yellow and blue font to indicate honest and opposite responding. Participants learned about the honest cue first and responded to each of the routine and mission questions in that color once in a random order. Afterward, the 20 questions were presented again in the color for opposite responding (practice block). In the following 10 blocks, honest and opposite cues appeared in a random sequence. The combination of 2 question types (routine vs. mission; with 10 questions each) × 2 activity-response correspondence (corresponding vs. noncorresponding) × 2 response positions (yes/left and no/right vs. no/left and yes/right) resulted in 80 trial combinations, presented once in each block.

Data treatment and analyses | All practice trials and the first trial of each block were excluded from all statistical analyses. Error rates were computed as in Experiment 8 and six participants were excluded because they had error rates of 50% or more in at

least one of the design cells. Thirty-eight participants remained for statistical analyses. Trials that used the same question as the trial before were excluded from all statistical analyses as well (3.9%).

Trials in which participants failed to touch the starting area within 1500 ms after question onset, failed to leave the starting area within 500 ms after touching it, or did not finish their movement in one of the response areas, were excluded prior to computing and analyzing error rates (18.6%; note that this rather high number reflects the emphasis on speeded responding that we sought to stress in this experiment). We selected reading duration, initiation time, movement time, and area under the curve as dependent variables to get a grasp on response initiation (reading duration, initiation time) and execution (movement time, area under the curve; for a detailed description of these variables, see Figure 17). The selection of those four variables was motivated by their high sensitivity to similar experimental manipulations in previous examinations (e.g., Pfister, Wirth, Schwarz, Steinhauser et al., 2016; Wirth, Dignath et al., 2016; Wirth, Pfister, Foerster et al., 2016). All error trials were excluded before analyzing those variables. Trials, where at least one of the dependent values deviated more than 2.5 SDs from the respective cell mean, were eliminated as outliers (7.8%).

All dependent variables were examined in separate 2 × 2 ANOVAs with the within-subjects factors question type (routine vs. mission) and activity-response correspondence (corresponding vs. noncorresponding). In the case of significant interactions, we used two-tailed paired *t*-tests to scrutinize the size of the correspondence effect for each question type.

5.3.2. Results

Error rate │ There was a non-significant trend toward more accurate responses to mission questions compared to routine questions (see Figure 18A), $F(1, 37) = 3.79$, $p = .059$, $\eta_p^2 = .09$. The main effect of activity-response correspondence was not significant, $F < 1$, whereas the interaction between both factors was significant, $F(1, 37) = 48.51$, $p < .001$, $\eta_p^2 = .57$. Non-corresponding responses were more error-prone than corresponding responses to routine questions, $t(37) = 6.01$, $p < .001$, $d_z = 0.97$, whereas non-corresponding responses were more accurate than corresponding responses to mission questions, $t(37) = -5.43$, $p < .001$, $d_z = -0.88$.

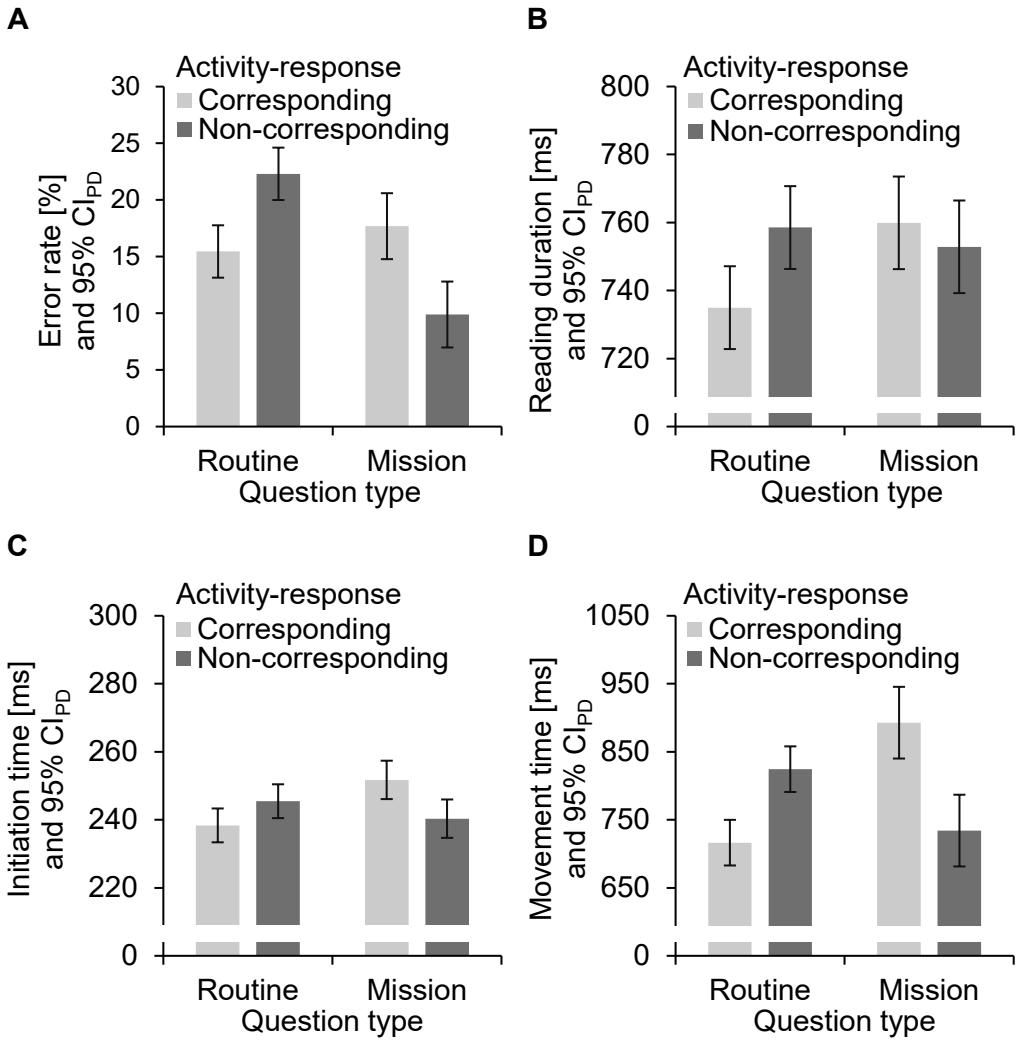

Figure 18 | Mean error rates (A), reading durations (B), initiation times (C) and movement times (D) of Experiment 9, plotted as function of activity-response correspondence and question type. Error bars represent the 95% confidence interval of paired differences (CI_{PD}), computed separately for routine and mission questions.

Reading duration | Participants spent more time before touching the starting area in mission than in routine trials (see Figure 18B), $F(1, 37) = 4.68$, $p = .037$, $\eta_p^2 = .11$. Moreover, non-corresponding trials increased reading durations compared to corresponding trials, $F(1, 37) = 4.17$, $p = .048$, $\eta_p^2 = .10$. A significant interaction qualified the main effects, $F(1, 37) = 9.63$, $p = .004$, $\eta_p^2 = .21$. Whereas non-corresponding trials increased reading durations compared to corresponding trials for routine questions, $t(37) = 3.92$, $p < .001$, $d_z = 0.64$, there was no effect of correspondence on reading durations for mission questions, $t(37) = -1.05$, $p = .301$, $d_z = -0.17$.

Initiation time | Participants took longer to leave the starting area when they responded to mission compared to routine questions (see Figure 18C), $F(1, 37) = 5.12$, $p = .030$, $\eta_p^2 = .12$. The main effect of activity-response correspondence was not significant, $F(1, 37) = 2.67$, $p = .111$, $\eta_p^2 = .07$, whereas there was a significant interaction of both factors, $F(1, 37) = 16.54$, $p < .001$, $\eta_p^2 = .31$. Initiation times were faster in corresponding than in non-corresponding trials for routine questions, $t(37) = 2.89$, $p = .006$, $d_z = 0.47$, and an opposite pattern of results was evident for mission questions, $t(37) = -4.09$, $p < .001$, $d_z = -0.66$.

Movement time | Movements from the starting area to the correct response area took longer for mission than for routine questions (see Figure 18D), $F(1, 37) = 13.50$, $p = .001$, $\eta_p^2 = .27$, and in corresponding compared to non-corresponding trials, $F(1, 37) = 6.89$, $p = .013$, $\eta_p^2 = .16$. The main effects were qualified by a significant interaction, $F(1, 37) = 46.47$, $p < .001$, $\eta_p^2 = .56$. Non-corresponding trials increased movement times for routine questions compared to corresponding trials, $t(37) = 6.53$, $p < .001$, $d_z = 1.06$, but an opposite pattern of results emerged for mission questions, $t(37) = -6.10$, $p < .001$, $d_z = -0.99$.

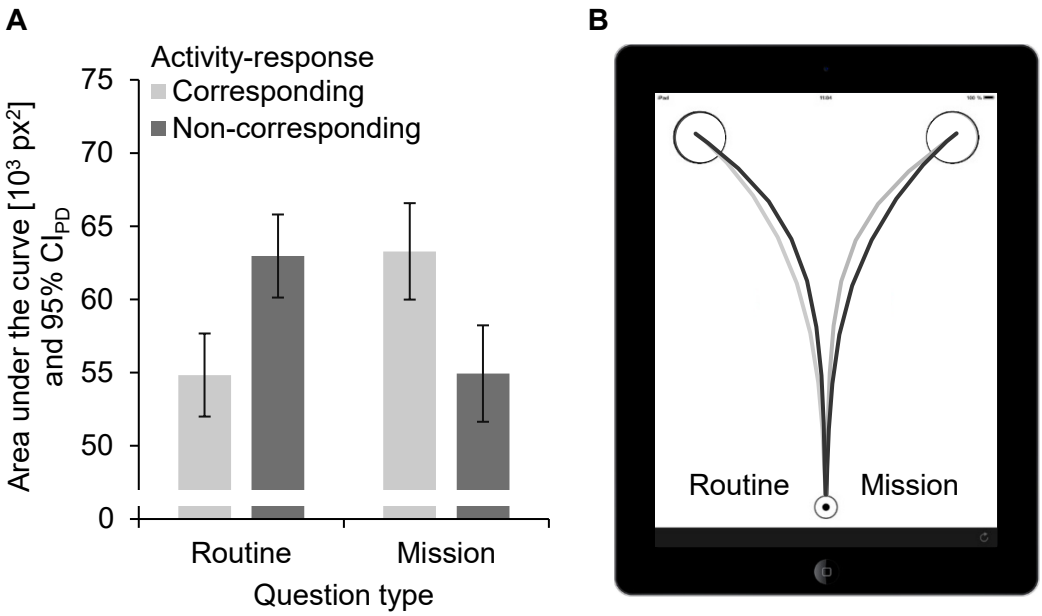

Figure 19 | Mean areas under the curve (A) and mean time-normalized movement trajectories (B) of Experiment 9 plotted as function of activity-response correspondence and question type. For simplicity, movement trajectories of routine questions are plotted to the left and mission questions are plotted to the right. Error bars represent the 95% confidence interval of paired differences (CI_{PD}), computed separately for routine and mission questions.

Area under the curve | The main effects of question type and activity-response correspondence were not significant (see Figure 19), *F*s < 1. The interaction of both factors was significant, however, $F(1, 37) = 41.16$, $p < .001$, $\eta_p^2 = .53$. Non-corresponding responses were bent more strongly toward the competing response area than corresponding movements for routine questions, $t(37) = 5.80$, $p < .001$, $d_z = 0.94$. For mission questions, however, the curve was bent more strongly toward the competing response area in corresponding than in non-corresponding trials, $t(37) = -5.13$, $p < .001$, $d_z = -0.83$.

5.3.3. Discussion

As in Experiment 8, participants received a false alibi before working on the inquiry during which they had to give a false impression about which activities they did or did not perform. In contrast to Experiment 8, the inquiry took place on an iPad that captured continuous responses to measure markers of response initiation and markers of response execution alike. Cues instructed honest and opposite *yes* and *no* responses to routine and mission questions. In accordance with their mission, participants answered mission questions with an honest response in the presence of opposite cues and with a dishonest response in the presence of honest cues while responding exactly as the cues instructed to routine questions.

Response initiation and execution replicated the strong impact of false alibis on dishonest responding. Responses to mission questions in accordance with the false alibi (i.e., negation of performed actions and affirmation of not performed actions) were less error-prone, initiated and executed faster, and less attracted by the opposite response label than responses in accordance with participants' actual activities (i.e., negation of not performed actions and affirmation of performed actions). Responses to routine questions again showed the traditional correspondence effect as noncorresponding responses were less accurate, slower initiated and executed, and more strongly bent toward the competing response side compared to corresponding responses (cf., Duran et al., 2010).

Together with the findings of Experiment 8, these results establish false alibis as a strong influence on dishonest processing, altering its behavioral signature substantially. Strikingly, a close look at the statistics reveals that for most dependent variables (error rate and RT of Experiment 8; initiation time and area under the curve of Experiment 9) the

interaction was so strong that the reversed correspondence effect in mission questions was just as large as the traditional correspondence effect in routine questions, indicated by a nonsignificant main effect of activity-response correspondence. After establishing that there is a strong impact of false alibis on dishonest responding, the next step is to scrutinize how processing of dishonest responding changes under false alibis in Experiment 10. Finally, Experiment 11 provides a control condition to assess whether the specific design of the mission questions (relative to routine questions) promoted the observed effects.

5.4 Experiment 10

The complete reversal of the correspondence effect by explicit false alibis might be taken to suggest that the dishonest rather than honest response to mission questions became activated by default. That is: The observed reversal suggests that alibi-related questions might trigger a dishonest response, which would have to be inhibited to respond honestly.

An appealing method to investigate automatic response activation in *instructed intention paradigms* was recently provided by Debey et al. (2014). As explained in the introduction, the authors presented truth or lie distractors (*yes* and *no*) with each question (for an illustration in the context of the current experiments, see Figure 20). That is, truth distractors would be *yes* for an affirmative response and *no* for a negation whereas lie distractors would be *no* for an affirmation and *yes* for a negation. They found not only honest but also dishonest responding to be facilitated by honest distractors. For example, if participants were to respond dishonestly with *yes*, responses were slower with *yes* distractors than with *no* distractors. In that study, the distractors seemed to have activated a response in the time window of the first process in dishonest responding, namely during honest response activation. As such, honest distractors facilitated this first process while dishonest distractors hampered it because of conflicting response activation.

In Experiment 10, we combined the false alibi manipulation with a computerized inquiry that featured distractors that either did or did not correspond to the participants' actual experiences in the mission. Responses that match the false alibi should again be faster and more accurate compared to responses that match the actual activities. We hypothesized that if false alibis change the response that is activated by default, then

noncorresponding instead of corresponding distractors should facilitate responding to mission questions. Responses to routine questions did not come with a false alibi and, thus, should be facilitated with corresponding distractors irrespective of whether participants responded honestly or dishonestly (Debey et al., 2014).

5.4.1. Method

Participants and overall procedure │ A sample size of at least 39 participants was required to detect a distractor effect of d_z = 0.46 (computed from RT data of Experiment 1 of Debey et al., 2014, Table 1, p. 328) with a power of 80% in a two-tailed test (α = 5%). In the tradition of the former experiment, we opted for a sample of 44 participants (age: mean = 21.6, SD = 3.55; 37 female; 39 right-handed) but decided to replace excluded participants as distractor effects are not as established as general effects of dishonesty and to allow a fair comparison of Experiment 10 and 11 based on same sample sizes. As we had to exclude six participants for the same criteria as in the former experiments, a total of 50 participants took part in Experiment 10. All participants gave written informed consent and received course credit for participation. Like in the former experiments, participants went through a mission and an inquiry. All procedures were as in Experiment 8 with the following changes.

Inquiry │ We adapted our trial procedure to the design of Debey et al. (2014; see Figure 20). Throughout all experimental trials, the first constant part of each question ("Hast du" – German for "did you") stayed centrally on the top of the screen as did the labels for *yes* and *no* (German: "ja" and "nein") in the bottom left and right of the screen as a reminder of the key-response assignment. The sentence fragment and response labels appeared in white font on black background. Each trial of the inquiry started with the presentation of these fixed features and after 1000 ms a white fixation cross appeared additionally in the center of the screen for 200 ms. After the offset of the fixation cross, the fixed features stayed on screen for 300 ms. In case of an early response before question onset, error feedback was provided for 1500 ms ("Zu früh!" – German for "too early!"). The fixed features were accompanied by a question and distractors. The distractors were either *yes* or *no* (German: "ja" and "nein") and occurred in a distance of 5% above and below the question, respectively (percentages refer to the vertical coordinate on the computer screen). The position of the question and its distractors was

determined randomly on each trial. They appeared centrally on one of four positions on the vertical axis (33%, 43%, 57%, and 67%). The font color of the distractors was white whereas yellow and blue font was again used for the central question fragment. The question and distractors stayed on the screen until a response was given or a time limit of 3000 ms was exceeded. In the latter case, error feedback was provided for 1500 ms ("Zu langsam!" – German for "too slow!").

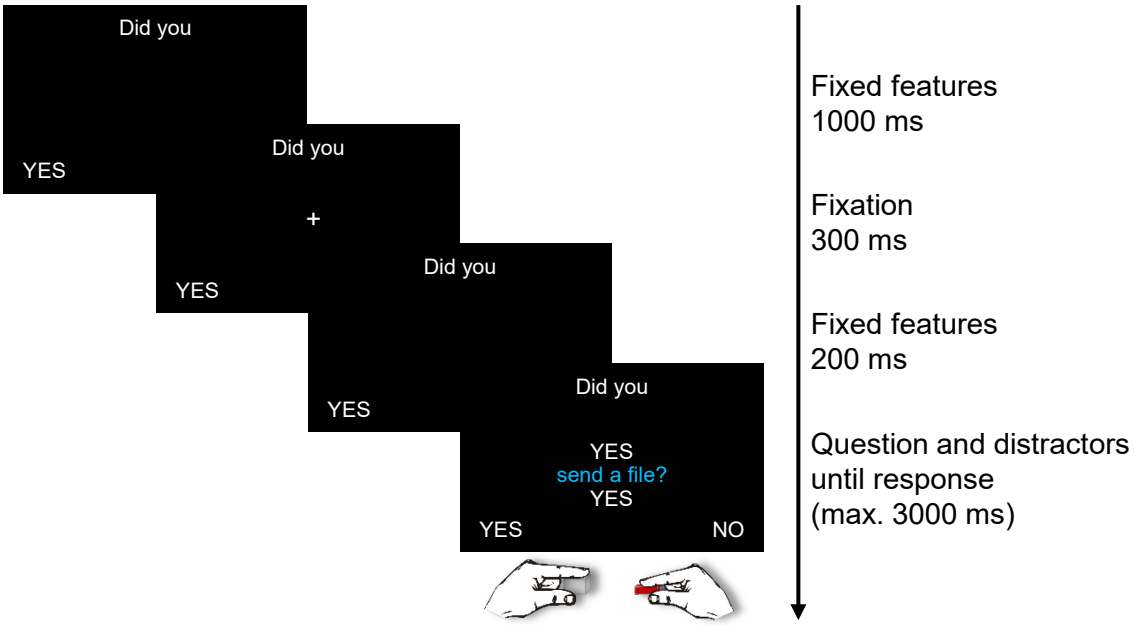

Figure 20 | Trial procedure of the inquiry in Experiment 10 and 11. Participants responded to routine and mission questions with *yes* or *no*. The font color of the question indicated whether participants should give an honest or an opposite to honest response. Distractors were either *yes* or *no* and like the responses, the distractors either did or did not correspond to the participants' actual experiences in the mission. For legibility, question, distractors, and response labels are not drawn to scale.

The combination of 2 question types (routine vs. mission; with 10 questions each) × 2 activity-response correspondence (corresponding vs. noncorresponding) × 2 activity-distractor correspondence (corresponding vs. noncorresponding) resulted in 80 trial types. Participants, therefore, went through nine blocks with 80 trials each, during which each combination was presented once. Participants could take self-paced breaks in between blocks.

Data treatment and analyses | The first block served as practice and was thus excluded from all statistical analyses, as was the first trial of each block. Trials that entailed question repetitions were excluded (3.8%). Trials in which participants gave an early response during fixation, did not respond, or responded with any other key than *D* or *K*

(2.3%), were excluded prior to computing and analyzing error rates. All erroneous trials were excluded before analyzing RTs, as were outliers (2.0%).

Error rates and RTs were examined in two separate 2 × 2 × 2 ANOVAs with the within-subjects factors question type (routine vs. mission), activity-response correspondence (corresponding vs. noncorresponding) and activity-distractor correspondence (corresponding vs. noncorresponding). In case of significant three-way and two-way interactions, we conducted separate 2 × 2 ANOVAs and two-tailed paired *t*-tests, respectively. Table 15 in Appendix 2 shows the mean error rates and RTs, computed separately for each combination of the three factors.

5.4.2. Results

Error rate │ The main effects of question type and activity-response correspondence were not significant (see Figure 21), *F*s < 1, but the two-way interaction of both factors was significant, $F(1, 43) = 51.67$, $p < .001$, $\eta_p^2 = .55$. Non-corresponding responses were more error-prone than corresponding responses to routine questions, $t(43) = 6.16$, $p < .001$, $d_z = 0.93$, but a reversed correspondence effect emerged for mission questions, $t(43) = -6.05$, $p < .001$, $d_z = -0.91$. The main effect of activity-distractor correspondence was not significant, $F(1, 43) = 3.23$, $p = .080$, $\eta_p^2 = .07$, however, the two-way interaction of activity-distractor correspondence and question type was significant, $F(1, 43) = 12.42$, $p = .001$, $\eta_p^2 = .22$, as responding was more accurate with corresponding than with non-corresponding distractors for routine questions, $t(43) = 3.93$, $p < .001$, $d_z = 0.59$, whereas distractors had no effect on mission questions, $t(43) = -1.63$, $p = .111$, $d_z = -0.25$. None of the remaining interactions were significant, *F*s < 1.

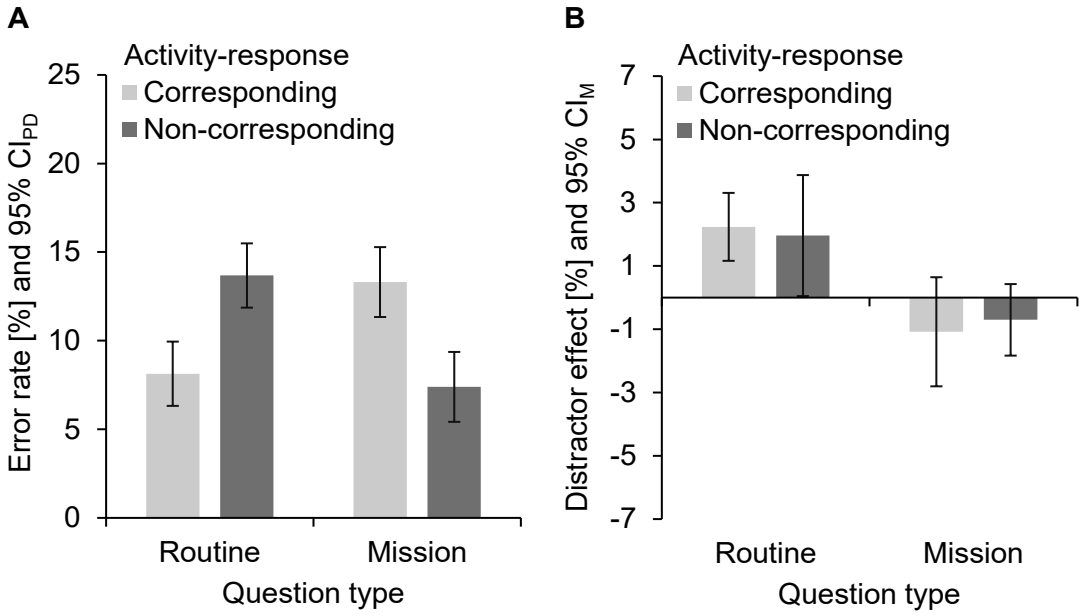

Figure 21 | Results of the analysis of error rates in Experiment 10. Mean error rates (A) and mean activity-distractor correspondence effects on error rate (B) are plotted as a function of activity-response correspondence and question type. Error rates of noncorresponding distractors were subtracted from error rates of corresponding distractors to compute distractor effects. Error bars represent the 95% confidence interval of paired differences (CI_{PD}), computed separately for routine and mission questions (A) whereas error bars around distractor effects show the 95% confidence interval around each of the four means (CI_M; B).

Response times | Responses to mission questions were slower than responses to routine questions (see Figure 22), $F(1, 43) = 44.13$, $p < .001$, $\eta_p^2 = .51$. The main effect of activity-response correspondence was not significant, $F < 1$, but the two-way interaction between activity-response correspondence and question type was significant, $F(1, 43) = 186.39$, $p < .001$, $\eta_p^2 = .81$, as responses to routine questions showed a typical correspondence effect, $t(43) = 13.19$, $p < .001$, $d_z = 1.99$, and a reversed correspondence effect was evident for mission questions, $t(43) = -12.11$, $p < .001$, $d_z = -1.83$. The main effect of activity-distractor correspondence was significant, $F(1, 43) = 11.20$, $p = .002$, $\eta_p^2 = .21$, but was further qualified by question type as mirrored in a significant two-way interaction of both factors, $F(1, 43) = 8.26$, $p = .006$, $\eta_p^2 = .16$. Responding to routine questions was easier with corresponding than with non-corresponding distractors, $t(43) = 4.94$, $p < .001$, $d_z = 0.74$, but distractors did not affect responding to mission questions, $t(43) = 0.74$, $p = .465$, $d_z = 0.11$. None of the remaining interactions were significant, Fs < 1.

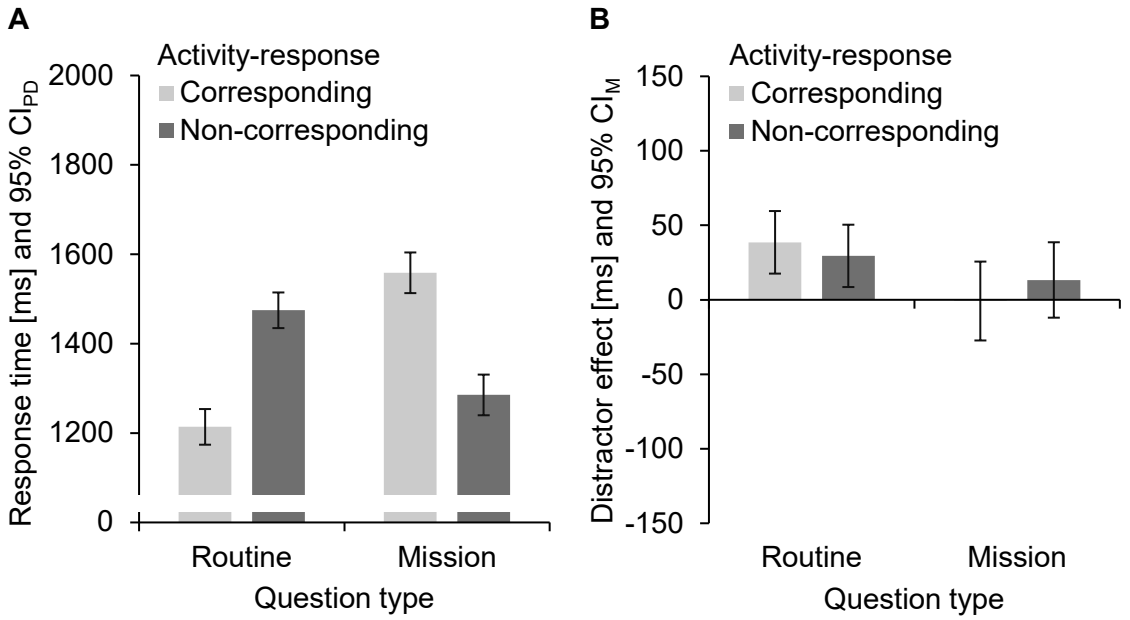

Figure 22 | Response time results of Experiment 10. Mean response times (A) and mean activity-distractor correspondence effects on response time (B) are plotted as function of activity-response correspondence and question type. Response times of noncorresponding distractors were subtracted from response times of corresponding distractors to compute distractor effects. Error bars represent the 95% confidence interval of paired differences (CI_{PD}), computed separately for routine and mission questions (A) whereas error bars around distractor effects show the 95% confidence interval around each of the four means (CI_M; B).

5.4.3. Discussion

Experiment 10 implemented corresponding and noncorresponding distractors in an inquiry with responses that did or did not correspond with participants' actual experiences to examine the underlying processes of dishonest responding with (mission questions) and without (routine questions) a false alibi. Replicating the results of the preceding experiments, noncorresponding responses to mission questions were delivered more easily than corresponding responses with a false alibi. In line with the literature, a traditional intention effect was again found in absence of a false alibi (e.g., Debey et al., 2012; Foerster, Wirth, Kunde et al., 2017; Spence et al., 2001). Distractor effects in routine questions also corroborated previous observations as distractors that corresponded with actual experiences facilitated responding in comparison to noncorresponding distractors (Debey et al., 2014). However, distractors did not affect responding to mission questions where participants had a false alibi.

This pattern of results suggests that false alibis have a potent impact on the automatic retrieval of responses as the usually observed distractor effect vanishes. The absence of any distractor effects despite large reversed activity-response correspondence effects indicates that false alibis do not simply substitute the noncorresponding response as an automatic default. A thorough discussion of this finding and potential theoretical explanations are postponed to the General Discussion, to establish with the following experiment whether the impact of false alibis on response and distractor correspondence effects do not stem from potential artifacts of the mission procedure.

5.5 Experiment 11

In Experiment 11, we used the same mission and inquiry as in the former experiments. Participants also still learned about the actions they were not to perform but, crucially, they were asked to admit which actions they actually did and did not perform in the inquiry. Accordingly, Experiment 11 aimed at establishing (a) that the observed reversed activity-response correspondence effects of the mission questions of the previous experiments were not an artifact of specific experimental parameters and (b) whether facilitating effects of corresponding distractors would emerge for mission questions without a false alibi. We hypothesized that without a false alibi, corresponding responses should be delivered faster and with fewer errors than noncorresponding responses in routine *and* mission questions. Likewise, corresponding compared to noncorresponding distractors should also facilitate responding for both question types.

5.5.1. Method

Participants and overall procedure │ Forty-four participants (age: M = 21.6, SD = 3.55; 37 female; 39 right-handed) of a sample size of 48 could be considered for statistical analyses. The sample size was based on the same criteria as Experiment 10, and four participants were excluded for the same criteria as in the former experiments. All participants gave written informed consent and received course credit as compensation. The procedure of Experiment 11 was almost the same as in Experiment 10, except that it featured slightly different instructions in the mission (see Appendix 4). Participants went through the same actions of the mission as the former participants. They were told, however, that most participants did five other actions and that these participants had to

conceal which action they engaged in. In contrast, their task was to accurately report in the upcoming inquiry which action they had performed in the mission.

Data treatment and analyses | The first block served as practice and was thus excluded from all statistical analyses as was the first trial of each block. Trials that entailed question repetitions were excluded (3.9%). Trials in which participants gave an early response during fixation, did not respond, or responded with any other key than *D* or *K* (2.2%), were excluded prior to computing and analyzing error rates. All erroneous trials were excluded before analyzing RTs. Trials with RTs that deviated more than 2.5 *SD*s from the respective cell mean were eliminated as outliers (2.1%).

Error rates and RTs were examined in two separate 2 × 2 × 2 ANOVAs with the within-subjects factors question type (routine vs. mission), activity-response correspondence (corresponding vs. noncorresponding) and activity-distractor correspondence (corresponding vs. noncorresponding). In case of significant three-way and two-way interactions, we conducted separate 2 × 2 ANOVAs and two-tailed paired *t*-tests, respectively. Table 15 in Appendix 2 shows the mean error rates and RTs, computed separately for each combination of the three factors.

5.5.2. Results

Error rates | Errors were more frequent in routine questions than in mission questions (see Figure 23), $F(1, 43) = 10.86$, $p = .002$, $\eta_p^2 = .20$. Non-corresponding responses were more error-prone than corresponding responses, $F(1, 43) = 66.25$, $p < .001$, $\eta_p^2 = .61$. The two-way interaction between activity-response correspondence and question type was significant, $F(1, 43) = 9.33$, $p = .004$, $\eta_p^2 = .18$. Non-corresponding responses were less accurate than corresponding responses for both question types but the effect was larger for mission questions, $t(43) = 7.53$, $p < .001$, $d_z = 1.14$, than for routine questions, $t(43) = 5.68$, $p < .001$, $d_z = 0.86$. More errors were made with non-corresponding than with corresponding distractors, $F(1, 43) = 16.08$, $p < .001$, $\eta_p^2 = .27$. The two-way interaction of activity-response correspondence and activity-distractor correspondence was significant, $F(1, 43) = 4.14$, $p = .048$, $\eta_p^2 = .09$, as the activity-distractor correspondence effect was stronger when activity and response did not correspond, $t(43) = 3.70$, $p = .001$, $d_z = 0.56$, than when they corresponded, $t(43) = 2.16$,

p = .036, d_z = 0.33. None of the remaining interactions were significant, Fs < 1.16, p ≥ .234.

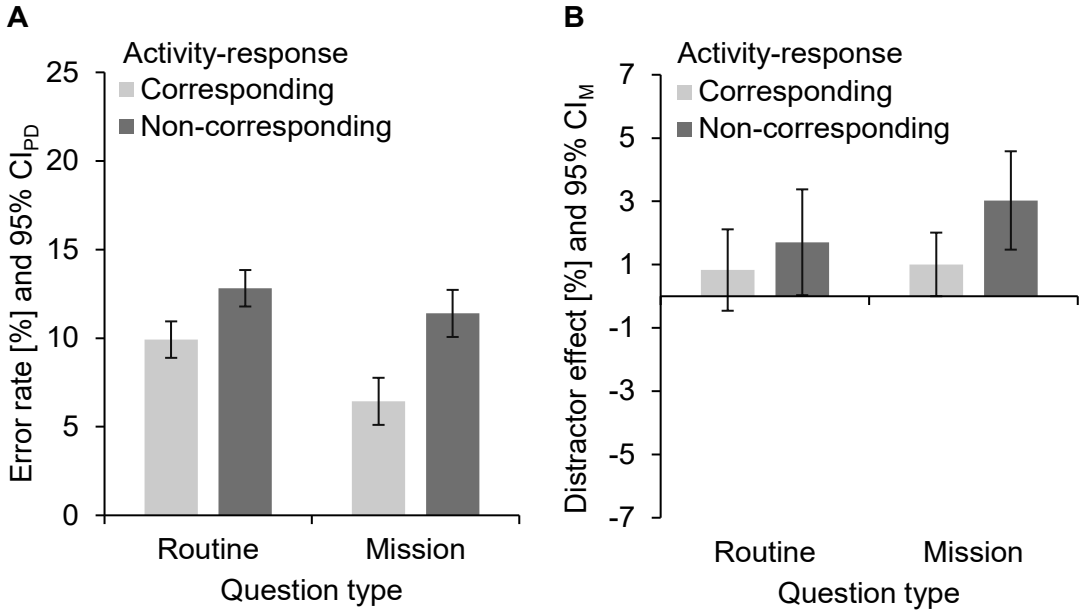

Figure 23 | Results of the analysis of error rates in Experiment 11. Mean error rates (A) and mean activity-distractor correspondence effects on error rate (B) are plotted as function of activity-response correspondence and question type. Error rates of noncorresponding distractors were subtracted from error rates of corresponding distractors to compute distractor effects. Error bars represent the 95% confidence interval of paired differences (CI_{PD}), computed separately for routine and mission questions (A) whereas error bars around distractor effects show the 95% confidence interval around each of the four means (CI_M; B).

Response times | Responses to mission questions were slower than responses to routine questions (see Figure 24), $F(1, 43) = 27.15$, $p < .001$, $\eta_p^2 = .39$. Non-corresponding responses were slower than corresponding responses, $F(1, 43) = 142.42$, $p < .001$, $\eta_p^2 = .77$. This effect was larger in mission questions, $t(43) = 12.95$, $p < .001$, $d_z = 1.95$, than in routine questions, $t(43) = 9.71$, $p < .001$, $d_z = 1.46$, as indicated by a significant two-way interaction between activity-response correspondence and question type, $F(1, 43) = 5.51$, $p = .024$, $\eta_p^2 = .11$. Responses were slower when activity and distractor did not correspond relative to when they corresponded, $F(1, 43) = 12.08$, $p = .001$, $\eta_p^2 = .22$. None of the remaining interactions were significant, Fs ≤ 1.03, ps ≥ .315.

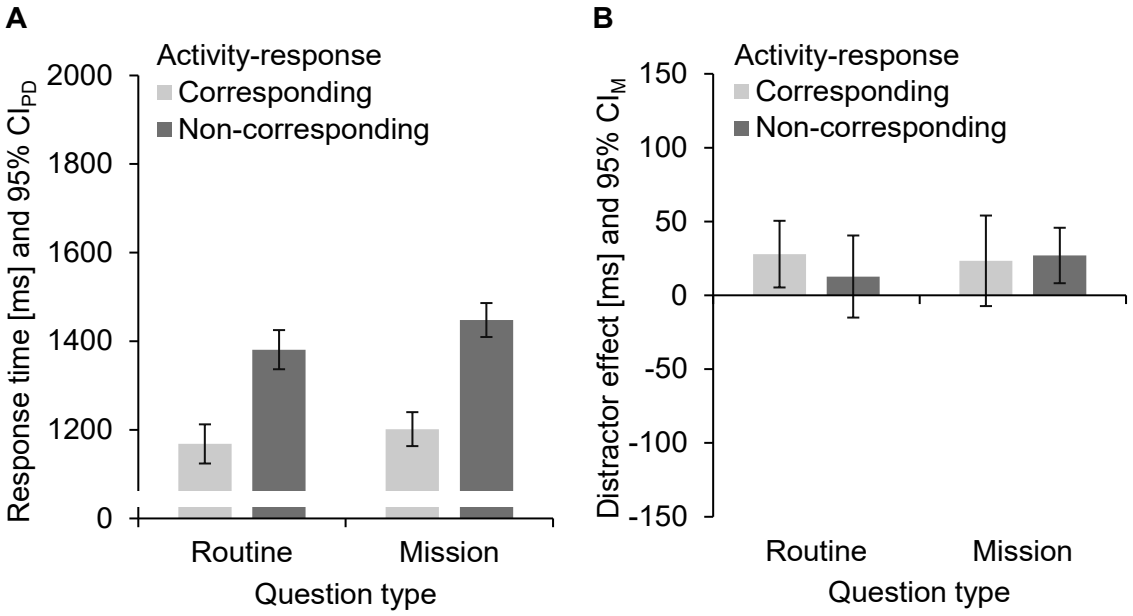

Figure 24 | Response time results of Experiment 11. Mean response times (A) and mean activity-distractor correspondence effects on response time (B) are plotted as function of activity-response correspondence and question type. Response times of noncorresponding distractors were subtracted from response times of corresponding distractors to compute distractor effects. Error bars represent the 95% confidence interval of paired differences (CI_{PD}), computed separately for routine and mission questions (A) whereas error bars around distractor effects show the 95% confidence interval around each of the four means (CI_M; B).

5.5.3. Discussion

In Experiment 11, participants performed actions in a mission and learned about other actions that were allegedly performed by most of the other participants. In an upcoming inquiry, they were to admit which actions they actually did and did not perform in the mission (mission questions) and on that day (routine questions) by responding honestly and oppositely. In line with our hypotheses and previous findings in the literature, responding in correspondence with actual experiences was easier than responding non-correspondingly in routine and mission questions (e.g., Debey et al., 2012; Foerster, Wirth, Kunde et al., 2017; Spence et al., 2001). Furthermore, corresponding distractors facilitated corresponding and non-corresponding responses (Debey et al., 2014). This suggests that mission questions activated their honest response, which had to be overcome to respond dishonestly, and that this process could be captured by competing response activation of distractors.

The former experiments featured opposite but in most cases equally sized activity-response correspondence effects in routine and mission questions. A potential conclusion from this pattern might have been confounded by the fact that mission questions were always accompanied by a false alibi whereas routine questions were not. Experiment 11 revealed that without a false alibi, the effect of more difficult responses in noncorresponding than in corresponding trials was even larger in mission than in routine questions. This is true for speed and accuracy alike. Accordingly, the reversing effect of false alibis appears all the more striking.

5.6 General discussion

In the present experiments, participants were led to believe that they had to fulfill a special mission that entailed to lie in an upcoming inquiry. The inquiry featured an instructed intention paradigm on a personal computer (Experiment 8, 10 and 11) or on an iPad (Experiment 9). Cues signaled participants to respond honestly or oppositely with *yes* and *no* to routine questions and to questions regarding their mission. To fulfill their special mission, participants were given a false alibi that specified a series of alternative actions and they were to answer according to this alibi in the inquiry. That is, participants had to lie in the presence of honest cues and had to be honest in the presence of opposite cues when responding to mission questions in Experiments 8 to 10. In Experiment 11, participants were not provided with a false alibi but they simply were to tell the truth in the presence of honest cues and had to be dishonest in the presence of opposite cues when responding to mission questions to control for potential confounding effects from different questions. Distractors were used in Experiment 10 and 11 to scrutinize how false alibis change dishonest processing.

5.6.1. The power of false alibis

Responses to routine questions in all experiments and to mission questions in Experiment 11 replicated common findings on dishonesty in the literature, as response initiation and execution took longer and were more error-prone when actual experiences and responses did not correspond as compared to corresponding responses (e.g., Debey et al., 2012; Foerster, Wirth, Kunde et al., 2017; Spence et al., 2001). Experiment 9 also revealed an impact of the corresponding response on movement trajectories when giving a noncorresponding response. Movement trajectories were more strongly attracted toward

the competing response side in noncorresponding trials than in corresponding trials for routine questions, which replicates and extends previous observations from studies on action dynamics during lying (Duran et al., 2010) and rule violations (Pfister, Wirth, Schwarz, Steinhauser et al., 2016; Wirth, Pfister, Foerster et al., 2016). Furthermore, distractors that corresponded with actual experiences facilitated honest and dishonest responding to routine questions as compared to noncorresponding distractors, again replicating existing evidence in the literature (Debey et al., 2014). The results of the mission questions from Experiments 8 to 10, however, came unexpectedly: Responses in accordance with the false alibi were, in fact, faster and more accurate than responses according to the participants' actual experience during the mission (i.e., corresponding responses). In particular, it was easier for participants to negate than to affirm activities they had actually experienced in the mission and it was easier to affirm than to negate activities they had not performed. The correspondence effect was also reversed in movement trajectories with a stronger attraction toward the competing response in corresponding trials.

These counterintuitive observations of faster and more accurate noncorresponding responses relative to corresponding responses suggest that false alibis change the typical honest default response to a dishonest default response. As such, these dishonest responses would be retrieved directly and automatically. However, this assumption predicts faster and more accurate responding with noncorresponding distractors when participants received a false alibi, and this prediction was not confirmed in Experiment 10.

Three theoretical possibilities suggest themselves to account for this result. First, participants might have succeeded in building a vivid mental model of the alibi actions, resulting in a strong representation of those actions (for a primer on mental models, see, e.g., Johnson-Laird, 2004). However, retrieval of the appropriate mental model of such actions could be effortful and time-consuming compared to honest response retrieval. Accordingly, false alibis could actually substitute the honest default response with the dishonest default response, but the retrieval of that dishonest default would take sufficient time for any distractor effects to level off due to a mandatory built-up of the mental model. Second, false alibis could implement an automatic inhibition of the corresponding response rather than an activation of the noncorresponding response. The sequence of initial activation and automatic inhibition would impair performance for corresponding

responses whereas it does not necessarily affect noncorresponding responses. Third, noncorresponding responses might be consistently activated in addition to the proposed automatic inhibition of the honest response on each trial. All three proposed mechanisms would produce reversed correspondence effects without necessarily showing effects of distractors. Absent distractor effects are in line with all three accounts by assuming that distractors were processed while the proposed automatic steps were still at work. Varying the temporal relation of question and distractors could give insight into whether the default response changed to the noncorresponding response (with a more effortful retrieval than for honest default responses) or corresponding responses were automatically inhibited and in the latter case, whether noncorresponding responses activation follows this inhibition process. Conceivably, dishonest processing could also operate less consistently with false alibis than without false alibis and entail a mixture of these mechanisms, thus, failing to show consistent distractor effects.

In any case, the present results raise the question of how basic the cognitive basis of dishonesty, as described in the introduction, really is. To recapitulate, contemporary models of the cognitive processes underlying dishonesty assume that the honest response to a question is retrieved automatically. When a question is always answered dishonestly, a stimulus-response association is built up that can be easily derived and makes lying as easy as being honest (Walczyk et al., 2009, 2012). An equal proportion of honest and dishonest responding to each question prevents such a creation of stimulus-response associations, like in the current experiments, where corresponding and noncorresponding responses differed in several behavioral measures. The present evidence suggests that false alibis change the automatically retrieved response and/or render the inhibition of honest default responses more efficient, thus, facilitating dishonest responding and interfering with honest responding.

In the current experiments, participants committed a considerable amount of errors which also led to the exclusion of several participants. Instructed lie paradigms without false alibi instructions already come with rather high error rates and participants frequently express that the task is difficult (e.g., Debey et al., 2014; Foerster, Wirth, Kunde et al., 2017 Debey et al., 2014). The applied character of our task did not allow giving error feedback in case of wrong response commissions. Accordingly, participants with difficulty to understand the inquiry task could not be corrected during the inquiry. Many of the

excluded participants failed to meet our predefined inclusion criteria by a considerable margin with error rates of at least 70% in one of the experimental cells of the mission questions. So, these participants responded consistently wrong and did not follow the alibi instruction but responded similarly to routine and mission questions.

One way to minimize data exclusion in future studies could be the use of a more accessible cover story, possibly combined with rewards to enhance motivation. This could give insight into whether false alibis are implemented more easily and successfully with higher motivation or whether false alibis always impose considerable difficulty. A challenge for researchers will be to establish a level of motivation in the laboratory that is comparable to the motivation of alleged criminals.

5.6.2. Implications for lie detection

The strong effects of false alibis cast doubt on whether cognitive tests may represent a feasible tool for lie detection, because of the clear-cut correspondence effects (standard or inverted) on the group level.[11] The present results further provide more insight into the conclusions of a similar, real-world investigation with a convicted woman (Spence, Kaylor-Hughes, Brook, Lankappa, & Wilkinson, 2008). The woman was already convicted of poisoning her child when she took part in the inquiry of the researchers. The authors created an *instructed intention paradigm* with questions relating to her case. According to a mapping rule, the woman responded to each question honestly, that is, analogously to her account of the events, or dishonestly, that is, analogously to the accusers' account of events. Enhanced RTs and hemodynamical activity in ventrolateral prefrontal and anterior cingulate regions emerged for responses in accordance with the accusers' account of the events. Based on the evidence in the literature that existed by then, the data supported the assumption that the woman was innocent of the crime. The present data shows, however, that a dishonest response can be well internalized and generated more promptly than an honest response. Hence, the *instructed intention paradigm* appears to be

[11] Large effects on the group level (like the correspondence effects here) are a necessary precondition for successful lie detection, but classification accuracy on a participant- and item-specific level has to be assessed to evaluate how a method is actually suited for lie detection (Franz & Luxburg, 2015).

impractical for lie detection because innocent and guilty persons might produce the same pattern of results.

If *instructed intention paradigms* do not seem to have particular utility for lie detection – is there an alternative approach to lie detection using RTs or similar behavioral parameters? One solution that has recently been proposed capitalizes on congruency effects as measured via the *autobiographical implicit association test* (Sartori et al., 2008). In this procedure, participants usually indicate with button presses whether sentences describe innocence (e.g., "I bought the CD") or guilt (e.g., "I stole the CD") or whether sentences are true (e.g., "I am reading a scientific manuscript") or false (e.g., "I am swimming in the Red Sea"). Participants have to categorize sentences of all four categories in a random sequence in each block. Crucially, in one block, innocence and truth share one response key and guilt and falsity share the other response key whereas in a second block, guilt and truth, and innocence and falsity share response keys, respectively. Innocent participants who did not steal a CD would respond faster in the first than in the second block whereas the opposite would be true for guilty participants (e.g., Agosta, Ghirardi, Zogmaister, Castiello, & Sartori, 2011).

Though there are indeed several promising reports, the validity and robustness of this measure are still under discussion. Countermeasures were identified that diminish the detection accuracy of the *autobiographical implicit association test* as the instruction of speeded responses and training in those blocks where innocent persons respond faster than guilty persons (Hu, Rosenfeld et al., 2012), and instructions to slow down responses in blocks where innocent persons respond slower than guilty persons (Verschuere, Prati, & Houwer, 2009). The results of the present study suggest false alibis as a potential countermeasure by representing performed actions strongly as being not performed and not-performed actions as being performed.

Another lie detection method, the concealed information test, relies on the fact that a crime stimulus, among more frequent but comparable neutral stimuli, is significant for the person who committed the crime but indistinguishable for innocent persons (e.g., Ben-Shakhar & Elaad, 2003). Accordingly, the crime stimulus produces detectable signatures because of significance only in guilty persons. False alibi stimuli could produce similar effects as crime stimuli in the concealed information test. As such, false alibis might be identified as true knowledge of the examined person. However, it seems less plausible that crime stimuli

become inseparable of neutral stimuli by imagining to not have interacted with those crime stimuli.

5.6.3. Conclusion

The current study adapted the *instructed intention paradigm* to observe alibi effects on lying performance in a forensically applicable design. Participants had to implement a false alibi by pretending to have engaged in plausible alibi actions and denying the involvement in actually experienced activities. They succeeded to a level where the fake story appeared as being true in all relevant measures. The data suggest that mere instruction can cause either dishonesty instead of honesty to become the default response (but a weaker default than the honest one) or dishonest processing to become the default in the sense that a question still triggers an honest response which is automatically inhibited, thus, facilitating (automatic) dishonest response retrieval. These mechanisms could also operate simultaneously. In all cases, honest responding would be more effortful – contrary to the usually assumed cognitive processes operating during honest and dishonest responding.

Theoretical integration: Disentangling dishonesty

6 Overcoming the truth

This thesis followed the logic of a central claim – with a focus on the activation of the truth and the need to overcome this initial response tendency for successful dishonest responding – *the truth will out*. Therefore, the experiments merged established paradigms from the literature on lying with methodology from sensorimotor approaches and cognitive control frameworks. The results provide valuable insights for our understanding of the cognitive architecture of responding, exposing the enduring process of overcoming the truth and its impressive flexibility. The following section provides a theoretical update based on the findings of this thesis complemented by suggestions to expand on these insights.

6.1 A flexible take on the truth

In line with the literature, behavioral differences between truth-telling and lying were of impressive size in the current experiments (see Suchotzki et al., 2017 for a recent meta-analysis). The *activation-decision-construction-action theory* is able to explain these differences with the pervasive impact of the truth on the cognitive processing of dishonest responses (Walczyk et al., 2014; for a former version of the theory, see Walczyk et al., 2003). The experiments of Chapter II provide strong empirical backup for this claim with a specific description of the roots of the involved processes. In particular, overcoming the truth in dishonest responding exerts its strongest impact on resource-limited processes of response selection, which is well in line with the proposed *action* component of the updated theory (Walczyk et al., 2014). Furthermore, this *action* component also proposes that liars monitor their behavior more thoroughly, and the data of this thesis suggest that this process is also capacity-limited and a consequence of the cognitive conflict in dishonest responding. However, the results also revealed early effects of dishonest responding on information processing that might be due to differences in the automatic activation or inhibition as well as threshold levels of responses between honest and dishonest responding. Further research is required to specify this process, which would then also need to be represented in cognitive theories of dishonest responding.

While the first set of experiments, thus, makes a strong case for an enduring and effortful process of overcoming the truth in dishonest responding, the experiments of Chapter II demonstrate means to diminish these large costs of dishonest responding. For one, the results support the assumption that dishonest responding triggers increased cognitive control, improving the selection of an upcoming dishonest response relative to an honest response. Second, the internalization of a false alibi even flipped the costs, rendering dishonest responses more efficient than honest responses in the current study. False alibis appear to have the power to introduce a qualitative change of the processing of honest and dishonest actions. A recent study corroborates these findings of this thesis in that performance differences between honest and dishonest responses vanished with the instruction of a false alibi (Suchotzki, Berlijn, Donath, & Gamer, 2018; see also Dhammapeera, Hu, & Bergström, 2019 for an examination of false alibis in the context of the *autobiographical implicit association test*). Notably, the impact of the false alibi differed between both studies, and the mechanisms behind these variations need to be addressed in future studies. Although the four components of the *activation-decision-construction-action theory* allow for a different order of its components than is suggested by the name, the theory does not propose that the *activation* component might be skipped entirely or that the automatic activation of a lie either replaces or happens in concert with the activation of the truth. However, especially the experiments on the impact of false alibis strongly suggest such a radical change in cognitive processing. As such, the observed flexibility of dishonest responding also introduces a huge practical problem. How should we detect lies based on the cognitive processing of dishonest responses if these processes adapt flexibly and thus produce very different measurable effects?

6.2 Efficiency and preference

If there are means to make lying more efficient, should that not also increase the willingness of agents to make use of these mechanisms? To my knowledge, there is no direct evidence for this assumption, but only correlational data on the relationship of both measures. For one, lying performance in an *instructed intention paradigm* showed a small correlation with self-reported daily lying during the last 24 hours (Debey, De Schryver et al., 2015). Similarly, smaller performance costs of rule violations compared to rule-based actions coincided with more rule violations; both measures where obtained in a laboratory task where participants navigated a virtual cyclist through a maze adhering to traffic rules

or violating them (Pfister et al., 2018; Pfister, Wirth, Schwarz, Steinhauser et al., 2016). Furthermore, even though convicted rule-breakers showed similar detrimental effects of rule violations as control participants in the planning phase of their actions, they did not show such a pattern when executing these actions whereas control participants continued to show that effect (Jusyte et al., 2017). These results could either reflect that agents who easily overcome the truth or a rule also make more use of lies and rule violations, but at the same time, they could also mean that frequent lying or rule violations render this behavior easier. Insight for the latter causal direction comes from the current thesis and the existing literature: Although Experiment 5 did only show an impact of the immediately preceding (dis)honest action but not of the overall frequency of dishonest responding on current dishonest responding, the latter impact has been shown elsewhere (Van Bockstaele et al., 2015). The rehearsal of specific lies also improves the generation of these lies (Hu, Chen et al., 2012; Walczyk et al., 2009, 2012).

The other direction, namely the impact of the efficiency of overcoming the truth on the probability of lying has not been put to a direct test, yet. One way to address this issue might be the manipulation of the strength of true memory traces by varying the number of repetitions of laboratory activities that participants are later asked about in a computerized inquiry. The honest response to a question might be activated more strongly for activities that had been repeated relatively often. A forced-choice phase of responding could validate whether dishonest responding becomes indeed more difficult with more repetitions of the inquired activity. A free-choice phase of responding would then probe for the frequency of dishonest responding depending on the number of repetitions of the activities. People seem to hold lay theories about the effort of honest and dishonest actions, associating honesty with effortlessness and dishonesty with effort (Lee, Ong, Parmar, & Amit, 2019). In the context of the proposed experiment, it would be interesting to scrutinize whether agents can adapt this intuition, anticipating that a stronger representation of the truth interferes more strongly with dishonest responding without having ever uttered a lie about this activity.

6.3 Inhibition or decay

A huge leap in the efficiency of dishonest responding occurred with the implementation of a false alibi in Experiments 8 to 10 of this thesis, and also in recent experiments of other researchers – even though the size of the impact differs substantially between studies (Suchotzki et al., 2018). Relatedly, although dishonest responding appears to be sensitive to changes in adaptive control, Experiment 6 and 7 found no transfer of cognitive control settings between dishonest responding and other similar behavioral conflicts. A crucial step toward a deeper understanding of how the implementation of false alibis or the experience of conflict in dishonest responding change processing of future dishonest responses seems to be an even more precise description of how agents overcome honest response activation in the first place. There is direct evidence for an initial activation of the truth (Experiment 10 and 11 of the current thesis; Debey et al., 2014), and there appears to be a continued influence of the honest response even during or possibly even after response execution (Experiment 3 and 4 of the current thesis; Duran et al., 2010).

However, it is not clear, yet, whether truth activation becomes weaker than dishonest response activation in a rather passive process of fading activation or whether there is active inhibition involved. The latter assumption received theoretical support (Walczyk et al., 2003, 2014) and has also been examined empirically with a wide range of methods. Researchers correlated intention effects with the inhibition ability measured in a stop-signal task (Debey, De Schryver et al., 2015) and assessed the distribution of intention effects in response times for participants with supposedly different inhibitory ability (Caudek, Lorenzino, & Liperoti, 2017; Debey, Ridderinkhof, Houwer, Schryver, & Verschuere, 2015). A recent study approached this issue via negative priming, assessing the accessibility of preceding truthful information in sequences of (dis)honest responding (Aïte, Houdé, & Borst, 2018). The results are mixed and the effects that support the presence of inhibition still leave room for alternative interpretations. An innovative method that allows for a clear-cut decision whether the truth is overcome by active means of inhibition or merely decays passively while the dishonest response becomes activated is needed. If there is inhibition involved, it might target conceptual or motoric representations of the truth. Getting more insight into the underlying mechanisms of unprepared lies would be the first step toward understanding how false alibis or experiences of dishonest

responding affect the efficiency of future dishonest responses, possibly through a boost of dishonest response activation or honest response inhibition, or both.

7 Accessing dishonest actions

A promising approach for a better understanding of how the truth is overcome in dishonest responding might come from the research field on learning and memory and research endeavors that bridge these branches with cognitive control theories. There are two particularly interesting perspectives. One highlights the instant connection of questions with their (dis)honest responses and the relation of these bindings to the intention of responding. The other concerns explicit remembering of dishonest and honest responses, with a special focus on the consequences of dishonest responding for the accuracy of remembering the underlying truth. Both perspectives suggest that when humans interact (dishonestly) with their environment, mnemonic processes already operate in the background, changing not only representations of dishonest but also of honest content, ultimately paving the way for a direct and smooth retrieval of dishonest responses. The following two sections elaborate on the two perspectives in more detail.

7.1 Hierarchical associations

Experimental data indicates that knowledge about having lied to a question is established by a direct association of the question with the dishonest intention of the response (Koranyi, Schreckenbach, & Rothermund, 2015). In this study, participants responded honestly and dishonestly to questions in an oral interview. The questions of that interview later served as primes preceding the target words *honest* and *dishonest* that had to be categorized by keypress. The presentation of prime questions that had been answered dishonestly in the preceding interview facilitated the categorization of the target word *dishonest* compared to the target word *honest*. The formation of such bindings between the dishonest intention and a question is one ingredient for telling consistent lies while the establishment of a link between a question and the specific lie appears just as important. Binding of a stimulus to a response is a hallmark of human action control; the process of associating a stimulus to its response and retrieving this response upon encountering this stimulus again is effortless, rendering action control particularly efficient (e.g., Henson, Eckstein, Waszak, Frings, & Horner, 2014; Hommel, 1998b; Logan, 1988).

The same associative mechanism seems to be responsible for the link between a dishonest intention to its question and the link between a question and its response, and both bindings appear to be connected hierarchically (Koranyi et al., 2015; Pfeuffer, Pfister, Foerster, Stecher, & Kiesel, 2019). In particular, empirical data suggests that a question only retrieves its previously associated response if the current intention matches the intention that it had have been associated with before (Pfeuffer et al., 2019). Responding honestly and dishonestly in a computerized size categorization of daily objects led to the retrieval of the same motor response upon repeating a previously presented object only if participants were instructed to respond with the same intention as in the preceding encounter of that stimulus (i.e., in honest – honest or dishonest – dishonest sequences). In other words, the retrieval of established stimulus-response associations is not mandatory upon stimulus encounter but guided by intentions. Interestingly, a detailed look at the data also reveals that the association between a question and a dishonest response is as strong as the association between a question and an honest response because retrieval effects were of similar size for both intentions (Experiments 2 and 3 in Pfeuffer et al., 2019).

This mechanism of retrieving honest and dishonest intentions and the appropriate responses hierarchically from stimulus presentations might lie at the heart of shifting dishonest responding from an effortful process of overcoming the truth to a direct retrieval of responses through rehearsal, enabling not only easier but also more consistent lying (Pfeuffer et al., 2019). Crucially, other studies also demonstrated that mere instructions can establish bindings between stimuli and responses that are later retrieved upon stimulus encounter (e.g., Cohen-Kdoshay & Meiran, 2009; Gollwitzer, 1999; Kunde, Kiesel, & Hoffmann, 2003; Wenke, Gaschler, & Nattkemper, 2007). Similarly, the association of stimuli and responses through instruction might also play a key role in the modulation of dishonest responding by false alibis. A more fine-grained manipulation of the intensity or the number of repetitions of the instruction of a false alibi action, including related questions and responses, might shed light on whether these instructions establish stimulus-response bindings that conform to the false alibi and thus facilitate the retrieval of these responses.

7.2 Explicit memory

The preceding studies demonstrated that dishonest actions are accompanied by an integration of the dishonest intention, the lied-about content and the dishonest response, and that this binding process seems to be similar for honest and dishonest responding. An open question is whether these elements are also remembered equally well across intentions or whether they diverge in this regard, and a large body of research is devoted to this question (for a review, see Otgaar & Baker, 2018). Of particular interest for the current study is the proposal that the increased cognitive effort of constructing and delivering a dishonest response is advantageous for memorizing dishonest responses compared to honest responses (Besken, 2018). In this study, participants saw questions about general knowledge (e.g., “What is the capital of Germany?”) and were instructed to respond honestly to half of the questions and dishonestly to the other half of the questions with varying levels of free-choice and forced-choice responses across experiments. If dishonest responses were slower than honest responses, participants showed improved free recall of their dishonest than of their honest responses, although participants assumed that they would remember honest responses better than dishonest responses. Notably, participants just had to recall as many of their responses as possible in a random fashion. That is, participants just listed responses that they recalled without indicating whether these responses had been honest or dishonest. In fact, lying can lead to memory errors like mistaking false information as true or losing access to true information (for a review, see Otgaar & Baker, 2018). The severity of inaccurate memory depends on the form of lying, and these authors also propose that the amount of cognitive effort lies at the heart of these differences.

From the perspective of the current thesis, it would be interesting to follow up on the study of Besken (2018) and explore the impact of dishonest responding on the ability to remember the *true* account of events in more detail, thus, allowing to infer the cognitive processes involved in overcoming the truth in dishonest acts from yet another angle. In the study, participants had to respond truthfully to all questions at the end of the experiment to assess whether they knew the correct answers to the questions. In two out of three experiments, participants’ accuracy rates were similar for questions they had answered honestly and the ones that they had responded to dishonestly. Just one experiment showed lower accuracy rates for formerly dishonest than honest questions.

However, if participants knew the correct answer to a question, it seems unlikely that dishonest responding could change access to ingrained general knowledge easily. In line with that are the overall high accuracy rates, which suggest that the questions tackled content that was well memorized. In contrast, memorizing actions and their corresponding situation offer richer and more diverse information, providing much more opportunity to forget or confuse details. As such, dishonest responding might exert an impact on the memory of honest actions. This could be tested by introducing experimentally manipulated activities in the laboratory and inquiring about these details instead of general knowledge.

Considering the literature on cognitive control processes and its impact on memory encourages two opposing hypotheses – the act of overcoming the truth in dishonest responding might reduce or increase memory for the truth itself. On the one hand, honest information takes a functional role in dishonest responding since the dishonest response cannot be accessed directly (Debey et al., 2014). As such, truthful content might be even more activated in dishonest than honest responding. In consequence, dishonest responding might also improve the encoding of truthful content. In Stroop-like tasks, a study showed indeed better memory for target stimuli that appeared together with incongruent rather than congruent distractor stimuli (e.g., Krebs, Boehler, Belder, & Egner, 2015; Rosner, D'Angelo, MacLellan, & Milliken, 2015). The interpretation of this finding is that conflict enhances attention to the relevant stimulus dimension (cf., Botvinick et al., 2001) improving encoding and later recall of target stimuli that appeared with incongruent rather than congruent distractors. However, in contrast to the Stroop task, both the representation of the honest and dishonest response is functional at one point in dishonest responding.

Accordingly, it might be more likely that delivering a dishonest response entails inhibition of the truth (Walczyk et al., 2003, 2014), and especially if this inhibition targets not only motoric but also conceptual representations, this might reduce the ability to remember it later (cf., Anderson, 2003). For one, retrieving information from a cue that also easily triggers other information impedes access to these incidentally activated information, a phenomenon that is known as retrieval-induced forgetting, probably because inhibition prevents the retrieval of automatically activated but currently inappropriate information (e.g., Anderson, Bjork, & Bjork, 1994). Furthermore, several experiments demonstrated that stimuli that appeared in a no-go instead of a go trial or in

a trial where participants stopped a prepared response instead of executing it were remembered less well (Chiu & Egner, 2015). The authors argued that inhibition decreased perceptual processing and supported this hypothesis with additional experimental data. In the case of dishonest responding, decreased perceptual processing because of inhibition seems to be less likely (cf. Experiment 1 and 2 of this thesis), whereas inhibition might rather target existing memory traces (i.e., honest responses) as has been shown for retrieval-induced forgetting.

In a nutshell, observing improved memory for the true account of experienced activities after delivering a dishonest compared to an honest response about this event would suggest that dishonesty strengthens the encoding of truthful content despite delivering an alternative response. The opposite result would be indicative of inhibition as a means to overcome dominant truthful representations of experienced activities for dishonest responding. There is a thesis that examined the impact of dishonest responding on memory for actively committed actions in the laboratory with mixed results (Li, 2015). Some of the data shows worse memory for events after dishonest than honest responding while there is no difference at other times. Furthermore, dishonest responding led to similar, worse and also to better memory performance than when participants neither responded honestly nor dishonestly about an event. Given the current empirical evidence, one could conclude that dishonest responding relative to honest responding rather leads to weaker than stronger memory traces about experienced activities, pointing to inhibition processes. The divergent results on the memory performance for dishonest responding compared to not responding about these events might reflect that these inhibition processes still work hand in hand with a considerable activation of honest representations.

8 Concluding remarks

The insights from this thesis on the cognitive architecture of dishonest responding, and especially the role of the truth in this process, root in a joint consideration of different branches of research on cognitive psychology. Together, the results demonstrate a pervasive impact of the truth on information processing of dishonest responses that can, however, be mitigated by adaptations of cognitive control settings and through false alibis. At the same time, our broad understanding of other behavioral conflicts does not seem to apply one-to-one to dishonest responding, which clearly motivates specific examinations

of the cognitive consequences of lies. Still, the findings do not only inform our understanding of lying in particular, but also contribute to the domains of sensorimotor stages of information processing and cognitive control in general. Connecting theories and insights from different fields of psychology will be mandatory to disentangle the processes that contribute to dishonest behavior – a truly complex task considering the adaptivity and flexibility in lying.

× References

Agosta, S., Ghirardi, V., Zogmaister, C., Castiello, U., & Sartori, G. (2011). Detecting fakers of the autobiographical IAT. *Applied Cognitive Psychology*, *25*(2), 299–306. https://doi.org/10.1002/acp.1691

Aïte, A., Houdé, O., & Borst, G. (2018). Stop in the name of lies: The cost of blocking the truth to deceive. *Consciousness and Cognition*, *65*, 141–151. https://doi.org/10.1016/j.concog.2018.07.015

Allport, A. D., Styles, E. A., & Hsieh, S. (1994). Shifting intentional set: Exploring the dynamic control of tasks. In C. A. Umiltà & M. Moscovitch (Eds.), *Attention and performance: Vol. 15. Conscious and nonconscious information processing* (pp. 421–452). Cambridge, Mass: MIT Press.

Anderson, M. C. (2003). Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of Memory and Language*, *49*(4), 415–445. https://doi.org/10.1016/j.jml.2003.08.006

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(5), 1063–1087. https://doi.org/10.1037/0278-7393.20.5.1063

Arico, A. J., & Fallis, D. (2013). Lies, damned lies, and statistics: An empirical investigation of the concept of lying. *Philosophical Psychology*, *26*(6), 790–816. https://doi.org/10.1080/09515089.2012.725977

Ben-Shakhar, G., & Elaad, E. (2003). The validity of psychophysiological detection of information with the Guilty Knowledge Test: A meta-analytic review. *Journal of Applied Psychology*, *88*(1), 131–151. https://doi.org/10.1037/0021-9010.88.1.131

Bereby-Meyer, Y., & Shalvi, S. (2015). Deliberate honesty. *Current Opinion in Psychology*, *6*, 195–198. https://doi.org/10.1016/j.copsyc.2015.09.004

Besken, M. (2018). Generating lies produces lower memory predictions and higher memory performance than telling the truth: Evidence for a metacognitive illusion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(3), 465–484. https://doi.org/10.1037/xlm0000459

Bhatt, S., Mbwana, J., Adeyemo, A., Sawyer, A., Hailu, A., & Vanmeter, J. (2009). Lying about facial recognition: An fMRI study. *Brain and Cognition*, *69*(2), 382–390. https://doi.org/10.1016/j.bandc.2008.08.033

Bloomquist, J. (2010). Lying, cheating, and stealing: A study of categorical misdeeds. *Journal of Pragmatics*, *42*(6), 1595–1605. https://doi.org/10.1016/j.pragma.2009.11.008

Bond, C. F., & DePaulo, B. M. (2008). Individual differences in judging deception: Accuracy and bias. *Psychological Bulletin*, *134*(4), 477–492. https://doi.org/10.1037/0033-2909.134.4.477

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. https://doi.org/10.1037/0033-295X.108.3.624

Braem, S., Abrahamse, E. L., Duthoo, W., & Notebaert, W. (2014). What determines the specificity of conflict adaptation? A review, critical analysis, and proposed synthesis. *Frontiers in Psychology*, *5*, 1134. https://doi.org/10.3389/fpsyg.2014.01134

Braem, S., Bugg, J. M., Schmidt, J. R., Crump, M. J. C., Weissman, D. H., Notebaert, W., & Egner, T. (2019). Measuring adaptive control in conflict tasks. *Trends in Cognitive Sciences.* https://doi.org/10.1016/j.tics.2019.07.002

Braem, S., Verguts, T., Roggeman, C., & Notebaert, W. (2012). Reward modulates adaptations to conflict. *Cognition*, *125*(2), 324–332. https://doi.org/10.1016/j.cognition.2012.07.015

Bugg, J. M. (2014). Conflict-triggered top-down control: Default mode, last resort, or no such thing? *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *40*(2), 567–587. https://doi.org/10.1037/a0035032

Capraro, V., Schulz, J., & Rand, D. G. (2019). Time pressure and honesty in a deception game. *Journal of Behavioral and Experimental Economics*, *79*, 93–99. https://doi.org/10.1016/j.socec.2019.01.007

Cassidy, H., Wyman, J., Talwar, V., & Akehurst, L. (2019). Exploring the decision component of the activation-decision-construction-action theory for different reasons to deceive. *Legal and Criminological Psychology*, *24*(1), 87–102. https://doi.org/10.1111/lcrp.12143

Caudek, C., Lorenzino, M., & Liperoti, R. (2017). Delta plots do not reveal response inhibition in lying. *Consciousness and Cognition*, *55*, 232–244. https://doi.org/10.1016/j.concog.2017.09.001

Chiu, Y.-C., & Egner, T. (2015). Inhibition-induced forgetting: When more control leads to less memory. *Psychological Science*, *26*(1), 27–38. https://doi.org/10.1177/0956797614553945

Christ, S. E., Van Essen, D. C., Watson, J. M., Brubaker, L. E., & McDermott, K. B. (2009). The contributions of prefrontal cortex and executive control to deception: Evidence from activation likelihood estimate meta analyses. *Cerebral Cortex*, *19*(7), 1557–1566. https://doi.org/10.1093/cercor/bhn189

Cohen-Kdoshay, O., & Meiran, N. (2009). The representation of instructions operates like a prepared reflex: Flanker compatibility effects found in first trial following S-R instructions. *Experimental Psychology*, *56*(2), 128–133. https://doi.org/10.1027/1618-3169.56.2.128

Coleman, L., & Kay, P. (1981). Prototype semantics: The English Word Lie. *Language*, *57*(1), 26–44. https://doi.org/10.1353/lan.1981.0002

Davidson, M. C., Amso, D., Anderson, L. C., & Diamond, A. (2006). Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia*, *44*(11), 2037–2078. https://doi.org/10.1016/j.neuropsychologia.2006.02.006

Debey, E., De Houwer, J., & Verschuere, B. (2014). Lying relies on the truth. *Cognition*, *132*(3), 324–334. https://doi.org/10.1016/j.cognition.2014.04.009

Debey, E., De Schryver, M., Logan, G. D., Suchotzki, K., & Verschuere, B. (2015). From junior to senior Pinocchio: A cross-sectional lifespan investigation of deception. *Acta Psychologica*, *160*, 58–68. https://doi.org/10.1016/j.actpsy.2015.06.007

Debey, E., Liefooghe, B., De Houwer, J., & Verschuere, B. (2015). Lie, truth, lie: The role of task switching in a deception context. *Psychological Research*, *79*(3), 478–488. https://doi.org/10.1007/s00426-014-0582-4

Debey, E., Ridderinkhof, R. K., Houwer, J. D., Schryver, M. de, & Verschuere, B. (2015). Suppressing the truth as a mechanism of deception: Delta plots reveal the role of response inhibition in lying. *Consciousness and Cognition*, *37*, 148–159. https://doi.org/10.1016/j.concog.2015.09.005

Debey, E., Verschuere, B., & Crombez, G. (2012). Lying and executive control: An experimental investigation using ego depletion and goal neglect. *Acta Psychologica*, *140*(2), 133–141. https://doi.org/10.1016/j.actpsy.2012.03.004

DePaulo, B. M., Kashy, D. A., Kirkendol, S. E., Wyer, M. M., & Epstein, J. A. (1996). Lying in everyday life. *Journal of Personality and Social Psychology*, *70*(5), 979–995. https://doi.org/10.1037/0022-3514.70.5.979

Dhammapeera, P., Hu, X., & Bergström, Z. M. (2019). Imagining a false alibi impairs concealed memory detection with the autobiographical implicit association test. *Journal of Experimental Psychology. Applied.* https://doi.org/10.1037/xap0000250

Dignath, D., & Eder, A. B. (2015). Stimulus conflict triggers behavioral avoidance. *Cognitive, Affective & Behavioral Neuroscience*, *15*(4), 822–836. https://doi.org/10.3758/s13415-015-0355-6

Dignath, D., Kiesel, A., & Eder, A. B. (2015). Flexible conflict management: Conflict avoidance and conflict adjustment in reactive cognitive control. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *41*(4), 975–988. https://doi.org/10.1037/xlm0000089

Duran, N. D., Dale, R., & McNamara, D. S. (2010). The action dynamics of overcoming the truth. *Psychonomic Bulletin & Review*, *17*(4), 486–491. https://doi.org/10.3758/PBR.17.4.486

Ekman, P. (1997). Deception, lying, and demeanor. In *States of mind: American and post-Soviet perspectives on contemporary issues in psychology* (pp. 93–105). New York, NY, US: Oxford University Press.

Ekman, P., & O'Sullivan, M. (1991). Who can catch a liar? *American Psychologist*, *46*(9), 913–920. https://doi.org/10.1037/0003-066X.46.9.913

Ferreira, V. S., & Pashler, H. (2002). Central bottleneck influences on the processing stages of word production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(6), 1187–1199. https://doi.org/10.1037/0278-7393.28.6.1187

Fischbacher, U., & Föllmi-Heusi, F. (2013). Lies in disguise - an experimental study on cheating. *Journal of the European Economic Association*, *11*(3), 525–547. https://doi.org/10.1111/jeea.12014

Foerster, A., Pfister, R., Schmidts, C., Dignath, D., & Kunde, W. (2013). Honesty saves time (and justifications). *Frontiers in Psychology*, *4*, 473. https://doi.org/10.3389/fpsyg.2013.00473

Foerster, A., Pfister, R., Schmidts, C., Dignath, D., Wirth, R., & Kunde, W. (2018). Focused cognitive control in dishonesty: Evidence for predominantly transient conflict adaptation. *Journal of Experimental Psychology. Human Perception and Performance*, *44*(4), 578–602. https://doi.org/10.1037/xhp0000480

Foerster, A., Wirth, R., Berghoefer, F. L., Kunde, W., & Pfister, R. (2018). Capacity limitations of dishonesty. *Journal of Experimental Psychology. General.* Advance online publication. https://doi.org/10.1037/xge0000510

Foerster, A., Wirth, R., Herbort, O., Kunde, W., & Pfister, R. (2017). Lying upside-down: Alibis reverse cognitive burdens of dishonesty. *Journal of Experimental Psychology. Applied*, *23*(3), 301–319. https://doi.org/10.1037/xap0000129

Foerster, A., Wirth, R., Kunde, W., & Pfister, R. (2017). The dishonest mind set in sequence. *Psychological Research*, *81*(4), 878–899. https://doi.org/10.1007/s00426-016-0780-3

Franz, V. H., & Luxburg, U. von (2015). No evidence for unconscious lie detection: A significant difference does not imply accurate classification. *Psychological Science*, *26*(10), 1646–1648. https://doi.org/10.1177/0956797615597333

Funes, M. J., Lupiáñez, J., & Humphreys, G. (2010). Sustained vs. transient cognitive control: Evidence of a behavioral dissociation. *Cognition*, *114*(3), 338–347. https://doi.org/10.1016/j.cognition.2009.10.007

Furedy, J. J., Davis, C., & Gurevich, M. (1988). Differentiation of deception as a psychological process: A psychophysiological approach. *Psychophysiology*, *25*(6), 683–688. https://doi.org/10.1111/j.1469-8986.1988.tb01908.x

Gerlach, P., Teodorescu, K., & Hertwig, R. (2019). The truth about lies: A meta-analysis on dishonest behavior. *Psychological Bulletin*, *145*(1), 1–44. https://doi.org/10.1037/bul0000174

Gneezy, U. (2005). Deception: The role of consequences. *American Economic Review*, *95*(1), 384–394. https://doi.org/10.1257/0002828053828662

Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. *American Psychologist*, *54*(7), 493–503. https://doi.org/10.1037/0003-066X.54.7.493

Gratton, G., Coles, M. G. H., & Donchin, E. (1992). Optimizing the use of information: Strategic control of activation of responses. *Journal of Experimental Psychology: General*, *121*(4), 480–506. https://doi.org/10.1037/0096-3445.121.4.480

Güldenpenning, I., Alaboud, M. A. A., Kunde, W., & Weigelt, M. (2018). The impact of global and local context information on the processing of deceptive actions in game sports. *German Journal of Exercise and Sport Research*, *48*(3), 366–375. https://doi.org/10.1007/s12662-018-0493-4

Halevy, R., Shalvi, S., & Verschuere, B. (2014). Being honest about dishonesty: Correlating self-reports and actual lying. *Human Communication Research*, *40*(1), 54–72. https://doi.org/10.1111/hcre.12019

Hazeltine, E., Lightman, E., Schwarb, H., & Schumacher, E. H. (2011). The boundaries of sequential modulations: Evidence for set-level control. *Journal of Experimental Psychology. Human Perception and Performance*, *37*(6), 1898–1914. https://doi.org/10.1037/a0024662

Henson, R. N., Eckstein, D., Waszak, F., Frings, C., & Horner, A. J. (2014). Stimulus–response bindings in priming. *Trends in Cognitive Sciences*, *18*(7), 376–384. https://doi.org/10.1016/j.tics.2014.03.004

Hilbig, B. E., & Hessler, C. M. (2013). What lies beneath: How the distance between truth and lie drives dishonesty. *Journal of Experimental Social Psychology*, *49*(2), 263–266. https://doi.org/10.1016/j.jesp.2012.11.010

Hilbig, B. E., & Thielmann, I. (2017). Does everyone have a price? On the role of payoff magnitude for ethical decision making. *Cognition*, *163*, 15–25. https://doi.org/10.1016/j.cognition.2017.02.011

Hilbig, B. E., & Zettler, I. (2015). When the cat's away, some mice will play: A basic trait account of dishonest behavior. *Journal of Research in Personality*, *57*, 72–88. https://doi.org/10.1016/j.jrp.2015.04.003

Hommel, B. (1998a). Automatic stimulus-response translation in dual-task performance. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(5), 1368–1384. https://doi.org/10.1037/0096-1523.24.5.1368

Hommel, B. (1998b). Event files: Evidence for automatic integration of stimulus-response episodes. *Visual Cognition*, *5*(1-2), 183–216. https://doi.org/10.1080/713756773

Hommel, B. (2004). Event files: Feature binding in and across perception and action. *Trends in Cognitive Sciences*, *8*(11), 494–500. https://doi.org/10.1016/j.tics.2004.08.007

Hommel, B. (2011). The Simon effect as tool and heuristic. *Acta Psychologica*, *136*(2), 189–202. https://doi.org/10.1016/j.actpsy.2010.04.011

Hu, X., Chen, H., & Fu, G. (2012). A repeated lie becomes a truth? The effect of intentional control and training on deception. *Frontiers in Psychology*, *3*, 488. https://doi.org/10.3389/fpsyg.2012.00488

Hu, X., Rosenfeld, J. P., & Bodenhausen, G. V. (2012). Combating automatic autobiographical associations: The effect of instruction and training in strategically concealing information in the autobiographical implicit association test. *Psychological Science*, *23*(10), 1079–1085. https://doi.org/10.1177/0956797612443834

Jentzsch, I., & Dudschig, C. (2009). Why do we slow down after an error? Mechanisms underlying the effects of posterror slowing. *Quarterly Journal of Experimental Psychology (2006)*, *62*(2), 209–218. https://doi.org/10.1080/17470210802240655

Jentzsch, I., Leuthold, H., & Ulrich, R. (2007). Decomposing sources of response slowing in the PRP paradigm. *Journal of Experimental Psychology. Human Perception and Performance*, *33*(3), 610–626. https://doi.org/10.1037/0096-1523.33.3.610

Johnson, R., Barnhardt, J., & Zhu, J. (2003). The deceptive response: Effects of response conflict and strategic monitoring on the late positive component and episodic memory-related brain activity. *Biological Psychology*, *64*(3), 217–253. https://doi.org/10.1016/j.biopsycho.2003.07.006

Johnson, R., Barnhardt, J., & Zhu, J. (2004). The contribution of executive processes to deceptive responding. *Neuropsychologia*, *42*(7), 878–901. https://doi.org/10.1016/j.neuropsychologia.2003.12.005

Johnson-Laird, P. N. (2004). The history of mental models. In K. Manktelow & M. C. Chung (Eds.), *Psychology of reasoning: Theoretical and historical perspectives* (pp. 179–212). Hove, England: Psychology Press.

Jusyte, A., Pfister, R., Mayer, S. V., Schwarz, K. A., Wirth, R., Kunde, W., & Schönenberg, M. (2017). Smooth criminal: Convicted rule-breakers show reduced cognitive conflict during deliberate rule violations. *Psychological Research*, *81*(5), 939–946. https://doi.org/10.1007/s00426-016-0798-6

Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A. M., & Koch, I. (2010). Control and interference in task switching—a review. *Psychological Bulletin*, *136*(5), 849–874. https://doi.org/10.1037/a0019842

Kleiman, T., Hassin, R. R., & Trope, Y. (2014). The control-freak mind: Stereotypical biases are eliminated following conflict-activated cognitive control. *Journal of Experimental Psychology. General*, *143*(2), 498–503. https://doi.org/10.1037/a0033047

Koch, I., Prinz, W., & Allport, A. D. (2005). Involuntary retrieval in alphabet-arithmetic tasks: Task-mixing and task-switching costs. *Psychological Research*, *69*(4), 252–261. https://doi.org/10.1007/s00426-004-0180-y

Koranyi, N., Schreckenbach, F., & Rothermund, K. (2015). The implicit cognition of lying: Knowledge about having lied to a question is retrieved automatically. *Social Cognition*, *33*(1), 67–84. https://doi.org/10.1521/soco.2015.33.1.67

Kornblum, S., Hasbroucq, T., & Osman, A. (1990). Dimensional overlap: Cognitive basis for stimulus-response compatibility—A model and taxonomy. *Psychological Review*, *97*(2), 253–270. https://doi.org/10.1037/0033-295X.97.2.253

Krebs, R. M., Boehler, C. N., Belder, M. de, & Egner, T. (2015). Neural conflict-control mechanisms improve memory for target stimuli. *Cerebral Cortex*, *25*(3), 833–843. https://doi.org/10.1093/cercor/bht283

Kunde, W., Kiesel, A., & Hoffmann, J. (2003). Conscious control over the content of unconscious cognition. *Cognition*, *88*(2), 223–242. https://doi.org/10.1016/S0010-0277(03)00023-4

Kunde, W., Pfister, R., & Janczyk, M. (2012). The locus of tool-transformation costs. *Journal of Experimental Psychology. Human Perception and Performance*, *38*(3), 703–714. https://doi.org/10.1037/a0026315

Kunde, W., Wirth, R., & Janczyk, M. (2018). The role of feedback delay in dual-task performance. *Psychological Research*, *82*(1), 157–166. https://doi.org/10.1007/s00426-017-0874-6

Kunde, W., & Wühr, P. (2006). Sequential modulations of correspondence effects across spatial dimensions and tasks. *Memory & Cognition*, *34*(2), 356–367. https://doi.org/10.3758/BF03193413

Leboe, J. P., Whittlesea, B. W. A., & Milliken, B. (2005). Selective and nonselective transfer: Positive and negative priming in a multiple-task environment. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *31*(5), 1001–1029. https://doi.org/10.1037/0278-7393.31.5.1001

Lee, J. J., Ong, M., Parmar, B., & Amit, E. (2019). Lay theories of effortful honesty: Does the honesty-effort association justify making a dishonest decision? *Journal of Applied Psychology*, *104*(5), 659–677. https://doi.org/10.1037/apl0000364

Levine, E., Hart, J., Moore, K., Rubin, E., Yadav, K., & Halpern, S. (2018). The surprising costs of silence: Asymmetric preferences for prosocial lies of commission and omission. *Journal of Personality and Social Psychology*, *114*(1), 29–51. https://doi.org/10.1037/pspa0000101

Levine, T. R. (2014). Truth-Default Theory (TDT). *Journal of Language and Social Psychology*, *33*(4), 378–392. https://doi.org/10.1177/0261927X14535916

Levine, T. R., Kim, R. K., & Hamel, L. M. (2010). People lie for a reason: Three experiments documenting the principle of veracity. *Communication Research Reports*, *27*(4), 271–285. https://doi.org/10.1080/08824096.2010.496334

Li, D. (2015). *Do liars come to believe their own lies? The effect of deception on memory* (Doctoral dissertation). Sydney, Australia. Retrieved from http://unsworks.unsw.edu.au/fapi/datastream/unsworks:39131/SOURCE02

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, *95*(4), 492–527. https://doi.org/10.1037/0033-295X.95.4.492

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. https://doi.org/10.3758/BF03197535

Lohse, T., Simon, S. A., & Konrad, K. A. (2018). Deception under time pressure: Conscious decision or a problem of awareness? *Journal of Economic Behavior & Organization*, *146*, 31–42. https://doi.org/10.1016/j.jebo.2017.11.026

Masip, J., Blandón-Gitlin, I., La Riva, C. de, & Herrero, C. (2016). An empirical test of the decision to lie component of the activation-decision-construction-action theory (ADCAT). *Acta Psychologica*, *169*, 45–55. https://doi.org/10.1016/j.actpsy.2016.05.004

Massaro, D. W., & Cowan, N. (1993). Information processing models: Microscopes of the mind. *Annual Review of Psychology*, *44*, 383–425. https://doi.org/10.1146/annurev.ps.44.020193.002123

McClelland, J. L. (1979). On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review*, *86*(4), 287–330. https://doi.org/10.1037/0033-295X.86.4.287

McCornack, S. A. (2015). Information manipulation theory. In C. R. Berger, M. E. Roloff, S. R. Wilson, J. P. Dillard, J. Caughlin, & D. Solomon (Eds.), *The International Encyclopedia of Interpersonal Communication* (Vol. 6, pp. 1–7). Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9781118540190.wbeic072

McCornack, S. A., Morrison, K., Paik, J. E., Wisner, A. M., & Zhu, X. (2014). Information manipulation theory 2. *Journal of Language and Social Psychology*, *33*(4), 348–377. https://doi.org/10.1177/0261927X14534656

Meyer, D. E., & Kieras, D. E. (1997). A computational theory of executive cognitive processes and multiple-task performance: Part 2. Accounts of psychological refractory-period phenomena. *Psychological Review*, *104*(4), 749–791. https://doi.org/10.1037/0033-295X.104.4.749

Miller, J. (2006). Backward crosstalk effects in psychological refractory period paradigms: Effects of second-task response types on first-task response latencies. *Psychological Research*, *70*(6), 484–493. https://doi.org/10.1007/s00426-005-0011-9

Miller, J., & Reynolds, A. (2003). The locus of redundant-targets and nontargets effects: Evidence from the psychological refractory period paradigm. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(6), 1126–1142. https://doi.org/10.1037/0096-1523.29.6.1126

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, *7*(3), 134–140. https://doi.org/10.1016/S1364-6613(03)00028-7

Notebaert, W., Gevers, W., Verbruggen, F., & Liefooghe, B. (2006). Top-down and bottom-up sequential modulations of congruency effects. *Psychonomic Bulletin & Review*, *13*(1), 112–117. https://doi.org/10.3758/BF03193821

Notebaert, W., & Verguts, T. (2008). Cognitive control acts locally. *Cognition*, *106*(2), 1071–1080. https://doi.org/10.1016/j.cognition.2007.04.011

Otgaar, H., & Baker, A. (2018). When lying changes memory for the truth. *Memory*, *26*(1), 2–14. https://doi.org/10.1080/09658211.2017.1340286

Paelecke, M., & Kunde, W. (2007). Action-effect codes in and before the central bottleneck: Evidence from the psychological refractory period paradigm. *Journal of Experimental Psychology. Human Perception and Performance*, *33*(3), 627–644. https://doi.org/10.1037/0096-1523.33.3.627

Pashler, H. (1984). Processing stages in overlapping tasks: Evidence for a central bottleneck. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(3), 358–377. https://doi.org/10.1037/0096-1523.10.3.358

Pashler, H. (1994a). Dual-task interference in simple tasks: Data and theory. *Psychological Bulletin*, *116*(2), 220–244. https://doi.org/10.1037/0033-2909.116.2.220

Pashler, H. (1994b). Graded capacity-sharing in dual-task interference? *Journal of Experimental Psychology: Human Perception and Performance*, *20*(2), 330–342. https://doi.org/10.1037/0096-1523.20.2.330

Pashler, H., & Johnston, J. C. (1989). Chronometric evidence for central postponement in temporally overlapping tasks. *The Quarterly Journal of Experimental Psychology Section A*, *41*(1), 19–45. https://doi.org/10.1080/14640748908402351

Pfeuffer, C. U., Pfister, R., Foerster, A., Stecher, F., & Kiesel, A. (2019). Binding lies: Flexible retrieval of honest and dishonest behavior. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(2), 157–173. https://doi.org/10.1037/xhp0000600

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

Pfister, R., Foerster, A., & Kunde, W. (2014). Pants on fire: The electrophysiological signature of telling a lie. *Social Neuroscience*, *9*(6), 562–572. https://doi.org/10.1080/17470919.2014.934392

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. https://doi.org/10.5709/acp-0133-x

Pfister, R., Wirth, R., Schwarz, K. A., Foerster, A., Steinhauser, M., & Kunde, W. (2016). The electrophysiological signature of deliberate rule violations. *Psychophysiology*, *53*(12), 1870–1877. https://doi.org/10.1111/psyp.12771

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. https://doi.org/10.1016/j.cognition.2015.11.009

Pfister, R., Wirth, R., Weller, L., Foerster, A., & Schwarz, K. A. (2018). Taking shortcuts: Cognitive conflict during motivated rule-breaking. *Journal of Economic Psychology, 71,* 138-147. https://doi.org/10.1016/j.joep.2018.06.005

Pittarello, A., Rubaltelli, E., & Motro, D. (2016). Legitimate lies: The relationship between omission, commission, and cheating. *European Journal of Social Psychology*, *46*(4), 481–491. https://doi.org/10.1002/ejsp.2179

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, *489*(7416), 427–430. https://doi.org/10.1038/nature11467

Reuss, H., Kiesel, A., Kunde, W., & Hommel, B. (2011). Unconscious activation of task sets. *Consciousness and Cognition*, *20*(3), 556–567. https://doi.org/10.1016/j.concog.2011.02.014

Rogers, T., Zeckhauser, R., Gino, F., Norton, M. I., & Schweitzer, M. E. (2017). Artful paltering: The risks and rewards of using truthful statements to mislead others. *Journal of Personality and Social Psychology*, *112*(3), 456–473. https://doi.org/10.1037/pspi0000081

Rosner, T. M., D'Angelo, M. C., MacLellan, E., & Milliken, B. (2015). Selective attention and recognition: Effects of congruency on episodic learning. *Psychological Research*, *79*(3), 411–424. https://doi.org/10.1007/s00426-014-0572-6

Rutschmann, R., & Wiegmann, A. (2017). No need for an intention to deceive? Challenging the traditional definition of lying. *Philosophical Psychology*, *30*(4), 438–457. https://doi.org/10.1080/09515089.2016.1277382

Sartori, G., Agosta, S., Zogmaister, C., Ferrara, S. D., & Castiello, U. (2008). How to accurately detect autobiographical events. *Psychological Science*, *19*(8), 772–780. https://doi.org/10.1111/j.1467-9280.2008.02156.x

Schindler, S., & Pfattheicher, S. (2017). The frame of the game: Loss-framing increases dishonest behavior. *Journal of Experimental Social Psychology*, *69*, 172–177. https://doi.org/10.1016/j.jesp.2016.09.009

Schneider, D. W., & Anderson, J. R. (2010). Asymmetric switch costs as sequential difficulty effects. *Quarterly Journal of Experimental Psychology*, *63*(10), 1873–1894. https://doi.org/10.1080/17470211003624010

Serota, K. B., Levine, T. R., & Boster, F. J. (2010). The prevalence of lying in America: Three studies of self-reported lies. *Human Communication Research*, *36*(1), 2–25. https://doi.org/10.1111/j.1468-2958.2009.01366.x

Shalvi, S., Eldar, O., & Bereby-Meyer, Y. (2012). Honesty requires time (and lack of justifications). *Psychological Science*, *23*(10), 1264–1270. https://doi.org/10.1177/0956797612443835

Shalvi, S., Eldar, O., & Bereby-Meyer, Y. (2013). Honesty requires time-a reply to Foerster et al. (2013). *Frontiers in Psychology*, *4*, 634. https://doi.org/10.3389/fpsyg.2013.00634

Simon, J. R., & Rudell, A. P. (1967). Auditory S-R compatibility: The effect of an irrelevant cue on information processing. *Journal of Applied Psychology*, *51*(3), 300–304. https://doi.org/10.1037/h0020586

Smith, E. E. (1968). Choice reaction time: An analysis of the major theoretical positions. *Psychological Bulletin*, *69*(2), 77–110. https://doi.org/10.1037/h0020189

Spapé, M. M., & Hommel, B. (2008). He said, she said: Episodic retrieval induces conflict adaptation in an auditory Stroop task. *Psychonomic Bulletin & Review*, *15*(6), 1117–1121. https://doi.org/10.3758/PBR.15.6.1117

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. https://doi.org/10.1097/00001756-200109170-00019

Spence, S. A., Kaylor-Hughes, C. J., Brook, M. L., Lankappa, S. T., & Wilkinson, I. D. (2008). 'Munchausen's syndrome by proxy' or a 'miscarriage of justice'? An initial application of functional neuroimaging to the question of guilt versus innocence. *European Psychiatry*, *23*(4), 309–314. https://doi.org/10.1016/j.eurpsy.2007.09.001

Steinhauser, M., Ernst, B., & Ibald, K. W. (2017). Isolating component processes of posterror slowing with the psychological refractory period paradigm. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *43*(4), 653–659. https://doi.org/10.1037/xlm0000329

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. *Acta psychologica*, *30*, 276–315. https://doi.org/10.1016/0001-6918(69)90055-9

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643–662. https://doi.org/10.1037/h0054651

Stürmer, B., Nigbur, R., Schacht, A., & Sommer, W. (2011). Reward and punishment effects on error processing and conflict control. *Frontiers in Psychology*, *2*, 335. https://doi.org/10.3389/fpsyg.2011.00335

Suchotzki, K., Berlijn, A., Donath, M., & Gamer, M. (2018). Testing the applied potential of the Sheffield lie test. *Acta Psychologica*, *191*, 281–288. https://doi.org/10.1016/j.actpsy.2018.10.011

Suchotzki, K., Crombez, G., Smulders, F. T. Y., Meijer, E., & Verschuere, B. (2015). The cognitive mechanisms underlying deception: An event-related potential study. *International Journal of Psychophysiology*, *95*(3), 395–405. https://doi.org/10.1016/j.ijpsycho.2015.01.010

Suchotzki, K., Verschuere, B., Van Bockstaele, B., Ben-Shakhar, G., & Crombez, G. (2017). Lying takes time: A meta-analysis on reaction time measures of deception. *Psychological Bulletin*, *143*(4), 428–453. https://doi.org/10.1037/bul0000087

Tabatabaeian, M., Dale, R., & Duran, N. D. (2015). Self-serving dishonest decisions can show facilitated cognitive dynamics. *Cognitive Processing*, *16*(3), 291–300. https://doi.org/10.1007/s10339-015-0660-6

Teper, R., & Inzlicht, M. (2011). Active Transgressions and Moral Elusions. *Social Psychological and Personality Science*, *2*(3), 284–288. https://doi.org/10.1177/1948550610389338

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. https://doi.org/10.1016/j.actpsy.2012.11.015

Torres-Quesada, M., Lupiáñez, J., Milliken, B., & Funes, M. J. (2014). Gradual proportion congruent effects in the absence of sequential congruent effects. *Acta Psychologica*, *149*, 78–86. https://doi.org/10.1016/j.actpsy.2014.03.006

Turri, A., & Turri, J. (2015). The truth about lying. *Cognition*, *138*, 161–168. https://doi.org/10.1016/j.cognition.2015.01.007

Turri, A., & Turri, J. (2019). Lying, fast and slow. *Synthese*, *63*(3), 368. https://doi.org/10.1007/s11229-018-02062-z

Turri, J., & Turri, A. (2016). Lying, uptake, assertion, and intent. *International Review of Pragmatics*, *8*(2), 314–333. https://doi.org/10.1163/18773109-00802006

Van Bockstaele, B., Verschuere, B., Moens, T., Suchotzki, K., Debey, E., & Spruyt, A. (2012). Learning to lie: Effects of practice on the cognitive cost of lying. *Frontiers in Psychology*, *3*, 526. https://doi.org/10.3389/fpsyg.2012.00526

Van Bockstaele, B., Wilhelm, C., Meijer, E., Debey, E., & Verschuere, B. (2015). When deception becomes easy: The effects of task switching and goal neglect on the truth proportion effect. *Frontiers in Psychology*, *6*, 1666. https://doi.org/10.3389/fpsyg.2015.01666

Van Steenbergen, H., Band, G. P. H., & Hommel, B. (2009). Reward counteracts conflict adaptation. Evidence for a role of affect in executive control. *Psychological Science*, *20*(12), 1473–1477. https://doi.org/10.1111/j.1467-9280.2009.02470.x

Verschuere, B., Prati, V., & Houwer, J. D. (2009). Cheating the lie detector: Faking in the autobiographical implicit association test. *Psychological Science*, *20*(4), 410–413. https://doi.org/10.1111/j.1467-9280.2009.02308.x

Verschuere, B., & Shalvi, S. (2014). The truth comes naturally! Does it? *Journal of Language and Social Psychology*, *33*(4), 417–423. https://doi.org/10.1177/0261927X14535394

Verschuere, B., Spruyt, A., Meijer, E., & Otgaar, H. (2011). The ease of lying. *Consciousness and Cognition*, *20*(3), 908–911. https://doi.org/10.1016/j.concog.2010.10.023

Walczyk, J. J., Griffith, D. A., Yates, R., Visconte, S. R., Simoneaux, B., & Harris, L. L. (2012). Lie detection by inducing cognitive load. *Criminal Justice and Behavior*, *39*(7), 887–909. https://doi.org/10.1177/0093854812437014

Walczyk, J. J., Harris, L. L., Duck, T. K., & Mulay, D. (2014). A social-cognitive framework for understanding serious lies: Activation-decision-construction-action theory. *New Ideas in Psychology*, *34*, 22–36. https://doi.org/10.1016/j.newideapsych.2014.03.001

Walczyk, J. J., Mahoney, K. T., Doverspike, D., & Griffith-Ross, D. A. (2009). Cognitive Lie Detection: Response Time and Consistency of Answers as Cues to Deception. *Journal of Business and Psychology*, *24*(1), 33–49. https://doi.org/10.1007/s10869-009-9090-8

Walczyk, J. J., Roper, K. S., Seemann, E., & Humphrey, A. M. (2003). Cognitive mechanisms underlying lying to questions: response time as a cue to deception. *Applied Cognitive Psychology*, *17*(7), 755–774. https://doi.org/10.1002/acp.914

Walczyk, J. J., Schwartz, J. P., Clifton, R., Adams, B., Wei, M. I. N., & Zha, P. (2005). Lying person-to-person about life events: A cognitive framework for lie detection. *Personnel Psychology*, *58*(1), 141–170. https://doi.org/10.1111/j.1744-6570.2005.00484.x

Walczyk, J. J., Tcholakian, T., Newman, D. N., & Duck, T. K. (2016). Impromptu decisions to deceive. *Applied Cognitive Psychology*, *30*(6), 934–945. https://doi.org/10.1002/acp.3282

Waszak, F., Pfister, R., & Kiesel, A. (2013). Top-down versus bottom-up: When instructions overcome automatic retrieval. *Psychological Research*, *77*(5), 611–617. https://doi.org/10.1007/s00426-012-0459-3

Welford, A. T. (1952). The 'psychological refractory period' and the timing of high-speed performance-a review and a theory. *British Journal of Psychology*, *43*(1), 2–19. https://doi.org/10.1111/j.2044-8295.1952.tb00322.x

Wenke, D., Gaschler, R., & Nattkemper, D. (2007). Instruction-induced feature binding. *Psychological Research*, *71*(1), 92–106. https://doi.org/10.1007/s00426-005-0038-y

Wiegmann, A., Samland, J., & Waldmann, M. R. (2016). Lying despite telling the truth. *Cognition*, *150*, 37–42. https://doi.org/10.1016/j.cognition.2016.01.017

Wirth, R., Dignath, D., Pfister, R., Kunde, W., & Eder, A. B. (2016). Attracted by rewards: Disentangling the motivational influence of rewarding and punishing targets and distractors. *Motivation Science*, *2*(3), 143–156. https://doi.org/10.1037/mot0000037

Wirth, R., Foerster, A., Rendel, H., Kunde, W., & Pfister, R. (2018). Rule-violations sensitise towards negative and authority-related stimuli. *Cognition & Emotion*, *32*(3), 480–493. https://doi.org/10.1080/02699931.2017.1316706

Wirth, R., Janczyk, M., & Kunde, W. (2018). Effect monitoring in dual-task performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(4), 553–571. https://doi.org/10.1037/xlm0000474

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. https://doi.org/10.1007/s00426-015-0690-9

Wirth, R., Pfister, R., Janczyk, M., & Kunde, W. (2015). Through the portal: Effect anticipation in the central bottleneck. *Acta Psychologica*, *160*, 141–151. https://doi.org/10.1016/j.actpsy.2015.07.007

Wirth, R., Pfister, R., & Kunde, W. (2016). Asymmetric transfer effects between cognitive and affective task disturbances. *Cognition & Emotion*, *30*(3), 399–416. https://doi.org/10.1080/02699931.2015.1009002

Wirth, R., Steinhauser, R., Janczyk, M., Steinhauser, M., & Kunde, W. (2018). Long-term and short-term action-effect links and their impact on effect monitoring. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(8), 1186–1198. https://doi.org/10.1037/xhp0000524

Wühr, P., Duthoo, W., & Notebaert, W. (2015). Generalizing attentional control across dimensions and tasks: Evidence from transfer of proportion-congruent effects. *Quarterly Journal of Experimental Psychology (2006)*, *68*(4), 779–801. https://doi.org/10.1080/17470218.2014.966729

× Appendices

Appendix 1: Question sets

Table 1 | Question set with German originals and English translations of Experiments 1 and 3 to 7.

Code	German original	English translation
1	Warst du Joggen?	Did you go for a run?
2	Bist du eine Treppe herunter gegangen?	Did you go down a staircase?
3	Bist du eine Treppe hoch gegangen?	Did you go up a staircase?
4	Hast du getankt?	Did you buy petrol?
5	Hast du Schokolade gegessen?	Did you eat chocolate?
6	Bist du Bus gefahren?	Did you take a bus?
7	Bist du Zug gefahren?	Did you take a train?
8	Hast du einen Mülleimer benutzt?	Did you use a dustbin?
9	Hast du ein Bad genommen?	Did you take a bath?
10	Hast du ein Toast zubereitet?	Did you make a sandwich?
11	Hast du einen Brief geschrieben?	Did you post a letter?
12	Hast du eine Tür geschlossen?	Did you close a door?
13	Warst du duschen?	Did you take a shower?
14	Hast du eine Zeitung gekauft?	Did you buy a newspaper?
15	Hast du eine Zeitschrift gekauft?	Did you buy a magazine?
16	Hast du ein Messer benutzt?	Did you use a knife?
17	Hast du einen Regenschirm benutzt?	Did you use an umbrella?
18	Hast du ein Medikament genommen?	Did you take a pill?
19	Hast du mit einem Polizisten gesprochen?	Did you speak to a police officer?
20	Hast du einen Apfel gegessen?	Did you eat an apple?
21	Hast du ein Fenster zerstört?	Did you break a window?
22	Hast du telefoniert?	Did you use a telephone?
23	Hast du eine SMS erhalten?	Did you receive a text?
24	Hast du einen Saft getrunken?	Did you drink fruit juice?
25	Hast du Radio gehört?	Did you listen to the radio?
26	Warst du im Internet?	Did you use the internet?
27	Hast du in einer Schlange angestanden?	Did you stand in a queue?
28	Hast du in einem Warteraum gesessen?	Did you sit in a waiting room?
29	Hast du dein Bett gemacht?	Did you make your bed?
30	Hast du deine Hände gewaschen?	Did you wash your hands?
31	Hast du ein Dokument unterzeichnet?	Did you sign a document?
32	Hast du Kaffee getrunken?	Did you drink coffee?
33	Hast du mit einem Kind gesprochen?	Did you speak to a child?

34	Hast du Fernsehen geschaut?	Did you watch television?
35	Hast du Zwiebeln gegessen?	Did you eat onions?
36	Hast du Wasser getrunken?	Did you drink water?
37	Hast du an einer Ampel gehalten?	Did you stop at a traffic light?
38	Warst du im Supermarkt?	Did you go to a supermarket?
39	Hast du Blumen gekauft?	Did you buy some flowers?
40	Hast du abgewaschen?	Did you do the dishes?
41	Bist du Fahrstuhl gefahren?	Did you take an elevator?
42	Hast du ein Fenster geputzt?	Did you clean a window?
43	Hast du eine Verabredung verschoben?	Did you reschedule an appointment?
44	Hast du ein Buch gelesen?	Did you read a book?
45	Hast du ein Moped abgestellt?	Did you park a moped?
46	Hast du eine Zitrone ausgepresst?	Did you squeeze a lemon?
47	Hast du eine Email verschickt?	Did you send an e-mail?
48	Hast du ein Tier gestreichelt?	Did you stroke a pet?
49	Hast du einen Mantel getragen?	Did you wear a coat?
50	Hast du einen Kühlschrank geöffnet?	Did you open a fridge?
51	Hast du einen Computer eingeschaltet?	Did you switch on a computer?
52	Hast du eine Zigarette geraucht?	Did you smoke a cigarette?
53	Hast du auf eine Uhr geschaut?	Did you look at a watch?
54	Hast du einen Wasserhahn geöffnet?	Did you open a water tap?
55	Hast du einen Toilettendeckel geöffnet?	Did you lift a toilet seat?
56	Bist du über einen Zebrastreifen gelaufen?	Did you use a pedestrian crossing?
57	Hast du einen Geldautomaten benutzt?	Did you use an ATM?
58	Hast du Geld gewechselt?	Did you change money?
59	Hast du einen Teppich abgesaugt?	Did you vacuum a carpet?
60	Hast du Hustensaft getrunken?	Did you drink cough syrup?
61	Hast du jemanden gegrüßt?	Did you greet someone?
62	Hast du geputzt?	Did you clean the house?
63	Hast du in deinen Briefkasten geschaut?	Did you check your mailbox?
64	Hast du deine Zähne geputzt?	Did you brush your teeth?
65	Hast du Musik gehört?	Did you listen to music?
66	Bist du Fahrrad gefahren?	Did you ride on a bicycle?
67	Hast du auf einer Leiter gestanden?	Did you stand on a ladder?
68	Hast du auf einem Stuhl gesessen?	Did you sit on a chair?
69	Hast du ein Stück Papier abgerissen?	Did you rip a piece of paper?
70	Hast du Blumen gegossen?	Did you water the plants?
71	Hast du deine Schlüssel benutzt?	Did you use your keys?
72	Hast du Wasser gekocht?	Did you boil some water?

Table 2 | Question set with German originals and English translations of Experiment 2. Explanations in brackets were not part of the question.

Set	German original	English translation
1	Hast du die Würfel gestapelt?	Did you stack the dice?
	Hast du die Bausteine getrennt?	Did you take apart the bricks?
	Hast du auf das Blatt gestempelt?	Did you stamp the piece of paper?
	Hast du die Stiftkappen vertauscht?	Did you swap the caps of the pens?
	Hast du in das Papiertuch getackert?	Did you staple the paper towel?
	Hast du die Klammer am Cent befestigt?	Did you clip the peg on the cent?
	Hast du den Kaffeefilter durchstochen?	Did you puncture the coffee filter?
	Hast du den Sticker auf den Teller geklebt?	Did you put the sticker on the plate?
	Hast du die Fliege [aus Papier] in die Schachtel gepackt?	Did you put the [paper] fly into the container?
	Hast du eine Schleife um die Gabel gebunden?	Did you tie a bow to the fork?
2	Hast du die Nudel zerbrochen?	Did you break the noodle?
	Hast du die Karte zerschnitten?	Did you cut the card?
	Hast du den Helikopter ausgemalt?	Did you color a helicopter?
	Hast du Reis in die Dose umgefüllt?	Did you decant the rice into the container?
	Hast du die Watte zur Kugel gerollt?	Did you form a ball from the cotton wool?
	Hast du den Draht vom Deckel entfernt?	Did you detach the wire from the cap?
	Hast du die Murmel in die Folie getan?	Did you put the marble into the transparent envelope?
	Hast du den Magneten aus der Kapsel geholt?	Did you take the magnet from the capsule?
	Hast du die Mutter von der Schraube gedreht?	Did you loosen the nut from the screw?
	Hast du Papier aus der Zeitschrift gerissen?	Did you rip paper from the magazine?

Table 3 | Question set with German originals and English translations of Experiments 8 to 11. The first five routine questions concern activities that were likely experienced on that day and the other five activities of the routine questions were very unlikely to be experienced on a common day. The top five questions of the mission type asked about the activities participants were instructed to do in the mission. The last five questions concern the alibi that participants got but were asked not to engage in (Experiments 8 to 10 or were told about after their mission (Experiment 11).

Activity status	Routine questions		Mission questions	
	German original	English translation	German original	English translation
Experienced	Hast du eine Straße überquert?	Did you cross a street?	Hast du ein Dreieck gezeichnet?	Did you draw a triangle?
	Hast du mit jemandem gesprochen?	Did you talk to somebody?	Hast du ein Blatt zerrissen?	Did you rip a sheet?
	Hast du eine Tür durchquert?	Did you walk through a door?	Hast du eine Box geöffnet?	Did you open a box?
	Hast du ein Dokument unterzeichnet?	Did you sign a document?	Hast du ein Papierstück versteckt?	Did you hide a piece of paper?
	Hast du dir Schuhe angezogen?	Did you put your shoes on?	Hast du einen Kreis gezeichnet?	Did you draw a circle?
Not experienced	Hast du ein Kamel gestreichelt?	Did you pet a camel?	Hast du eine Email verfasst?	Did you write an email?
	Hast du im Lottospiel gewonnen?	Did you win the lottery?	Hast du eine Datei gesendet?	Did you send a file?
	Hast du ein Fenster zerstört?	Did you destroy a window?	Hast du eine Tabelle geöffnet?	Did you open a table?
	Hast du die Polizei angerufen?	Did you call the police?	Hast du den USB-Stick benutzt?	Did you use a USB stick?
	Hast du einen Pilz gesammelt?	Did you pick mushrooms?	Hast du den PC eingeschaltet?	Did you turn on the PC?

Appendix 2: Descriptive Statistics

Table 4 | Mean (*M*) error rates and mean differences (Δ = dishonest – honest) in percent with standard deviations (*SD*s) in parentheses for each combination of task, stimulus onset asynchrony (SOA), and intention of Experiment 1.

Task	SOA	Intention	Error rate	
			***M* (*SD*)**	**Δ*M* (*SD*)**
Tone Task 1	150 ms	Honest	2.1 (2.57)	-0.1 (1.83)
		Dishonest	2.0 (3.08)	
	1500 ms	Honest	1.9 (1.86)	0.1 (1.70)
		Dishonest	1.9 (2.04)	
(Dis)honest Task 2	150 ms	Honest	10.0 (5.99)	5.4 (6.58)
		Dishonest	15.4 (8.21)	
	1500 ms	Honest	9.7 (7.43)	4.8 (7.62)
		Dishonest	14.5 (8.81)	

Table 5 | Mean (*M*) response times and mean differences (Δ = dishonest – honest) in milliseconds with standard deviations (*SD*s) in parentheses for each combination of task, stimulus onset asynchrony (SOA) and intention of Experiment 1.

Task	SOA	Intention	Response time	
			***M* (*SD*)**	**Δ*M* (*SD*)**
Tone Task 1	150 ms	Honest	740 (201.4)	26 (73.1)
		Dishonest	765 (231.2)	
	1500 ms	Honest	614 (133.5)	5 (31.6)
		Dishonest	619 (149.8)	
(Dis)honest Task 2	150 ms	Honest	1456 (309.7)	150 (128.8)
		Dishonest	1606 (308.5)	
	1500 ms	Honest	1147 (184.7)	183 (104.0)
		Dishonest	1331 (215.3)	

Table 6 | Mean (*M*) error rates and mean differences (Δ = dishonest – honest) in percent with standard deviations (*SD*s) in parentheses for each combination of task, stimulus onset asynchrony (SOA), and intention of Experiment 2.

Task	SOA	Intention	Error rate	
			***M* (*SD*)**	**Δ*M* (*SD*)**
Tone task 1	150 ms	Honest	2.7 (4.46)	0.3 (3.18)
		Dishonest	3.0 (3.54)	
	1500 ms	Honest	2.2 (1.95)	0.1 (2.42)
		Dishonest	2.3 (3.17)	
(Dis)honest task 2	150 ms	Honest	8.7 (9.74)	7.7 (6.57)
		Dishonest	16.4 (9.26)	
	1500 ms	Honest	7.8 (8.96)	8.8 (7.09)
		Dishonest	16.6 (11.74)	

Table 7 | Mean (*M*) response times and mean differences (Δ = dishonest – honest) in milliseconds with standard deviations (*SD*s) in parentheses for each combination of task, stimulus onset asynchrony (SOA) and intention of Experiment 2.

Task	SOA	Intention	Response time	
			***M* (*SD*)**	**Δ*M* (*SD*)**
Tone task 1	150 ms	Honest	754 (197.1)	1 (106.9)
		Dishonest	755 (203.7)	
	1500 ms	Honest	626 (139.8)	3 (32.9)
		Dishonest	629 (141.4)	
(Dis)honest task 2	150 ms	Honest	1470 (287.1)	132 (134.4)
		Dishonest	1602 (308.8)	
	1500 ms	Honest	1178 (174.8)	181 (117.0)
		Dishonest	1359 (207.6)	

Table 8 | Mean (*M*) error rates and mean differences (Δ = dishonest – honest) in percent with standard deviations (*SD*s) in parentheses for each combination of task, stimulus onset asynchrony (SOA), and intention of Experiment 3.

Task	SOA	Intention	Error rate	
			***M* (*SD*)**	**Δ*M* (*SD*)**
(Dis)honest task 1	150 ms	Honest	6.4 (6.06)	3.8 (4.66)
		Dishonest	10.2 (5.96)	
	1500 ms	Honest	6.6 (5.52)	4.1 (5.79)
		Dishonest	10.7 (7.32)	
Tone task 2	150 ms	Honest	7.3 (4.31)	-0.6 (4.64)
		Dishonest	6.7 (4.79)	
	1500 ms	Honest	3.9 (4.89)	0.5 (3.39)
		Dishonest	4.4 (3.67)	

Table 9 | Mean (*M*) response times and mean differences (Δ = dishonest – honest) in milliseconds with standard deviations (*SD*s) in parentheses for each combination of task, stimulus onset asynchrony (SOA) and intention of Experiment 3.

Task	SOA	Intention	Response time	
			***M* (*SD*)**	**Δ*M* (*SD*)**
(Dis)honest task 1	150 ms	Honest	1397 (282.3)	179 (142.0)
		Dishonest	1576 (305.0)	
	1500 ms	Honest	1426 (375.1)	207 (140.3)
		Dishonest	1633 (398.8)	
Tone task 2	150 ms	Honest	1664 (296.9)	211 (163.6)
		Dishonest	1875 (323.0)	
	1500 ms	Honest	661 (211.6)	111 (109.3)
		Dishonest	772 (258.3)	

Table 10 | Mean (*M*) error rates and mean differences (Δ = dishonest – honest) in percent with standard deviations (*SD*s) in parentheses for each combination of task, response-stimulus interval (RSI), and intention of Experiment 4.

Task	RSI	Intention	Error rate	
			***M* (*SD*)**	**Δ*M* (*SD*)**
(Dis)honest task 1	0 ms	Honest	8.6 (7.48)	7.0 (7.62)
		Dishonest	15.5 (10.80)	
	1000 ms	Honest	8.7 (7.25)	6.4 (7.10)
		Dishonest	15.1 (10.77)	
Tone task 2	0 ms	Honest	4.4 (3.36)	-0.6 (3.10)
		Dishonest	3.9 (3.57)	
	1000 ms	Honest	2.5 (2.30)	0.0 (2.72)
		Dishonest	2.5 (2.50)	

Table 11 | Mean (*M*) response times and mean differences (Δ = dishonest – honest) in milliseconds with standard deviations (*SD*s) in parentheses for each combination of task, response-stimulus interval (RSI) and intention of Experiment 4.

Task	SOA	Intention	Response time	
			***M* (*SD*)**	**Δ*M* (*SD*)**
(Dis)honest task 1	0 ms	Honest	1083 (239.2)	153 (113.2)
		Dishonest	1235 (252.4)	
	1000 ms	Honest	1101 (261.2)	132 (100.4)
		Dishonest	1233 (259.4)	
Tone task 2	0 ms	Honest	584 (89.4)	4 (29.4)
		Dishonest	588 (96.5)	
	1000 ms	Honest	445 (74.2)	4 (23.3)
		Dishonest	449 (72.1)	

Table 12 | Mean (*M*) error rates and mean differences (Δ = dishonest – honest) in percent with standard deviations (*SD*s) in parentheses for each combination of task, response-stimulus interval (RSI), and intention of the follow-up experiment of Experiment 2.

Task	SOA	Intention	Error rate	
			***M* (*SD*)**	**Δ*M* (*SD*)**
Tone task 1	150 ms	Honest	3.1 (2.42)	0.2 (2.58)
		Dishonest	3.3 (3.85)	
	1500 ms	Honest	1.8 (1.74)	0.5 (1.64)
		Dishonest	2.3 (2.18)	
(Dis)honest task 2	150 ms	Honest	7.8 (3.41)	10.3 (6.12)
		Dishonest	18.1 (7.00)	
	1500 ms	Honest	6.2 (3.03)	11.4 (6.09)
		Dishonest	17.6 (7.36)	

Table 13 | Mean (*M*) response times and mean differences (Δ = dishonest – honest) in milliseconds with standard deviations (*SD*s) in parentheses for each combination of task, response-stimulus interval (RSI) and intention of the follow-up experiment of Experiment 2.

Task	SOA	Intention	Response time	
			***M* (*SD*)**	**Δ*M* (*SD*)**
Tone task 1	150 ms	Honest	933 (305.1)	38 (97.6)
		Dishonest	971 (356.5)	
	1500 ms	Honest	944 (334.6)	35 (73.8)
		Dishonest	979 (349.5)	
(Dis)honest task 2	150 ms	Honest	1682 (359.4)	199 (127.7)
		Dishonest	1880 (413.8)	
	1500 ms	Honest	1539 (397.4)	208 (112.9)
		Dishonest	1748 (419.5)	

Table 14 | Mean (*M*) error rates and response times with standard deviations (*SD*s) in parentheses for each combination of item, dishonest proportion and current congruency and the mean differences (Δ; *SD*s in parentheses) between honest and dishonest responses of Experiment 5.

Item	Dishonest proportion	Current congruency	Error rate		Response time	
			***M* (*SD*)**	**Δ*M* (*SD*)**	***M* (*SD*)**	**Δ*M* (*SD*)**
Inducer	Low	Honest	6.5 (4.73)	4.4 (7.68)	1066.0 (165.25)	259.5 (185.45)
		Dishonest	10.9 (9.96)		1325.5 (241.76)	
	High	Honest	11.1 (9.46)	2.8 (9.42)	1197.1 (211.54)	68.2 (146.49)
		Dishonest	13.9 (8.00)		1265.3 (196.20)	
Probe	Low	Honest	5.3 (4.05)	7.1 (7.77)	1082.6 (166.10)	213.1 (142.33)
		Dishonest	12.4 (7.53)		1295.7 (201.79)	
	High	Honest	8.2 (5.85)	3.9 (6.49)	1178.2 (217.15)	110.8 (101.66)
		Dishonest	12.1 (6.97)		1289.0 (218.20)	

Table 15 | Mean error rates, response times and their standard deviations in parentheses for each combination of question type, activity-response correspondence, and activity-distractor correspondence of Experiments 10 and 11.

Experiment	Question type	Activity-response correspondence	Activity-distractor correspondence	Error rate	Response time
Experiment 10	Routine	Corresponding	Corresponding	7.0 (7.63)	1195 (249)
			Noncorresponding	9.3 (8.39)	1233 (254)
		Noncorresponding	Corresponding	12.7 (11.23)	1460 (282)
			Noncorresponding	14.7 (11.48)	1489 (303)
	Mission	Corresponding	Corresponding	13.8 (12.74)	1559 (311)
			Noncorresponding	12.8 (12.34)	1558 (333)
		Noncorresponding	Corresponding	7.7 (9.20)	1279 (256)
			Noncorresponding	7.0 (8.34)	1292 (266)
Experiment 11	Routine	Corresponding	Corresponding	9.5 (9.81)	1154 (231)
			Noncorresponding	10.3 (9.20)	1182 (235)
		Noncorresponding	Corresponding	12.0 (9.58)	1375 (284)
			Noncorresponding	13.7 (10.12)	1387 (289)
	Mission	Corresponding	Corresponding	5.9 (8.15)	1190 (243)
			Noncorresponding	6.9 (8.53)	1213 (262)
		Noncorresponding	Corresponding	9.9 (8.27)	1434 (292)
			Noncorresponding	12.9 (9.89)	1461 (286)

Appendix 3: Follow-up experiment of Chapter II

Data treatment and analyses. We preregistered this experiment publicly (osf.io/hdqyx). The intention effect for the short SOA of 150 ms in Experiment 2 amounted to d_z = 0.98. A sample size of about 13 participants ensures a power of 90% with an alpha of 5% to detect this effect size. Because of counterbalancing and potential exclusion of participants, we recruited a sample of 16 participants. Data were treated and analyzed as in Experiment 2. One participant committed at least 50% commission errors in one of the design cells and was excluded. We excluded post-error trials (17.3%). Other errors than commission errors in the Tone Task 1 were excluded (0.3%) before analyzing error rates of the Tone Task 1. Error rate analysis of the (Dis)honest Task 2 was restricted to trials with correct tone responses and we then excluded other errors than commission errors of the (Dis)honest Task 2 (1.0%). Only correct trials with inter-response intervals above 100 ms (2.3% excluded) and RTs within 2.5 *SD*s of the corresponding cell mean (3.7% excluded) entered RT analyses of both tasks. Descriptive statistics of the error rates are presented in Table 12 and of the RTs in Table 13 in Appendix 2 and in Figure 25.

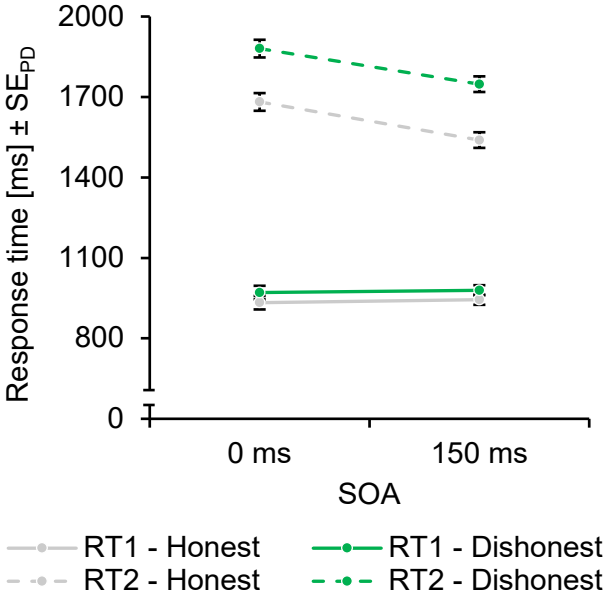

Figure 25 | Mean RTs of the Tone Task 1 (RT1; solid lines) and of the (Dis)honest Task 2 (RT2; dashed lines) of the additional experiment. Light gray lines constitute honest trials and dark green (dark gray) lines dishonest trials. Error bars represent standard errors of paired differences (SE_{PD}), computed separately for each SOA and task.

Tone Task 1 │ Tone RTs showed a nonsignificant trend toward slower responses in dishonest compared to honest trials (Δ = 36 ms), $F(1, 14) = 3.84$, $p = .070$, $\eta_p^2 = .22$, and the other effects were also not significant, Fs < 1. Error rates were higher with short than long SOAs, $F(1, 14) = 6.12$, $p = .027$, $\eta_p^2 = .30$ (Δ = 1.1%). The main effect of intention and the interaction were not significant in error rates, Fs < 1.37, ps > .261.

(Dis)honest Task 2 │ Dishonest responses were slower than honest responses (Δ = 204 ms), $F(1, 14) = 51.75$, $p < .001$, $\eta_p^2 = .79$, and with short SOAs than with long SOAs (Δ = 137 ms), $F(1, 14) = 77.57$, $p < .001$, $\eta_p^2 = .85$. The interaction of both factors was not significant, $F < 1$. Dishonest responses were less accurate than honest responses (Δ = 10.9%), $F(1, 14) = 66.67$, $p < .001$, $\eta_p^2 = .83$. The main effect of SOA and the interaction were not significant in error rates, Fs < 1.

Appendix 4: Translated mission instructions of Experiments 8 to 11

[Front page of all experiments]

Dear participant,

Please engage in all of the following activities in the given sequence. To ensure that the experiment can be completed successfully, it is important that you do not share with the experimenter which activities you engaged in.

Please engage in the following activities:

1. Sit down at the desk.
2. Take a sheet of paper and the pen from the box on the desk.
3. Draw a triangle and a circle on this sheet of paper.
4. Put the pen back in the box on the table.
5. Tear the sheet of paper in half.
6. Hide one piece of the sheet of paper **under** the stack of paper in the box on the desk and the other one in the box **below** the desk.

Please turn the page when you have engaged in all of the activities!

[Back page of Experiments 8 to 10]

Thank you for performing all activities. In a moment, you are going to be questioned about these activities in a computerized inquiry. The majority of participants had to engage in different activities and had to be honest about them in the following inquiry. In contrast, you are on a special mission and have to hide the true activities you just engaged in. That helps us to learn more about lie detection.

Accordingly, you are going to deny that you used the sheet of paper, the pen, and the boxes. Instead, you are going to pretend that you engaged in other activities. Hence, you need an alibi. Your alibi activities are those that the majority of participants experienced. You are going to pretend that you

- switched on the computer,
- used the USB stick,
- opened a file called “table”,
- wrote an email,
- and sent the file via email.

Please do not engage in any of these activities!

In the inquiry, you are going to be asked about activities you could or could not have engaged in today (**routine questions**) and about your secret activities in this room (**mission questions**). In addition, the color of the questions is going to indicate how you have to respond: either **honestly** or **oppositely to that honest response**.

The following is very important: You are going to follow these instructions exactly as told when you respond to **routine questions**. When you have engaged in the activity today, you respond with „yes“ when you are to respond honestly and with “no” when you are to respond oppositely. When you have not engaged in the activity today, you respond with “no” when you are to respond honestly and with “yes” when you are to respond oppositely.

When you encounter a **mission question**, however, you always have to lie when the color indicates to be honest, similar as a criminal would respond to the police. Accordingly, you always need to pretend that you have engaged in the alibi activities and have not used the pen, paper, and box. Example: When you are asked whether you wrote an email, you respond with “yes” when you are to respond honestly and with “no” when you are to respond oppositely. When you are asked whether you hid a sheet of paper, you respond with “no” when you are to respond honestly and with “yes” when you are to respond oppositely.

Please take your time to memorize these instructions and the activities of your alibi as you will need this information shortly. When you are ready, insert this letter in the box under the table. Then go to room H9 for the inquiry. Do not talk about your activities to the experimenter.

[Back page of Experiment 11]

Thank you for performing all activities. In a moment, you are going to be questioned about these activities in a computerized inquiry. Most other participants had to engage in different activities and were not allowed to reveal them in the following inquiry. In contrast, you are on a special mission and have to admit the true activities you just engaged in. That helps us to learn more about lie detection.

Accordingly, you are going to admit that you used the sheet of paper, the pen, and the boxes. The majority of participants engaged in the following activities. They

- switched on the computer,
- used the USB stick,
- opened a file called “table”,
- wrote an email,
- and sent the file via email.

Please do not engage in any of these activities!

In the inquiry, you are going to be asked about activities you could or could not have engaged in today (**routine questions**) and about your secret activities in this room (**mission questions**). In addition, the color of the questions is going to indicate how you have to respond: either **honestly** or **oppositely to that honest response**.

When you respond to **routine questions** and you have engaged in the activity today, you respond with „yes“ when you are to respond honestly and with “no” when you are to respond oppositely. When you have not engaged in the activity today, you respond with “no” when you are to respond honestly and with “yes” when you are to respond oppositely.

The same applies to **mission questions**. Example: When you are asked whether you wrote an email, you respond with “no” when you are to respond honestly and with “yes” when you are to respond oppositely. When you are asked whether you hid a sheet of paper, you respond with “yes” when you are to respond honestly and with “no” when you are to respond oppositely.

Please take your time to memorize these instructions and the activities of your alibi as you will need this information shortly. When you are ready, insert this letter in the box under the table. Then go to room H9 for the inquiry. Do not talk about your activities to the experimenter.