



## UvA-DARE (Digital Academic Repository)

### The Achilles' heel of the truth bias? High personal stakes reduce vulnerability to false information

Pantazi, M.; Klein, O.; Kissine, M.

**DOI**

[10.1002/ejsp.3086](https://doi.org/10.1002/ejsp.3086)

**Publication date**

2024

**Document Version**

Final published version

**Published in**

European Journal of Social Psychology

**License**

CC BY-NC-ND

[Link to publication](#)

**Citation for published version (APA):**

Pantazi, M., Klein, O., & Kissine, M. (2024). The Achilles' heel of the truth bias? High personal stakes reduce vulnerability to false information. *European Journal of Social Psychology*, 54(6), 1416-1429. <https://doi.org/10.1002/ejsp.3086>

**General rights**

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

**Disclaimer/Complaints regulations**

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

*UvA-DARE is a service provided by the library of the University of Amsterdam (<https://dare.uva.nl>)*

## RESEARCH ARTICLE

# The Achilles' heel of the truth bias? High personal stakes reduce vulnerability to false information

Myrto Pantazi<sup>1,2</sup> | Olivier Klein<sup>2</sup> | Mikhail Kissine<sup>3,4,5</sup>

<sup>1</sup>Social Psychology Programme Group, University of Amsterdam, Amsterdam, The Netherlands

<sup>2</sup>Center for Social and Cultural Psychology, Université Libre de Bruxelles, Brussels, Belgium

<sup>3</sup>Center of Linguistic Research LaDisco, Université Libre de Bruxelles, Brussels, Belgium

<sup>4</sup>Department of Philosophy, Classics, History of Art and Ideas, University of Oslo, Oslo, Norway

<sup>5</sup>ULB Neuroscience Institute, Université libre de Bruxelles, Brussels, Belgium

## Correspondence

Myrto Pantazi, Social Psychology Programme Group, University of Amsterdam, Nieuwe Achtergracht 129, Postbus 15900, 1001 NK, Amsterdam, The Netherlands.  
Email: [m.pantazi@uva.nl](mailto:m.pantazi@uva.nl)

Olivier Klein and Mikhail Kissine share the last authorship.

## Funding information

Université libre de Bruxelles : Mini-ARC Project grant "At the Sources of Faith"; BA/Leverhulme small research grant, Grant/Award Number: SRG19\190779; Fonds David et Alice Van Buuren; Jaumotte-Demoulin Foundation; Wallonia-Brussels Federation Concerted Research Action grant "The Socio-Cognitive Impact of Literacy"

## Abstract

While, by default, people tend to believe communicated content, it is also possible that they become more vigilant when personal stakes increase. A lab ( $N = 72$ ) and an online ( $N = 284$ ) experiment show that people make judgements affected by explicitly tagged false information and that they misremember such information as true – a phenomenon dubbed the 'truth bias'. However, both experiments show that this bias is significantly reduced when personal stakes – instantiated here as a financial incentive – become high. Experiment 2 also shows that personal stakes mitigate the truth bias when they are high at the moment of false information processing, but they cannot reduce belief in false information a posteriori, that is once participants have already processed false information. Experiment 2 also suggests that high stakes reduce belief in false information whether participants' focus is directed towards making accurate judgements or correctly remembering information truthfulness. We discuss the implications of our findings for models of information validation and interventions against real-world misinformation.

## KEYWORDS

belief formation, information validation, language, misinformation, truth bias

## 1 | INTRODUCTION

How much did you learn today because others told you so, and how much did you learn through first-hand experience? For most of us, the answers are likely 'everything' and 'nothing', respectively. Language

is one of the main means of knowledge and belief exchange and formation (Clément, 2010; Lackey, 2007), as it greatly increases the amount of information people can acquire in their lifetime. However, language may be less directly connected to factual truth, especially as compared to information acquired through other senses, like vision

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2024 The Author(s). *European Journal of Social Psychology* published by John Wiley & Sons Ltd.

(Zahavi, 1993). How people validate – if at all – verbally transmitted information is an old philosophical question, also addressed by experimental work in psychology and linguistics (Clark, 1974; Gilbert et al., 1990; Richter & Rapp, 2014). In this paper, we both confirm previous findings that by default people tend to believe verbally communicated contents, even if these are explicitly tagged as false but also identify high personal stakes as a novel factor that may reduce vulnerability to false information.

In doing so, we bring together two major theoretical perspectives on how and why people believe and validate communicated information.

On the one hand, cultural evolution and social learning theories propose that vigilance or epistemic mechanisms *must* have evolved to protect us from deception – precisely because verbal communication makes people vulnerable to deception or unintentional misinformation (Henrich, 2009; Sperber et al., 2010). On the other hand, it is likely that language emerged in small groups bound by kinship, within which cooperation drastically minimized deceptive behaviours (Hurlford, 2007; McNally & Jackson, 2013; Reber & Unkelbach, 2010; Serota et al., 2010). In such cooperative contexts, by default believing communicated contents – rather than spending time and energy to check their accuracy – appears an efficient adaptive communication mechanism (Forgas & Baymeister, 2019; Kissine & Klein, 2013; Levine, 2014; Millikan, 2005).

While some experimental (Hasson et al., 2005; Richter et al., 2009) and modelling studies (Henrich, 2009) have revealed conditions under which addressees may be vigilant and resistant to false information, multiple studies also indicate that, by default, people tend to believe communicated information. This assumption is consistent with dominant pragmatics theories (Grice, 1975) but also corroborated empirically. For example, people are not particularly good at detecting lies (Bond & Depaulo, 2006; Ekman et al., 1999; ten Brinke et al., 2014). They are also easily suggestible to incorrectly presupposed information (Loftus, 2005) and prone to form false memories (Fiedler et al., 1996; Garcia-Marques et al., 2010).

The strongest evidence for the permeability of beliefs to communicated information comes from paradigms where participants are presented with statements containing *explicitly labelled* false information, tagged by red font colour (Gilbert et al., 1993), the word ‘false’ (Gilbert et al., 1990) or the speaker’s gender (Pantazi et al., 2018). These paradigms are a strict test of people’s proneness to believe false information because they do not rely on participants’ prior knowledge or critical thinking and cognitive skills to assess the veracity of provided information, but instead offer a readily available and explicit cue signalling that the information should be disbelieved. Yet, even in such paradigms, participants base their judgements on the false information and tend to misremember false information as true to a much greater extent than they misremember true information as false, a phenomenon dubbed the truth bias.

While early accounts of this phenomenon were based on the fact that cognitive load made participants more prone to believe explicitly labelled false information, more recent research showed that this tendency robustly appeared even in the absence of distraction (Pantazi et al., 2018). These findings suggest a more generalized truth bias

than suggested by earlier work, one that is independent of cognitive load and depth of processing. Crucially, the truth bias has been shown to be robust to certain conditions that enhance vigilance. For example, factors that could be expected to improve how diligently people process and validate information, such as high accountability (Lerner & Tetlock, 1999) or year-long expertise (Klein et al., 2017), as is the case for professional judges, have failed to reduce this bias (Pantazi et al., 2020).

Since deeper processing alone does not seem to have a significant impact in and of itself, a crucial question, then, is what factors may reduce people’s tendency to believe false information. This question is particularly timely as misinformation is a pressing societal issue from public health to politics and education (Lazer et al., 2018). The two lines of research presented above are jointly in line with a model of verbal communication characterized by a default truth bias, ‘a tendency to perceive messages as truthful’ (Bond & Depaulo, 2006, p. 217). This bias may have co-evolved with epistemic vigilance mechanisms that protect hearers from falsehoods (Kissine & Klein, 2013). Because the activation of these vigilance mechanisms has a cost, it is rather exceptional and contingent on contextual circumstances. In line with this model, in this paper we predict that high personal stakes may increase vigilance towards inaccurate content, a hypothesis consistent with cultural evolution and social learning studies showing that increased payoffs increase people’s reliance on social meta-information (Arbilly et al., 2011; Muthukrishna et al., 2016, 2012).

Our theoretical frame has similarities with the truth-default theory (Levine, 2014). However, the latter is a theory of the detection of deceitful intent and centres on people’s assessment of meta-information (communicators’ intents). Our focus is rather on the socio-cognitive factors that determine the validation of verbally communicated content itself (instead of meta-information). This distinction is crucial as people readily utilize primary content available in their environment without successfully assessing available ‘meta-information’ concerning its history, sources and accuracy (Fiedler, 2012).

Our position also differs from Gilbert’s ‘Spinozan’ model, according to which belief formation is an automatic, inevitable step of information processing (Gilbert et al., 1990, 1993). First, the ‘automatic belief’ Spinozan account was based on evidence that cognitive load made people more likely to believe false information. More recent evidence suggests that not only are people truth-biased in the absence of distraction or high cognitive load but that such factors do not actually determine the magnitude of the truth bias (Pantazi et al., 2018). Thus, we propose that the tendency to believe is not an intrinsic step of the cognitive process of statement comprehension but rather a strong default option. Second, in line with Sperber et al. (2010) and with empirical studies highlighted above, we recognize that epistemic vigilance plays a central role in efficient communication and we, thus, predict that the truth bias may be overridden. However, we expect such vigilance mechanisms to be exceptional and contingent on contextual circumstances. We posit that high personal stakes may be such a contextual factor that can lead people to resist false information.

To present participants with true and false information, we relied on a previously validated task, where participants make judgements based

on crime reports containing true and false information (Gilbert et al., 1993; Pantazi et al., 2018). To raise personal stakes, we used monetary rewards, a robust personal incentive (Heyman & Ariely, 2004). In Experiment 1, we asked whether the tendency to use false information to form judgements and to misremember false information as true is reduced by financially rewarding rejection of falsehood. In Experiment 2, we determined whether the effect of high personal stakes operates during information encoding or whether it reflects belief correction during judgement and recall tasks.

## 2 | EXPERIMENT 1

### 2.1 | Method

We borrowed the materials from Pantazi et al. (2018; Experiment 1). Participants listened to two fictional crime reports, each containing 27 true and seven false statements with truth values assigned by their source: Participants were told that the male (female) speaker provided true (false) information, while the female (male) speaker provided false (true) information.<sup>1</sup> According to pre-tests (Pantazi et al., 2018), the true information in the two reports described crimes of equivalent gravity, that is, two armed robberies, but the false information in one report aggravated the crime (e.g., described the use of violence by the perpetrator) while the false information in the other attenuated it (e.g., signalled good prior record of the perpetrator). The full-text reports appear in the [Supporting Information](#). This experiment also involves a 'high-stakes' group that was offered extra financial incentives to make accurate judgements based on the reports and a control group that was only offered the participation payment.

#### 2.1.1 | Measures

As per prior research (Gilbert et al., 1993; Pantazi et al., 2018, 2020), we employed two complementary measures of belief in false information or 'truth bias'. First, participants proposed a prison term per perpetrator from 0–10 years and judged them on several dimensions: prison term severity (0 'extremely lenient'–7 'extremely strict'), dangerousness (0 'not at all dangerous'–7 'extremely dangerous') and probability to recidivate (0 'extremely unlikely'–7 'extremely likely'). The difference in the judgements of the two perpetrators indicates to what extent participants believed the false (aggravating vs. attenuating) information. The design for this judgement-based truth-bias measure was 2 (*false information*: aggravating vs. attenuating)  $\times$  2 (*group*: 'high-stakes' vs. control) mixed design.

Second, participants took a memory test containing 4 true, 4 false statements and 16 new statements per crime report which they had to classify as 'true', 'false' or 'new'. Unlike in deception-detection research, we measured the truth bias by comparing the proportion of believed false statements to the proportion of disbelieved true statements. This

memory-based truth-bias index was expected to be larger in the control than in the high-stakes groups. The design for the analysis of the memory-based truth-bias measure was a 2 (*group*: 'high-stakes' vs. control)  $\times$  2 (*statement type*: true vs. false)  $\times$  3 (*response type*: correctly identified, confounded with the opposite type and misidentified as 'new' (see the Analysis section for details). Thus, the two combined measures robustly measured not only whether participants remembered the false information to be true but also whether they actually believed the false information, as reflected in the judgements they made.

To verify that high stakes indeed increase participants' resistance to false information and did not simply change response biases in the memory test, we also present an analysis of the responses for the new items following Pantazi et al. (2018). Lastly, in the [Supporting Information](#), we present analyses on participants' reaction time in the judgement and memory tasks, which we analysed using similar models as those used for the main judgement and memory truth-bias index analyses.

#### 2.1.2 | Participants and procedure

The experiment received ethics approval from a University Ethics Review Board. We conducted an a priori power analysis based on Hogarth et al. (1991; Experiment 1). For the lenient conditions in their experiments (where financial incentives were expected to play a role), these authors report mean performance differences in a cognitive task between an incentivized and a non-incentivized condition amounting to  $d = .92$  amounting to a between-groups effect of  $f = 0.45$  (DeCoster, 2012). Assuming that in our paradigm, participants form their judgements based on their belief in the false information as also attested by the correlated nature of the judgement and memory measures, we expected incentives to have analogous effects on the two measures. According to G\*power (Faul et al., 2007) to detect a similar effect of incentives on either the judgement or the memory measure truth-bias index in a one-way analysis of variance (ANOVA)<sup>3</sup> with two groups, and .90 power at the .05 alpha level, we would need 68 participants. We recruited 72 Belgian volunteers ( $M_{\text{age}} = 21$ ,  $SD_{\text{age}} = 2.31$ ; 51 female, 17 male, four with unreported gender), randomly assigned to two groups. Participants were invited to participate in a study on 'Language and Communication' where they could earn between 5 and 15 euros via in-camp flyers, e-mail lists and relevant social media groups.

<sup>2</sup> We chose this study because its quota-based incentives scheme is similar to ours. Using Lakens' (2013) calculator for between-groups effects, we calculated the size of the mean difference between the incentivized and non-incentivized groups in the lenient condition separately for Round 1 and Round 2 based on their reported mean performance (Round 1:  $M = 331$  for the incentivized,  $M = 263$  for the control; Round 2:  $M = 386$  for the incentivized,  $M = 314$  for the control), standard deviation (Round 1:  $SD = 73$  for the incentivized,  $SD = 74$  for the control; Round 2:  $SD = 74$  for the incentivized,  $SD = 84$  for the control) and sample sizes ( $N = 20$  per group). This revealed effect sizes of  $d = .92$  and  $d = .90$ . We now realize that this effect likely overestimates the actual impact of incentives even though our power analysis at the time of recruitment was based on it.

<sup>3</sup> We used an ANOVA power analysis given the complexity of calculating power for mixed models. If anything, the mixed-model analyses that we report have more statistical power than an ANOVA (Quené & van den Bergh, 2008).

<sup>1</sup> The gender/truth value tag combinations were counterbalanced across participants.

Participants came to the lab in groups of a maximum of eight per session and completed the tasks individually in computer booths. After signing an informed-consent form, they read the instructions informing them that they would receive 5 euros for participating, that they would listen to two crime reports, and that, in each report, one speaker would provide true information while the other would provide false information. All participants were urged to listen as carefully as possible because they would be asked related questions and would have to come up with a fair prison term per perpetrator taking into account the correct information in favour and against the perpetrator. The 'high-stakes' group was additionally promised a generous (see Khan et al., 2020) bonus payment of 5 euros per report if they proposed such a fair prison term. Participants listened to the two reports (approximately 2 min each) through headphones and responded to the judgement and memory questions and demographic questions administered through *E-prime* (2.0).<sup>4</sup> The order of the reports was counterbalanced, so that half participants listened to the aggravating report first and the other half to the attenuating report first. The judgement question always preceded the memory questions. All participants were paid 15 euros, the maximum a participant could theoretically earn regardless of their group and responses, and were fully debriefed.

## 2.2 | Results

Data and analysis scripts for Study 1 are available on the study's OSF link ([https://osf.io/u854t/?view\\_only=b41e951c9e364cc5af1bdbd4c629067f](https://osf.io/u854t/?view_only=b41e951c9e364cc5af1bdbd4c629067f)).

### 2.2.1 | Judgements

We treated the four judgements per perpetrator ( $\alpha = .86$  for aggravated;  $\alpha = .88$  for attenuated) as repeated measures of the truth-bias index – belief about the severity of the crime – and analysed judgements using a mixed model with *false information*, *group* and their interaction as fixed factors and intercepts of subjects, judgement and subject-by-judgement as random effects (*mixed* command; SPSS 23). There was one missing value and no outliers based on the criterion of three median absolute deviations from the median (Leys et al., 2013). Cohen's *d* for repeated measures is calculated and reported following Lakens (2013).

Figure 1 shows average judgements per condition. Participants were truth-biased as they judged the (falsely) aggravated perpetrator more severely than the attenuated one ( $F(1, 498.20)$ <sup>5</sup> = 5.06,  $p = .025$ ,  $d_{rm} = 0.13$ ), but this effect was moderated by group ( $F(1,$

498.20) = 5.91,  $p = .025$ ; group alone did not have a main effect ( $F(1, 69.86) = 0.47$ ,  $p = .494$ ). Bonferroni-corrected pairwise comparisons showed that while participants in the control group were affected by the false information ( $t(498.20) = 3.26$ ,  $p = .001$ , *Meandiff* = 0.74, 95% CI [0.29, 1.20],  $d_{rm} = 0.30$ ) participants in the high-stakes condition were not ( $t(498.20) = .13$ ,  $p = .897$ , *Meandiff* = -0.02, 95% CI [-0.46, 0.40],  $d_{rm} = 0.008$ ).<sup>6</sup>

### 2.2.2 | Memory

#### *True and false statements*

There were three types of responses for the true and false statements in the memory test: correctly identified, confounded with the opposite value (true mistaken for false; false mistaken for true) and misidentified as 'new'. We treated these as three repeated measures of each statement, and in each trial coded them as '1' if they reflected participants' actual response and '0' otherwise. The response data were thus binomial. We then ran a generalized linear mixed model for binomial data (GENLINMIXED procedure, logit transformed – LOGIT link Target\_Option – in SPSS 23; see Quené & van den Bergh, 2008). We included *response type*, *statement type*, *group* and all two and three-way interactions as fixed factors. We also included intercepts of subjects and statements as random factors.<sup>7</sup>

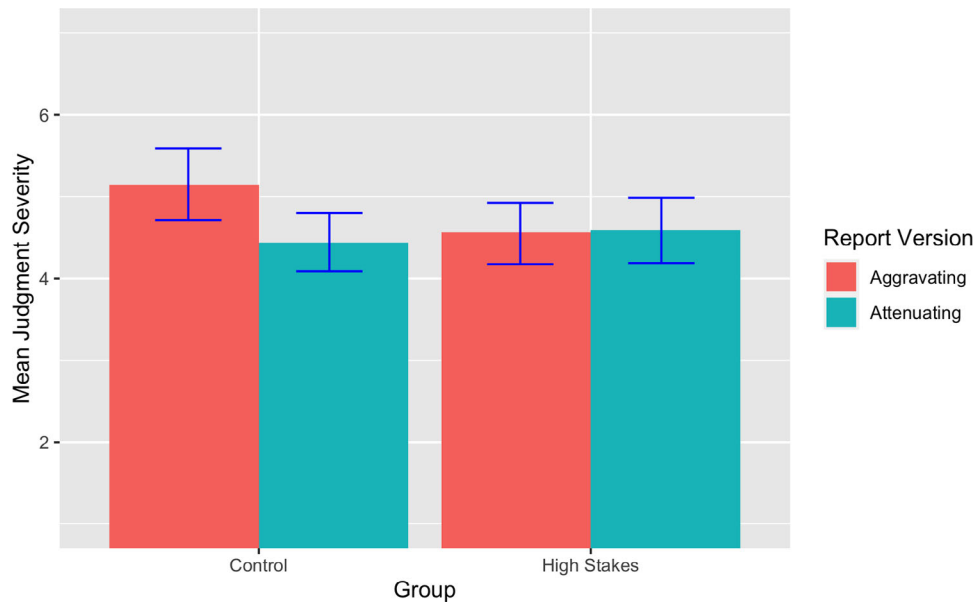
Percentage responses are presented in Figure 2. A main effect of *response type* ( $F(2, 3443) = 73.26$ ,  $p < .001$ ) indicated that participants understood the instructions and performed decently, identifying more true and false statements correctly ( $M = 0.66$ ,  $SD = 0.47$ ) than they confounded ( $M = 0.14$ ,  $SD = 0.34$ );  $t(3444) = 13.54$ ,  $p < .001$ , *Meandiff* = 0.52, 95% CI [0.50, 0.54],  $d_{rm} = 1.40$ ) or misidentified as new ( $t(3444) = 11.10$ ,  $p < .001$ , *Meandiff* = 0.46, 95% CI [0.44, 0.48],  $d_{rm} = 1.13$ ). A *statement type* × *response type* interaction ( $F(2, 3443) = 12.35$ ,  $p < .001$ ) showed that false statements were more confounded with true than true statements with false ( $t(3444) = 2.28$ ,  $p = .022$ , *Meandiff* = 0.08, 95% CI [0.01, 0.15],  $d_{rm} = 0.20$ ). The three-way interaction was non-significant ( $F(2, 3443) = 2.74$ ,  $p = .065$ ), but Bonferroni-corrected pairwise comparisons indicated that false statements were more confounded than the true statements in the control ( $t(3443) = 4.23$ ,  $p < .001$ , *Meandiff* = 0.04, 95% CI [0.003, .08],  $d_{rm} = 0.13$ ) but not in the 'high-stakes' group ( $t(3443) = 1.54$ ,  $p = .124$ , *Meandiff* = 0.0, 95% CI [-0.04, 0.04],  $d_{rm} = 0.0$ ). There was no significant difference between the false statements misidentified as new between the two groups ( $t(3443) = .69$ ,  $p = .543$ , *Meandiff* = 0.02, 95% CI [-0.08, 0.04],  $d = .05$ ). There was no significant *group* × *response type* interaction ( $F(2, 3443) = 0.57$ ,  $p = .567$ ) to suggest that high stakes affected accuracy overall.

<sup>4</sup> After the primary tasks, participants completed an *N*-back task (Kane et al., 2007). Because this task could not affect the main measures and targeted a different research question, we omit it.

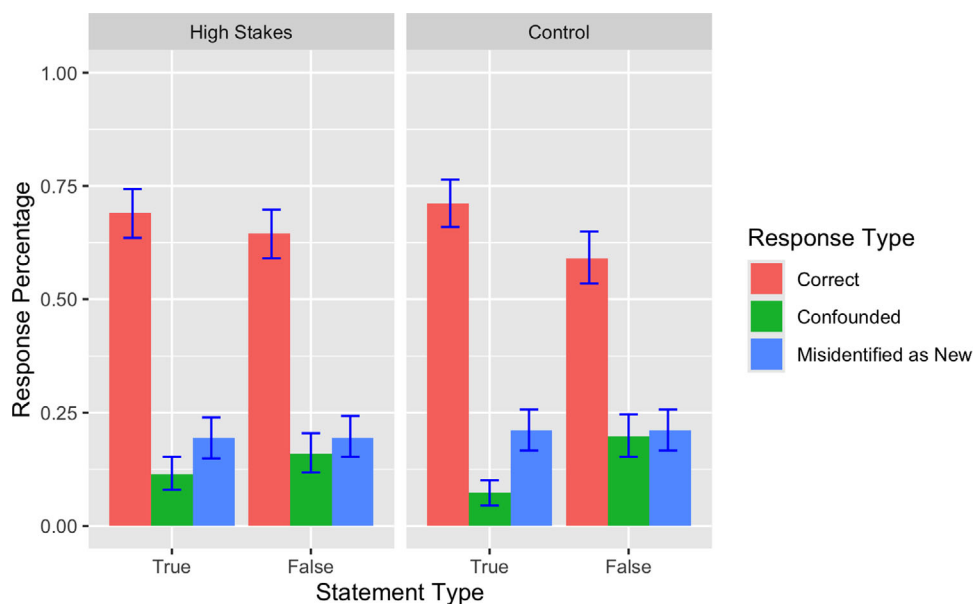
<sup>5</sup> Degrees of freedom in the judgement analyses are non-integers because in the mixed model design we used these were calculated based on the Satterthwaite approximation.

<sup>6</sup> Additional analyses of response times, reported in the Supporting Information do not suggest any difference in the time incentivized and control participants took to respond. (see p. S1).

<sup>7</sup> We further tested whether memory differed for false aggravating and attenuating, which we report in the supplementary materials for both studies (see p. S3 in the Supporting Information).



**FIGURE 1** Mean judgement severity per *report version* and *group* in Experiment 1. Error bars represent 95% confidence intervals.

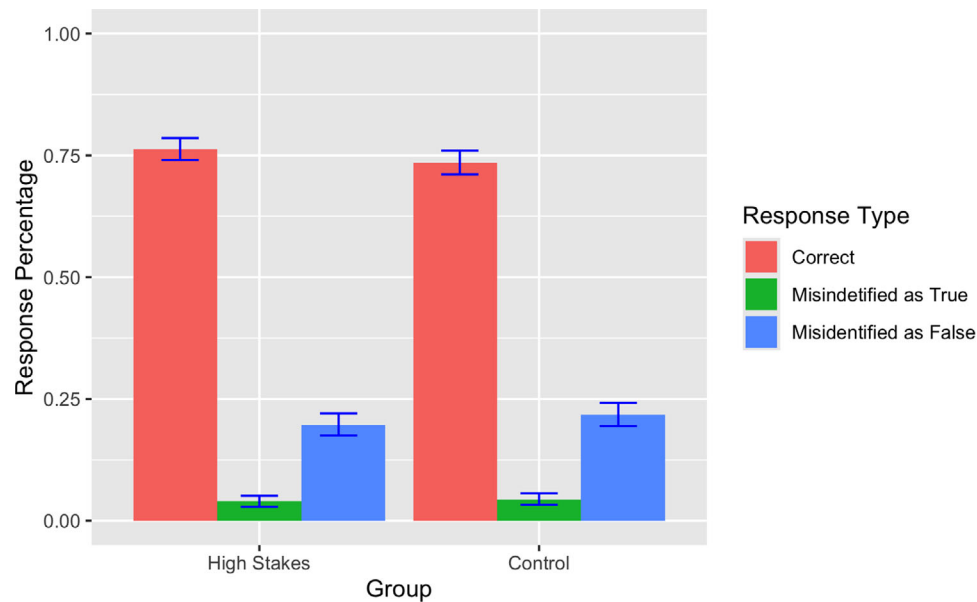


**FIGURE 2** Response percentages for the true and false statements per response type and group, in Experiment 1. Error bars represent 95% confidence intervals.

### New statements

Mean response percentages for the new items are presented in Figure 3. We ran a similar generalized linear mixed model for binomial data on the memory responses for the new items, with *response type* (correct vs. misidentified as 'true' vs. misidentified as 'false')  $\times$  *group* (high stakes vs. control) and their two-way interaction as fixed factors. Intercepts of subjects and statements were also included as random factors. There was only an effect of response type ( $F(1, 6906) = 1005.61, p < .001$ ), signalling that participants were predom-

inantly accurate in correctly identifying new items as new ( $M = 0.75, SD = 0.43$ ) at a significantly higher rate than they misidentified them as true ( $M = 0.04, SD = 0.20, t(6906) = 68.04, p < .001, Meandiff = 0.71, 95\% \text{ CI } [0.70, 0.72], d_{rm} = 2.11$ ) or false ( $M = 0.21, SD = 0.41, t(6906) = 43.70, p < .001, Meandiff = 0.54, 95\% \text{ CI } [0.26, 0.82], d_{rm} = 1.28$ ). Participants also appeared to misidentify new statements as true to a lesser extent than they misidentified them as false ( $t(6906) = -15.82, p < .001, Meandiff = 0.17, 95\% \text{ CI } [0.15, 0.19], d_{rm} = 0.40$ ). The group did not exert a significant effect ( $F(1,$



**FIGURE 3** Response percentages for the new statements per response type and group. Error bars represent 95% confidence intervals.

6906) = 0.03,  $p = .864$ ), and neither did its interaction with response type ( $F(1, 6906) = 2.01, p = .134$ ).

### 2.2.3 | Judgement–memory correlation

Participants' tendency to confound false statements with true ones (the difference in 'confounded' responses between true and false statements) correlated positively with the impact of false information on their judgements (i.e., the difference in judgements for the aggravated vs. the attenuated perpetrator;  $r(72) = .239, p = .043$ ).

## 2.3 | Discussion

In line with our hypotheses and previous studies, Experiment 1 attested to the truth bias: Control participants based their judgements on false information and tended to misremember false information as true more than they misremembered true information as false. The correlation between the two measures suggested a consistent within-subject tendency to believe false statements (see Gilbert et al., 1993; Pantazi et al., 2018). These two tendencies, however, disappeared when participants received financial rewards for ignoring false information, indicating that high personal stakes may increase vigilance towards false content. An absence of observed differences across groups in the response pattern for the new items in the memory test rules out an alternative explanation of the memory truth-bias pattern based on a change in response bias in the high-stakes conditions. In addition, recall that when participants were wrong on new statements, they were more likely to misidentify them as previously seen false information, rather than previously seen true information. This result implies that the truth bias does not merely reflect a response

bias that could apply to any statement regardless of its novelty. Rather, the results of the old and new statements considered jointly suggest that participants truly believed the false statements they reported to be true. Overall, Experiment 1 strongly supports the idea that high personal stakes reduce vulnerability to false information.

## 3 | EXPERIMENT 2

In addition to replicating Experiment 1, Experiment 2 aimed to further confirm that high stakes genuinely elicit more vigilant information processing as opposed to simply prompting participants to adjust their responses after having processed the false information. For this reason, we manipulated the moment when stakes increased so that they either preceded or succeeded in the presentation of the true and false information. Moreover, in Experiment 1, the high-stakes manipulation was intricately linked to participants' judgements. Although Experiment 1 suggested that both judgement and memory measures were contingent on high stakes, an impression formation goal may make people remember information better compared to a memorization goal (Chartrand & Bargh, 1996; Fiedler et al., 2009). It is important to ensure, therefore, that high stakes, instead of instructions to form impressions about the defendants, attenuated the truth bias. Experiment 2 involved both groups where the high-stakes manipulation was linked to participants' judgements and groups where the high-stakes manipulation was linked to participants' memory. This allowed us to test whether incentives mitigate the truth bias regardless of whether participants focus on their judgements or their memory.

Experiment 2 also examined two alternative explanations of the findings of Experiment 1. First, the recruitment ad in Experiment 1 suggested that participants could earn up to 15 euros. The observed pattern might then reflect a deteriorated performance of control

participants, following disappointment or resentment during the task for being offered 5 euros, instead of increased vigilance elicited in the high-stakes group. The study ad now only informed participants about their flat-rate payment. Second, in Experiment 1 true information outnumbered false information. Although the proportion of false versus true statements does not eliminate the truth-bias effect (Pantazi et al., 2018), this may have specifically encouraged participants to adopt a 'truth-biased' processing. To more stringently test our hypothesis, the reports participants read in Experiment 2 contained equal numbers of true and false statements.

### 3.1 | Method

Participants listened to reports that were similar to those used in Experiment 1 but contained an equal number of true and false statements (Pantazi et al., 2018; Experiment 3). The full text of the reports appears in the [Supporting Information](#). As in Experiment 1, the truth value of the information in the report was signalled by the gender of the speaker while participants listened to the true and false information in the reports, and false information was aggravating in one report and attenuating in the other. This experiment involved four groups. The control group was the same as in Experiment 1 as they received no high-stakes manipulation. All other three groups involved a high-stakes manipulation. Participants in the 'Judgement-high-stakes-before' group were informed, before listening to the reports that they would receive a bonus payment if they proposed a fair prison term. Thus, the high-stakes manipulation in this group took place *before* participants listened to the reports and was linked to participants' judgements. Participants in the 'Judgement-high-stakes-after' group were offered the same incentive but after listening to both reports but before completing the judgement and memory tasks. Finally, the 'Memory-high-stakes' group was promised, before listening to the reports, that they would receive a bonus if they accurately remembered which statements were true and which were false. Thus, the high-stakes manipulation in this group took place before participants listened to the reports and was linked to participants.

#### 3.1.1 | Measures

The two main measures, judgement and memory, were the same as in Experiment 1. We also included four questions measuring motivation to form accurate judgements or memories (e.g., How much effort did you dedicate to (a) proposing a fair prison term (b) distinguishing true information from the false?). Lastly, given that this study was run online through Qualtrics, we measured total completion time.<sup>8</sup>

<sup>8</sup> We also included a question on conspiracy beliefs. Because this question formed part of a related ongoing project and was not related to the present core hypotheses, we do not report its results here.

#### 3.1.2 | Participants and procedure

The experiment received ethics approval from a University Departmental Research Ethics Committee. The sample size calculation was based on a power analysis using the effect size from Experiment 1 and pre-registered on OSF ([https://osf.io/cd9hs/?view\\_only=44d11fe85d644a8a8d46f20ff9c12bea](https://osf.io/cd9hs/?view_only=44d11fe85d644a8a8d46f20ff9c12bea)).<sup>9</sup> Based on our previous studies, we expect an interaction of  $f = 0.135$  between the report version and group for the judgement measures. According to G\*Power, to detect such an effect with .9 power given four groups, two within-subject measures, and a .288 correlation between the within-subject measures we would need 284 participants. Likewise, we expected an interaction between statement type and group for the memory measure of  $f = 0.155$ . According to G\*Power, to detect an effect of this size with a power of .9, a repeated measures correlation of .098 (estimated from the previous studies), given the four groups and two measures, and a .05 alpha level, we would need 276 participants.

We recruited 287 participants through Prolific for an online study on 'Language and Communication' ( $M_{age} = 29.12$ ,  $SD_{age} = 29.12$ ; 163 male, 123 female; four with unreported gender). According to a sensitivity power analysis, with this sample we could detect a between-within interaction effect in a mixed ANOVA analysis as small as  $f = 0.15$  with .80 power, given alpha .05, the four groups, 2 within-subject measures and inter-item correlations as large as .288 (the correlation observed in the first study; G\*Power; Faul et al., 2007). Participants were informed that they would earn 3 euros and only those in the three high-stakes groups were informed about the possibility of earning three extra euros per prison term, contingent on their instructions. All participants ultimately received 9 euros.

The tasks and procedures were similar to Experiment 1 but implemented online through Qualtrics. The only two differences with Experiment 1 were that the 'Memory-high-stakes' group and half of the control group completed the memory task before the judgement and the four questions measuring participants' motivation, was asked after the main task. The report audio files were presented for a limited time to ensure that participants listened to each report once.

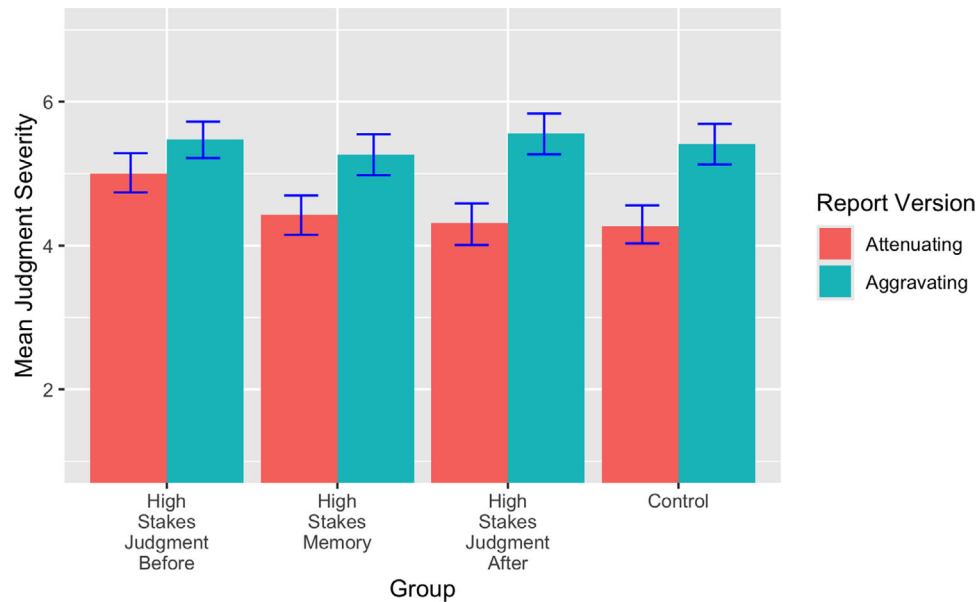
### 3.2 | Results

Data and analysis scripts are available on the OSF ([https://osf.io/cd9hs/?view\\_only=44d11fe85d644a8a8d46f20ff9c12bea](https://osf.io/cd9hs/?view_only=44d11fe85d644a8a8d46f20ff9c12bea)).

#### 3.2.1 | Judgements

Judgment distributions are presented in Figure 4. No outliers were detected (Leys, et al., 2013). We used the same mixed model as in Experiment 1 ( $\alpha = .86$  for the aggravated version judgements;  $\alpha = .88$

<sup>9</sup> While in our pre-registration we planned to run Bayesian analyses in case our frequentist analyses did not provide evidence for a difference between the memory and control groups, this proved uncalled for as our analyses provided robust evidence for a significant difference between the high-stakes memory and control groups.



**FIGURE 4** Mean judgement severity per report version and group in Experiment 2. Error bars represent 95% confidence intervals. HS, high stakes.

for the attenuated version judgements). As in Experiment 1, participants judged the aggravated perpetrator ( $M = 5.43$ ,  $SD = 2.40$ ) more severely than the attenuated one ( $M = 4.51$ ,  $SD = 2.35$ ,  $F(1, 2002) = 140.91$ ,  $p = .001$ ,  $Meandiff = 0.92$ , 95% CI [0.74, 1.10],  $d_{rm} = 0.30$ ), and this effect was moderated by the *group*. Four Bonferroni-corrected pairwise comparisons showed that while participants in all groups were affected by the false information, this difference was larger in the control ( $t(2002) = 3.07$ ,  $p = .002$ ,  $Meandiff = 1.14$ , 95% CI [0.82, 1.45],  $d_{rm} = 0.41$ ) and 'High-stakes-judgement-after groups' ( $t(2002) = 7.80$ ,  $p < .001$ ,  $Meandiff = 0.125$ , 95% CI [0.87, 1.63],  $d_{rm} = 0.39$ ) than in the 'Judgement-high-stakes-before' ( $t(2002) = 7.47$ ,  $p = .002$ ,  $Meandiff = 0.47$ , 95% CI [0.13, 0.81],  $d_{rm} = 0.16$ ) and 'Memory-high-stakes' groups ( $t(2002) = 5.38$ ,  $p < .001$ ,  $Meandiff = 0.84$ , 95% CI [0.47, 1.21],  $d_{rm} = 0.26$ ).

### 3.2.2 | Memory

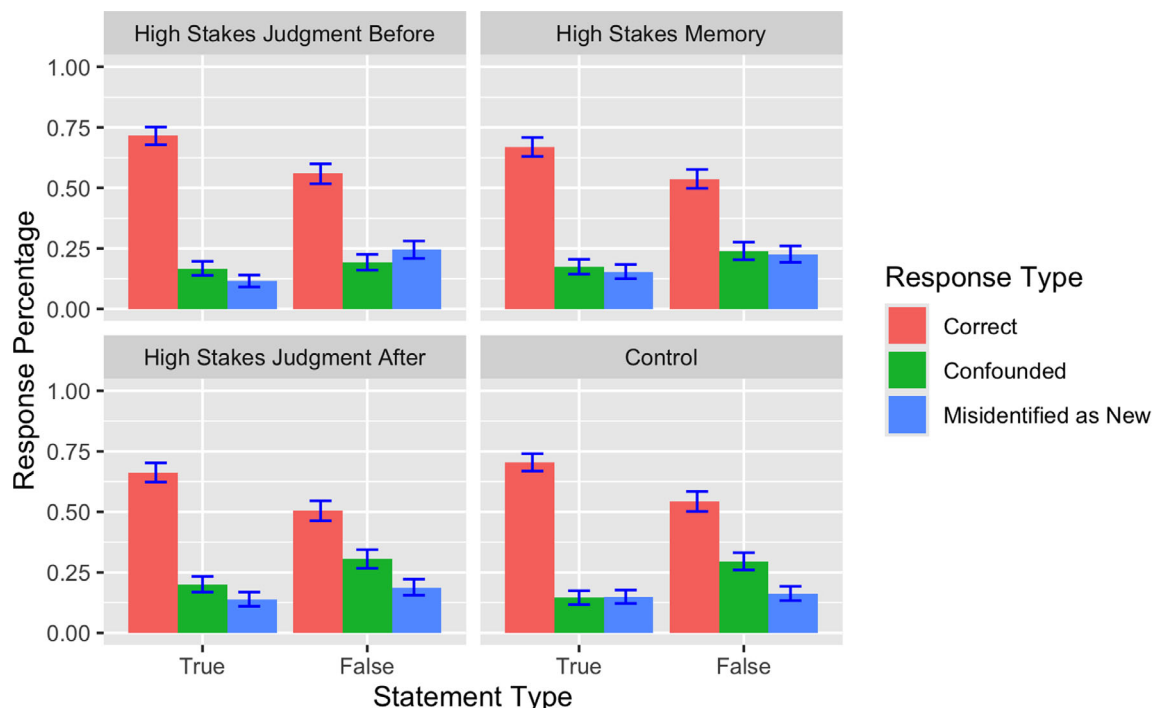
#### Old statements

As in Experiment 1, we ran a generalized linear mixed model for binomial data on the true and false statements contained in the memory test. Figure 5 presents participants' memory responses.

Again a main effect of *response type* ( $F(2, 13752) = 1102.50$ ,  $p < .001$ ) indicates that participants identified more true and false statements correctly ( $M = 0.61$ ,  $SD = 0.49$ ) than they confounded ( $M = 0.21$ ,  $SD = 0.41$ ;  $t(13752) = 42.52$ ,  $p < .001$ ,  $Meandiff = 0.40$ , 95% CI [0.39, 0.41],  $d_{rm} = 1.11$ ) or misidentified as new ( $M = 0.17$ ,  $SD = 0.38$ ;  $t(13752) = 48.44$ ,  $p < .001$ ,  $Meandiff = 0.44$ , 95% CI [0.43, 0.45],  $d_{rm} = 1.01$ ). Furthermore, statements were more confounded than misidentified as new ( $t(13752) = 4.96$ ,  $p < .001$ ,  $Meandiff = 0.04$ , 95% CI [0.03, 0.55],  $d_{rm} = 0.08$ ). A *statement type*  $\times$  *response type* interac-

tion ( $F(2, 13752) = 95.69$ ,  $p < .001$ ) indicated that true statements were more correctly identified ( $M = 0.69$ ,  $SD = 0.46$ ) than false ones ( $M = 0.54$ ,  $SD = 0.50$ ;  $t(13752) = 10.66$ ,  $p < .001$ ,  $Meandiff = 0.15$ , 95% CI [0.12, 0.18],  $d_{rm} = 0.23$ ).

False statements were more confounded with true ( $M = 0.26$ ,  $SD = 0.43$ ) than true statements with false ( $M = 0.17$ ,  $SD = 0.46$ ,  $t(13752) = 7.08$ ,  $p < .001$ ,  $Meandiff = 0.09$ , 95% CI [0.14, 0.20],  $d_{rm} = 0.14$ ). While the *group*  $\times$  *response type* interaction was significant ( $F(6, 13752) = 4.72$ ,  $p < .001$ ), it was qualified by a three-way interaction ( $F(6, 13752) = 4.56$ ,  $p < .001$ ) signalling between-group differences in responses of false versus true information. Bonferroni-corrected pairwise comparisons suggested that false statements were more likely to be confounded for true than true statements for false in the control ( $t(13752) = 6.36$ ,  $p < .001$ ,  $Meandiff = 0.16$ , 95% CI [0.13, 0.19],  $d_{rm} = 0.49$ ) the 'Judgement-high-stakes-after' groups ( $t(13752) = 4.09$ ,  $p < .001$ ,  $Meandiff = 0.11$ , 95% CI [0.08, 0.14],  $d_{rm} = 0.30$ ), and the 'Memory-high-stakes' groups ( $t(13752) = 4.09$ ,  $p < .001$ ,  $Meandiff = 0.07$ , 95% CI [0.02, 0.11],  $d_{rm} = 0.12$ ) groups, but not in the 'Judgement-high-stakes-before' ( $t(13752) = 1.15$ ,  $p = .250$ ,  $Meandiff = 0.01$ , 95% CI [-0.03, 0.05],  $d_{rm} = 0.02$ ). Furthermore, false statements were more likely to be confounded for true in the control group ( $M = 0.30$ ,  $SD = 0.46$ ) than in participants in both the 'Memory-high-stakes' ( $M = 0.24$ ,  $SD = 0.43$ ;  $t(13752) = 2.17$ ,  $p = .030$ ,  $Meandiff$  95% CI [0.009, 0.11],  $d = .13$ ) and the 'Judgement-high-stakes-before' ( $t(13752) = 4.12$ ,  $p < .001$ ,  $Meandiff$  95% CI [0.06, 0.16],  $d = .25$ ). Similarly, false statements were more likely to be confounded for true in the 'Judgement-high-stakes-after' group ( $M = 0.30$ ,  $SD = 0.46$ ) as compared to both the 'Memory-high-stakes' ( $M = 0.24$ ,  $SD = 0.43$ ;  $t(13752) = 2.47$ ,  $p = .014$ ,  $Meandiff$  95% CI [0.02, 0.12],  $d = .16$ ) and the 'Judgement-high-stakes-before' ( $t(13752) = 4.73$ ,  $p < .001$ ,  $Meandiff$  95% CI [0.06, 0.16],  $d = .26$ ). The 'Judgement-high-stakes-after' did not



**FIGURE 5** Response percentages for the true and false statements per response type and group, in Experiment 2. Error bars represent 95% confidence intervals. HS, high stakes.

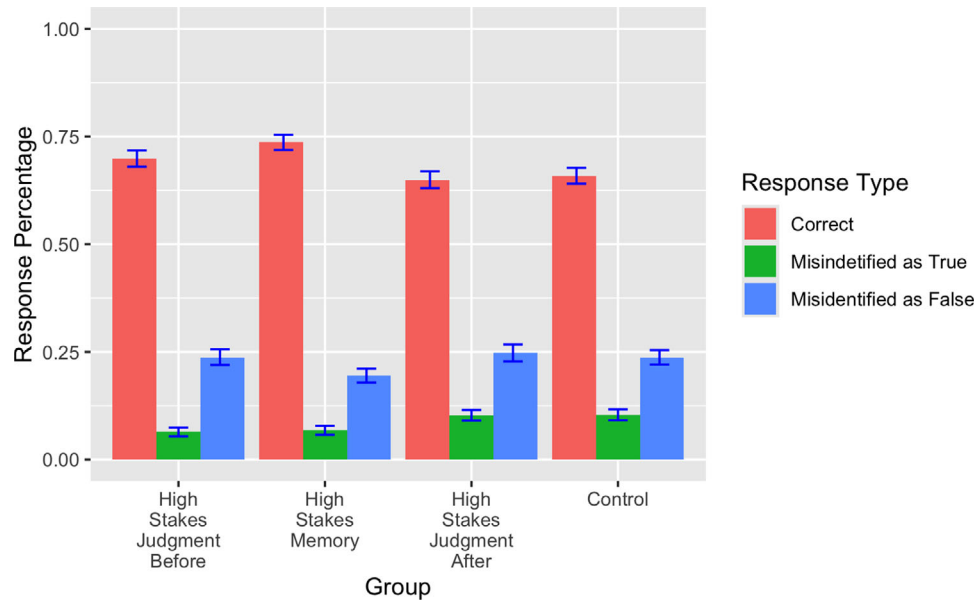
differ from the control group in that respect ( $t(13752) = 0.35, p = .727$ , *Meandiff* 95% CI  $[-0.05, 0.05]$ ,  $d = 0$ ) and neither did the 'Judgement-high-stakes-before' from the 'Memory-high-stakes' ( $t(13752) = 1.92$ ,  $p = .054$ , *Meandiff* 95% CI  $[0.002, 0.10]$ ,  $d = .12$ ).

The control group misidentified less false statements as 'new' compared to the 'Memory-high-stakes' ( $t(13752) = 2.70, p = .007$ , *Meandiff* 95% CI  $[-0.10, -0.01]$ ,  $d = .15$ ), and the 'Judgement-high-stakes-before' group ( $t(13752) = 3.64, p < .001$ , *Meandiff* 95% CI  $[-0.14, -0.04]$ ,  $d = .22$ ). The 'Judgement-high-stakes-after' group misidentified less false statements as 'new' compared to the 'Judgement-high-stakes-before' group ( $t(13752) = 2.43, p = .015$ , *Meandiff* 95% CI  $[-0.12, -0.02]$ ,  $d = .17$ ) but not the 'Memory-high-stakes' ( $t(13752) = 1.56, p = .127$ , *Meandiff* 95% CI  $[-0.08, 0.01]$ ,  $d = .10$ ) or the control group ( $t(13752) = 1.12, p = .259$ , *Meandiff* 95% CI  $[-0.02, 0.06]$ ,  $d = .05$ ). Lastly the 'Judgement-high-stakes-before' group and the 'Memory-high-stakes' did not differ in terms of the number of false statements they misidentified as new ( $t(13752) = 0.914, p = .360$ , *Meandiff* 95% CI  $[-0.09, 0.01]$ ,  $d = .09$ ). It is also noteworthy that the only between-group differences in responses on the true statements was that the control group confounded true statements less than the 'Judgement-high-stakes-after' group ( $t(13752) = 2.38, p = .017$ , *Meandiff* 95% CI  $[-0.10, -0.02]$ ,  $d = .16$ ).

#### New statements

Mean response percentages for the new items in Experiment 2 are presented in Figure 6. As in Experiment 1, we ran a generalized linear mixed model for binomial data on the memory responses for the new items, with *response type* (correct vs. misidentified as 'true' vs. misidentified as 'false')  $\times$  *group* (high stakes vs. control) and their two-way

interaction as fixed factors, while including also intercepts of subjects and statements as random factors. There was a main effect of response type ( $F(1, 27540) = 3251.50, p < .001$ ) qualified by a response type by condition interaction ( $F(1, 27540) = 16.78, p < .001$ ). Participants identified significantly more new statements as 'new' ( $M = 0.68, SD = 0.46$ ) than as 'true' ( $M = 0.08, SD = 0.28, t(27540) = 105.04, p < .001$ , *Meandiff* = 0.60, 95% CI  $[0.59, 0.50]$ ,  $d_{rm} = 1.40$ ) or 'false' ( $M = 0.23, SD = 0.42, t(27540) = 69.76, p < .001$ , *Meandiff* = 0.45, 95% CI  $[0.44, 0.45]$ ,  $d_{rm} = 1.00$ ) and more new statements as 'false' than 'true' (*Meandiff* = 0.15, 95% CI  $[0.14, 0.16]$ ,  $d_{rm} = 1.00$ ). The 'Judgement-high-stakes-before' ( $M = 0.70, SD = 0.46$ ) and the 'Memory-high-stakes' group ( $M = 0.73, SD = 0.44$ ) exhibited a significantly better identification of new statements than the 'Judgement-high-stakes-after' ( $M = 0.65, SD = 0.48; t(27540) = 3.54, p < .001, Meandiff$  95% CI  $[0.02, 0.08]$ ,  $d = .11; t(27540) = 6.35, p < .001, Meandiff$  95% CI  $[0.05, 0.10]$ ,  $d = .18$ , respectively) and the control group ( $M = 0.66, SD = 0.47; t(27540) = 2.91, p = .004, Meandiff$  95% CI  $[0.01, 0.07]$ ,  $d = .09; t(27540) = 5.78, p < .001, Meandiff$  95% CI  $[0.04, 0.09]$ ,  $d = .15$ , respectively). The performance of the 'Memory-high-stakes' group was also better than that of the 'Judgement-high-stakes-before' ( $t(27540) = 2.91, p = .004, Meandiff$  95% CI  $[0.004, 0.06]$ ,  $d = .07$ ). The 'Judgement-high-stakes-before' group tended less to identify new statements as true ( $M = 0.06, SD = 0.24$ ) than the 'Judgement-high-stakes-after' group ( $M = 0.10, SD = 0.30; t(27540) = -3.93, p < .001, Meandiff$  95% CI  $[0.02, 0.06]$ ,  $d = .15$ ) and the control group ( $M = 0.10, SD = 0.30; t(27540) = -4.17, p < .001, Meandiff$  95% CI  $[0.02, 0.06]$ ,  $d = .15$ ) but not the 'Memory-high-stakes' group ( $M = 0.07, SD = 0.25; t(27540) = -0.42, p = .676, Meandiff$  95% CI  $[0.004, 0.02]$ ,  $d = .04$ ). Similarly, the 'Memory-high-stakes' group misidentified less



**FIGURE 6** Response percentages for the new statements per response type and group, in Experiment 2. Error bars represent 95% confidence intervals. HS, high stakes.

new statements as true than the *Judgement-high-stakes-after* group ( $t(27540) = -3.51, p < .001, \text{Meandiff } 95\% \text{ CI } [0.01, 0.05], d = .10$ ) and the control group ( $t(27540) = -3.74, p < .001, \text{Meandiff } 95\% \text{ CI } [0.01, 0.05], d = .11$ ). Lastly, the *Memory-high-stakes* group also identified less new statements as 'false' ( $M = 0.19, SD = 0.40$ ) than all three other groups ( $M = 0.24, SD = 0.42, t(27540) = -3.46, p = .001, \text{Meandiff } 95\% \text{ CI } [-0.07, -0.03], d = .112$  for the *Judgement-high-stakes-before* group;  $M = 0.25, SD = 0.43, t(27540) = -4.26, p < .001, \text{Meandiff } 95\% \text{ CI } [-0.08, -0.04], d = .14$ , for the *Judgement-high-stakes-after* group;  $M = 0.24, SD = 0.42, t(27540) = -3.45, p = .001, \text{Meandiff } 95\% \text{ CI } [-0.07, -0.03], d = .14$  for the control group).

### 3.2.3 | Additional analyses

Again the difference in 'confounded' responses between true and false statements correlated positively with the difference in judgements for the two reports ( $r(285) = .435, p < .001$ ).

Participants' self-reported motivation to provide a fair prison term and distinguish true from false information were submitted to one-way ANOVAs. The condition had a significant effect on motivation to provide accurate judgements ( $F(3, 283) = 3.59, p = .014, \text{partial } \eta^2 = .037$ ). The 'Judgement-high-stakes-before' reported being more motivated ( $M = 9.63, SD = 1.94$ ) than the control group ( $M = 8.46, SD = 2.02; p = .15, \text{Meandiff} = 95\% \text{ CI } [0.52, 1.82], d = .59$ ), but not relative to the 'Memory-high-stakes' ( $M = 8.87, SD = 2.11; p = .121, \text{Meandiff} = 95\% \text{ CI } [0.09, 1.42], d = .37$ ) or the 'Judgement-high-stakes-after' groups ( $M = 9.25, SD = 1.64, \text{Meandiff} = 95\% \text{ CI } [-0.22, 0.98], d = .21, p > .999$ ). There was no effect on participants' motivation to form accurate memories ( $F(3, 283) = 1.17, p = .320, \text{partial } \eta^2 = .012$ ). At the same time, self-reported motivation for accurate memory correlated negatively

with participants' tendency to be affected by the false information in their judgements ( $r(285) = -.12, p = .037$ ) and their tendency to misremember false statements as true ( $r(285) = -.35, p < .001$ ). On the contrary, motivation for fair judgements did not predict the judgement ( $r(285) = .05, p = .363$ ) or the memory ( $r(285) = -.050, p = .411$ ) indices. Again, there was no evidence that the truth-bias reduction correlated with more time to complete the task ( $F(3, 286) = 1.48, p = .221, \text{partial } \eta^2 = .015$ ).

### 3.3 | Discussion

Experiment 2 provided additional evidence for a high-stakes effect on the truth bias in a new and different sample and based on better controlled materials and recruitment strategy than Experiment 1: While control participants exhibited a truth bias, this bias was lower if personal stakes were raised by a financial incentive. Importantly, the three different high-stakes groups in Experiment 2 allowed us to clarify the mechanisms of the effects of incentives. First, it was observed that high stakes led to a reduced judgement and memory-based truth bias when these were high before participants processed the false information, but no such effect was observed when stakes increased after participants had processed the information. This finding suggests that high stakes can increase resistance to false information only before one has processed the false information but cannot lead to significant belief correction after processing. The results observed in the *Judgement-high-stakes-after* group also refute the possibility of incentives-based demand characteristics operating in our experiment whereby participants who are not incentivized to respond correctly merely rely on false information to validate our hypotheses. In addition, increased personal stakes triggered epistemic vigilance regardless

of whether participants were urged to make accurate judgements or form accurate memories. This finding precludes the possibility that the incentive manipulation is tied to an information processing goal. Rather, it allows for a more generalized conclusion of the effect of incentives on resistance to misinformation.

While in Experiment 2, there were some observable differences in the performance of the incentivized groups on the new statements, changes in response biases cannot account for the effect of incentives on memory performance. It is true that both the *Judgement-high-stakes-before* and the *Memory-high-stakes* group were less inclined to identify new statements as true compared to the other two groups: This could hint at a reduction in 'true' responses in the memory test. However, the *Memory-high-stakes* group was also less inclined to identify new statements as 'false' compared to all three other groups. Thus, it is unlikely that the reduced truth bias, at least in the *Memory-high-stakes* group, is merely attributable to a change in response bias.

Interestingly, contrary to the results of Experiment 1, the two incentivized groups tended to identify more false statements as 'new' compared to the control or the *Judgement-high-stakes-after* group. This pattern suggests that one of the strategies people in the high-stakes situations may opt for is to abstain from listening to false information. We discuss the implications of such a possibility in the General Discussion.

A noteworthy difference compared to Experiment 1 is that in this study incentives made participants in the *Judgement-high-stakes-before* group more severe towards the attenuated perpetrator as compared to the control group. In Experiment 1, incentives rather lead to more lenient judgements of the falsely aggravated perpetrator. Although the crucial comparisons for our hypotheses are the ones between aggravated and attenuated perpetrators within-subjects/groups, which we report, this difference is puzzling. It may have to do with the addition of more false statements in Experiment 2. These may have been more neutral than the false statements in Experiment 1, which were all particularly aggravating/attenuating. Alternatively, the different instructions for the control group in Experiment 2, which urged participants to focus either on providing accurate judgements or on providing accurate memory, may have also implicitly modified participants' judgements.

## 4 | GENERAL DISCUSSION

We speculated that a bias towards believing communicated information by default is inherent in human verbal communication but also co-evolved with vigilance mechanisms. We empirically confirmed two key predictions of such a communication model: (1) people are influenced by false information, even if it is explicitly marked as such; (2) this truth bias *can* be mitigated if high personal stakes in the communicative setting activate vigilance towards false information.

Previous studies have shown that prior knowledge or informativeness can help people disregard false information (Hasson et al., 2005; Richter et al., 2009). However, they were largely based on the seminal work by Gilbert and colleagues and have thus operated on the assumption that distraction is a necessary condition for the truth bias.

In addition, much past work relied on truth valuation of isolated statements, with no obvious relevance for the participants. The results of our two experiments confirm previous observations that distraction is neither necessary nor sufficient for the truth bias to operate (Fiedler et al., 1996; Pantazi et al., 2018). They also show that personal stakes increase vigilance in processing linguistic inputs that are otherwise likely to result in false beliefs. In a way, then, high stakes seem to improve people's ability to filter false information out under more stringent conditions (i.e., absence of distraction) than informativeness or prior knowledge.

It is worth emphasizing that in our study high stakes did not lead to an overall improved accuracy as they did not alter participants' accuracy on the true statements. Rather the interaction patterns, we observe suggest that high stakes specifically reduced participants' tendency to believe false information. Hence, our results cannot be explained by a depth-of-processing account as the latter should affect all information (regardless of truthfulness). Instead, it looks like high personal stakes in a conversational setting specifically change participants' threshold of acceptance or belief.

An open question also remains on the exact mechanisms through which people respond in the judgement and memory tasks in our paradigm. The memory and judgement truth bias indexes were positively correlated in both studies. Yet, it is equally possible that our participants made memory-based judgements or instead based their memory responses on judgements they made about the perpetrators online, while listening to the reports (see Hastie & Park, 1986).

Based on the two studies reported above, vigilance mechanisms may be conceived of as a solution to a system of communication otherwise biased towards trust or belief. While the view that people are truth-biased by default is not inherently antagonistic with the assumption of efficient mechanisms of epistemic vigilance, many studies point towards a strong tendency to believe (Asp & Tranel, 2013; Clare & Levine, 2019; Levine, 2014; Pantazi et al., 2018) and those documenting an ability to disbelieve (Hasson et al., 2005; Richter et al., 2009; Schroeder et al., 2008; Wertgen & Richter, 2018) make it look like they are. The present results make it clear that while people may prove to be 'gullible' in communicative settings where the stakes are not trivial, there are circumstances where the stakes are sufficiently high to enable resistance to false information. While Experiment 2 showed that epistemic vigilance is successful when participants are exposed to false information, the exact strategy and processing mechanisms employed by participants following the activation of epistemic vigilance are still unclear. One possibility is that high stakes may undo the default truth-biased processing of information presumed by Gilbert and colleagues (1990, 1993). In that case, while our results support the main behavioural outcome of 'gullibility' supported by this seminal work, they also challenge the strict Spinozan model where belief is an inevitable component of information comprehension.

An alternative strategy that people may have employed, nevertheless, is to simply avoid or ignore the false information altogether. Although our research did not directly test this, the memory results can provide indirect evidence of this. While in Experiment 1 the incentivized group did not appear to misidentify more false statements as

new, in Experiment 2, overall, the incentivized groups that showed evidence of successful epistemic vigilance tended to also misidentify more false statements as new. This might suggest that one strategy adopted following epistemic vigilance activation is avoidance of false information. In that case, our studies would not contradict the processing assumptions of the Spinozan model but simply reveal strategies that can be effective within an otherwise Spinozan model. This, it would be informative for future studies to take a closer look at the processing of true versus false information under high- versus low-stakes situations.

In our studies, we experimentally manipulated the stakes for the participants (i.e., the self), a choice that emanated from our theoretical framework, viewing epistemic vigilance as a costly mechanism activated when the stakes become high for an individual. An open question remains whether increased stakes for others (e.g., in our paradigm the perpetrator) might have similar effects. While no studies have directly tested this, past research documenting the operation of the truth bias despite increased accountability manipulations or among professional judges (who, presumably, have internalized the significance of their decision-making for defendants; Pantazi et al., 2020) suggest that other-oriented stakes are not as impactful self-oriented ones. Future studies should directly address this possibility.

Turning to implications for fighting misinformation, previous studies suggest that high stakes alone do not automatically improve lie detection (Levine et al., 2014). Our results indicate that high personal stakes trigger vigilance mechanisms *if* combined with explicit meta-information, signalling content falsity or inaccuracy. It is unknown whether high stakes may increase resistance to false information in the absence of additional meta-information that signals information truth value. For example, financial incentives do not effectively moderate the illusory truth effect (Speckmann & Unkelbach, 2022), which corroborates the idea that incentives alone may not automatically help shield people against misinformation. Our results suggest that high stakes are important at the level of processing *communicated* false information, granted that people possess the necessary tools to exhibit vigilance towards it. From a psychological perspective, fighting against misinformation requires to capitalize on ability as well as motivation.

Conversely, our results suggest that simply warning readers and news consumers that certain pieces of information are inaccurate or outright false is unlikely to ward off misinformation. For such warnings to be efficient, people should perceive the news consumption process as personally relevant and care more about the repercussions of consuming inaccuracies. Lastly, high stakes can effectively trigger epistemic vigilance at the time of processing but cannot undo the effects of a truth-biased processing a posteriori. Accordingly, and in line with suggestions by others (see Greifeneder et al., 2020; Lewandowsky et al., 2012) interventions against misinformation should be much more proactive than retroactive.

To be sure, epistemic vigilance could be conditional on other contextual factors than high personal stakes. For example, a motive for deception by a communicator is considered a vigilance trigger in truth-default theory (Levine, 2014) and is always an important consideration in judicial contexts. The fact that we relied on crime reports may have enhanced the salience of this factor. Future studies should

test how high stakes combine with other potential vigilance triggers as well as non-financial operationalizations of high stakes, such as social reputation or self-esteem. Previous work, however, suggests that accountability and professional expertise are not efficient vigilance triggers (Pantazi et al., 2020). Finally, our findings should be replicated with different materials, conversational settings and populations. Still, our results shed some light in the tunnel of information processing, which is so crucial to illuminate in our 'post-truth' times.

## ACKNOWLEDGEMENTS

This research was supported by the Mini-ARC 'Project' grant, *At the sources of faith*, from the Université libre de Bruxelles, by the BA/Leverhulme small research grant 'Mitigating the truth bias: a psychological approach to misinformation effects' (SRG19-190779) and by the Fonds David et Alice Van Buuren and the Jaumotte-Demoulin Foundation.

## CONFLICT OF INTEREST STATEMENT

The authors have no conflicts of interest to report.

## ETHICS STATEMENT

The studies reported here received ethics approval by the Université libre de Bruxelles and the Oxford Internet Institute.

## TRANSPARENCY STATEMENT

Data and analysis scripts for Study 1 are available on OSF ([https://osf.io/u854t/?view\\_only=b41e951c9e364cc5af1bdbd4c629067f](https://osf.io/u854t/?view_only=b41e951c9e364cc5af1bdbd4c629067f)).

Preregistration, data and analysis script for Study 2 are available on OSF ([https://osf.io/cd9hs/?view\\_only=44d11fe85d644a8a8d46f20ff9c12bea](https://osf.io/cd9hs/?view_only=44d11fe85d644a8a8d46f20ff9c12bea)).

## REFERENCES

- Arbilly, M., Motro, U., Feldman, M. W., & Lotem, A. (2011). Evolution of social learning when high expected payoffs are associated with high risk of failure. *Journal of the Royal Society Interface*, 8(64), 1604–1615. <https://doi.org/10.1098/rsif.2011.0138>
- Asp, E., & Tranel, D. (2013). False Tagging theory: Toward a unitary account of prefrontal cortex function. In D. Stuss & R. Knight (Eds.), *Principles of frontal lobe function* (pp. 383–416). Oxford University Press. <https://doi.org/10.1093/med/9780199837755.003.0029>
- Bond, C. F., & Depaulo, B. M. (2006). Accuracy of deception judgments. *Personal and Social Psychology Review*, 10(3), 214–234. <https://doi.org/10.1207/s15327957pspr1003>
- Chartrand, T., & Bargh, J. (1996). Automatic activation of impression formation and memorization goals: Nonconscious goal priming reproduces effects of explicit task instructions. *Journal of Personality and Social Psychology*, 71(3), 464–478.
- Clare, D. D., & Levine, T. R. (2019). Documenting the truth-default: The low frequency of spontaneous unprompted veracity assessments in deception detection. *Human Communication Research*, 45(3), 286–308. <https://doi.org/10.1093/hcr/hqz001>
- Clark, H. H. (1974). Semantics and comprehension. In *Current trends in linguistics* (pp. 1291–1428). Mouton. [https://doi.org/10.1016/0024-3841\(75\)90076-5](https://doi.org/10.1016/0024-3841(75)90076-5)
- Clément, F. (2010). To trust or not to trust? Children's social epistemology. *Review of Philosophy and Psychology*, 1(4), 531–549. <https://doi.org/10.1007/s13164-010-0022-3>

- DeCoster, J. (2012). Spreadsheet for converting effect size measures. Retrieved October 13, 2015, from <http://www.stat-help.com/spreadsheets/Converting%20effect%20sizes%202012-06-19.xls>
- Ekman, P., O'Sullivan, M., & Frank, M. G. (1999). A few can catch a liar. *Psychological Science*, 10(3), 263–268. <https://doi.org/10.1111/1467-9280.00147>
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G\*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175–191.
- Fiedler, K. (2012). Meta-cognitive myopia and the dilemmas of inductive-statistical inference. *Psychology of Learning and Motivation*, 57, 1–55. <https://doi.org/10.1016/B978-0-12-394293-7.00001-7>
- Fiedler, K., Armbruster, T., Nickel, S., Walther, E., & Asbeck, J. (1996). Constructive biases in social judgment: Experiments on the self-verification of question contents. *Journal of Personality and Social Psychology*, 71(5), 861–873. <https://doi.org/10.1037/0022-3514.71.5.861>
- Fiedler, K., Kaczor, K., Haarmann, S., Stegmüller, M., & Maloney, J. (2009). Impression-formation advantage in memory for faces: When eyewitnesses are interested in targets' likeability, rather than their identity. *European Journal of Social Psychology*, 39, 793–807. <https://doi.org/10.1002/ejsp>
- Forgas, J. P., & Baymeister, R. (2019). *The social psychology of gullibility: Conspiracy theories, fake news and irrational beliefs*. Routledge.
- Garcia-Marques, L., Ferreira, M. B., Nunes, L. D., Garrido, M. V., Garcia-marques, T., & Lisboa, U. D. (2010). False memories and impressions of personality. *Analysis*, 28(4), 556–568. <https://doi.org/10.1521/soco.2010.28.4.556>
- Gilbert, D. T., Krull, D. S., & Malone, P. S. (1990). Unbelieving the unbelievable: Some problems in the rejection of false information. *Journal of Personality and Social Psychology*, 59(4), 601–613.
- Gilbert, D. T., Tafarodi, R. W., & Malone, P. S. (1993). You can't not believe everything you read. *Journal of Personality and Social Psychology*, 65(2), 221–233.
- Greifeneder, R., Jaffé, M. E., Newman, E. J., & Schwarz, N. (Eds.). (2020). *The psychology of fake news: Accepting, sharing, and correcting misinformation*. Routledge.
- Grice, P. (1975). Logic and conversation. In P. Cole, & J. Morgan (Eds.), *Syntax and semantics. 3: Speech acts* (pp. 41–58). Academic Press.
- Hasson, U., Simmons, J. P., & Todorov, A. (2005). Believe it or not: On the possibility of suspending belief. *Psychological Science*, 16(7), 566–571. <https://doi.org/10.1111/j.0956-7976.2005.01576.x>
- Hastie, R., & Park, B. (1986). The relationship between memory and judgement depends on whether the judgement task is memory-based or on-line. *Psychological Review*, 93(3), 258–268.
- Henrich, J. (2009). The evolution of costly displays, cooperation and religion: Credibility enhancing displays and their implications for cultural evolution. *Evolution and Human Behavior*, 30(4), 244–260. <https://doi.org/10.1016/j.evolhumbehav.2009.03.005>
- Heyman, J., & Ariely, D. (2004). Effort for payment. *Psychological Science*, 15(11), 787–793.
- Hogarth, R. M., Gibbs, B. J., McKenzie, C. R., & Marquis, M. A. (1991). Learning from feedback: Exactness and incentives. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17(4), 734–752. <https://doi.org/10.1037/0278-7393.17.4.734>
- Hurford, J. R. (2007). The origin of noun phrases: Reference, truth and communication. *Lingua*, 117, 527–542. <https://doi.org/10.1016/j.lingua.2005.04.004>
- Kane, M. J., Conway, A. R., Miura, T. K., & Colflesh, G. J. H. (2007). Working memory, attention control, and the N-back task: A question of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(3), 615–622. <https://doi.org/10.1037/0278-7393.33.3.615>
- Khan, U., Goldsmith, K., & Dhar, R. (2020). When does altruism trump self-interest? The moderating role of affect in extrinsic incentives. *Journal of the Association for Consumer Research*, 5(1), 44–55. <https://doi.org/10.1086/706512>
- Kissine, M., & Klein, O. (2013). Models of communication, epistemic trust and epistemic vigilance. In J. Laszlo, J. Forgas, & O. Vincze (Eds.), *Social cognition and communication* (pp. 139–154). Psychology Press.
- Klein, G., Shneiderman, B., Hoffman, R. R., & Ford, K. M. (2017). Why expertise matters: A response to the challenges. *Intelligent Systems*, 32(6), 67–73. <https://doi.org/10.1109/MIS.2017.4531230>
- Lackey, J. (2007). Why we don't deserve credit for everything we know. *Synthese*, 158(3), 345–361. <https://doi.org/10.1007/s11229-006-9044-x>
- Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4, 1–12. <https://doi.org/10.3389/fpsyg.2013.00863>
- Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, 359(6380), 1094–1096. <https://doi.org/10.1126/science.aao2998>
- Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, 125(2), 255–275. <https://doi.org/10.1037/0033-2909.125.2.255>
- Levine, T. R. (2014). Truth-default theory (TDT): A theory of human deception and deception detection. *Journal of Language and Social Psychology*, 33(4), 378–392. <https://doi.org/10.1177/0261927X14535916>
- Levine, T. R., Clare, D. D., Green, T., Serota, K. B., & Park, H. S. (2014). The effects of truth-lie base rate on interactive deception detection accuracy. *Human Communication Research*, 40(3), 350–372. <https://doi.org/10.1111/hcre.12027>
- Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3), 106–131. <https://doi.org/10.1177/1529100612451018>
- Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviations around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4), 764–766. <https://doi.org/10.1016/j.jesp.2013.03.013>
- Loftus, E. F. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory*, 12, 361–366. <https://doi.org/10.1101/lm.94705>
- McNally, L., & Jackson, A. L. (2013). Cooperation creates selection for tactical deception. *Proceedings of the Royal Society B: Biological Sciences*, 280(1762), 1–7. <https://doi.org/10.1098/rspb.2013.0699>
- Millikan, R. G. (2005). *Language: A biological model*. Clarendon Press. <https://doi.org/10.1093/0199284768.001.0001>
- Muthukrishna, M., Morgan, T. J. H., & Henrich, J. (2016). The when and who of social learning and conformist transmission. *Evolution and Human Behavior*, 37(1), 10–20. <https://doi.org/10.1016/j.evolhumbehav.2015.05.004>
- Nakahashi, W., Wakano, J. Y., & Henrich, J. (2012). Adaptive social learning strategies in temporally and spatially varying environments: How temporal vs. spatial variation, number of cultural traits, and costs of learning influence the evolution of conformist-biased transmission, payoff-biased transmission, and Individual Learning. *Human Nature*, 23(4), 386–418. <https://doi.org/10.1007/s12110-012-9151-y>
- Pantazi, M., Kissine, M., & Klein, O. (2018). The power of the truth bias: false information affects memory and judgment even in the absence of distraction. *Social Cognition*, 36(2), 167–198. <https://doi.org/10.1521/soco.2018.36.2.167>
- Pantazi, M., Klein, O., & Kissine, M. (2020). Is justice blind or myopic? An examination of the effects of meta-cognitive myopia and truth bias on mock jurors and judges. *Judgment and Decision Making*, 15(2), 214–229.
- Quené, H., & van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of*

- Memory and Language*, 59(4), 413–425. <https://doi.org/10.1016/j.jml.2008.02.002>
- Reber, R., & Unkelbach, C. (2010). The epistemic status of processing fluency as source for judgments of truth. *Review of Philosophy and Psychology*, 1(4), 563–581. <https://doi.org/10.1007/s13164-010-0039-7>
- Richter, T., & Rapp, D. N. (2014). Comprehension and validation of text information: Introduction to the special issue. *Discourse Processes*, 51(1–2), 1–6. <https://doi.org/10.1080/0163853X.2013.855533>
- Richter, T., Schroeder, S., & Wöhrmann, B. (2009). You don't have to believe everything you read: Background knowledge permits fast and efficient validation of information. *Journal of Personality and Social Psychology*, 96(3), 538–558. <https://doi.org/10.1037/a0014038>
- Schroeder, S., Richter, T., & Hoever, I. (2008). Getting a picture that is both accurate and stable: Situation models and epistemic validation. *Journal of Memory and Language*, 59(3), 237–255. <https://doi.org/10.1016/j.jml.2008.05.001>
- 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>
- Speckmann, F., & Unkelbach, C. (2022). Monetary incentives do not reduce the repetition-induced truth effect. *Psychonomic Bulletin and Review*, 29(3), 1045–1052. <https://doi.org/10.3758/s13423-021-02046-0>
- Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, 25(4), 359–393. <https://doi.org/10.1111/j.1468-0017.2010.01394.x>
- ten Brinke, L., Stimson, D., & Carney, D. R. (2014). Some evidence for unconscious lie detection. *Psychological Science*, 25(5), 1098–1105. <https://doi.org/10.1177/0956797614524421>
- Wertgen, A., & Richter, T. (2018). Source credibility modulates the validation of implausible information. *Memory & Cognition*, 48(8), 1359–1375.
- Zahavi, A. (1993). The fallacy of conventional signalling. *Philosophical Transactions - Royal Society of London, B*, 340(1292), 227–230. <https://doi.org/10.1098/rstb.1993.0061>

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Pantazi, M., Klein, O., & Kissine, M. (2024). The Achilles' heel of the truth bias? High personal stakes reduce vulnerability to false information. *European Journal of Social Psychology*, 54, 1416–1429. <https://doi.org/10.1002/ejsp.3086>