



## UvA-DARE (Digital Academic Repository)

### The use-the-best heuristic facilitates deception detection

Verschuere, B.; Lin, C.-C.; Huismann, S.; Kleinberg, B.; Willemse, M.; Mei, E.C.J.; van Goor, T.; Löwy, L.H.S.; Appiah, O.K.; Meijer, E.

**DOI**

[10.1038/s41562-023-01556-2](https://doi.org/10.1038/s41562-023-01556-2)

**Publication date**

2023

**Document Version**

Final published version

**Published in**

Nature Human Behaviour

**License**

Article 25fa Dutch Copyright Act (<https://www.openaccess.nl/en/policies/open-access-in-dutch-copyright-law-taverne-amendment>)

[Link to publication](#)

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

Verschuere, B., Lin, C.-C., Huismann, S., Kleinberg, B., Willemse, M., Mei, E. C. J., van Goor, T., Löwy, L. H. S., Appiah, O. K., & Meijer, E. (2023). The use-the-best heuristic facilitates deception detection. *Nature Human Behaviour*, 7(5), 718-728. <https://doi.org/10.1038/s41562-023-01556-2>

**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>)*






# The use-the-best heuristic facilitates deception detection

Received: 17 June 2022

Accepted: 10 February 2023

Published online: 20 March 2023

 Check for updates

Bruno Verschuere <sup>1</sup>✉, Chu-Chien Lin<sup>1</sup>, Sara Huisman<sup>1</sup>, Bennett Kleinberg <sup>2,3</sup>, Marleen Willemse<sup>1</sup>, Emily Chong Jia Mei<sup>1</sup>, Thierry van Goor <sup>1</sup>, Leonie H. S. Löwy<sup>1</sup>, Obed Kwame Appiah <sup>1</sup> & Ewout Meijer <sup>4</sup>

Decades of research have shown that people are poor at detecting deception. Understandably, people struggle with integrating the many putative cues to deception into an accurate veracity judgement. Heuristics simplify difficult decisions by ignoring most of the information and relying instead only on the most diagnostic cues. Here we conducted nine studies in which people evaluated honest and deceptive handwritten statements, video transcripts, videotaped interviews or live interviews. Participants performed at the chance level when they made intuitive judgements, free to use any possible cue. But when instructed to rely only on the best available cue (detailedness), they were consistently able to discriminate lies from truths. Our findings challenge the notion that people lack the potential to detect deception. The simplicity and accuracy of the use-the-best heuristic provides a promising new avenue for deception research.

Being able to make a correct lie–truth judgement touches personal lives (for example, infidelity in relationships), legal practice (for example, detecting false allegations and deceptive denials) and society at large (for example, does a regime really possess weapons of mass destruction?). However, deception detection is notoriously difficult, with people performing barely better than the chance level. A meta-analytic estimate of 24,483 people found their average accuracy in lie–truth discrimination to be only four percentage points higher than what would be achieved by random guessing. This poor deception-detection performance is not restricted to ordinary people, but also found in professionals who routinely engage in deception detection<sup>1</sup>. But why is deception detection so challenging?

There are two prominent explanations of why people fail at deception detection. First, people rely on the wrong cues. Global surveys have shown that people's beliefs about cues to deception are strong, but wrong<sup>2</sup>. A particularly persistent stereotype is that liars avert their eye gaze, despite meta-analytic evidence showing they do not<sup>3</sup>. Second, and arguably more importantly, most cues are, at best, only weakly predictive of deception<sup>4</sup>. A meta-analysis found the median standardized

effect size (Cohen's  $d$ ) of 88 behavioural cues to be only  $d = 0.10$  (ref. <sup>5</sup>). Put differently, liars and truth tellers display 96% overlap on these behavioural variables. The diagnostic value of most cues was close to zero, and only very few cues—such as richness in detail—show actual promise as cues to deception<sup>6</sup>.

The current approach to improve deception detection is often to combine many cues. The Aberdeen Report Judgement Scales, for instance, requires three weeks of training for people to be able to use 52 cues for deception detection<sup>7</sup>. Rolled out after 9/11, the controversial US\$900 million programme called Screening Passengers by Observation Technique trained airport security personnel to screen passengers on 92 cues<sup>8</sup>. These training programmes, however, have limited success in improving the ability to detect deception<sup>9</sup>. We see two main challenges with such a 'many cues' approach. First, as the overall effect size is small, the many cues approach necessarily involves the inclusion of weak cues. Second, people will struggle with combining the many, and often conflicting, cues into a binary veracity judgement<sup>10</sup>. And statistically, combining many cues can lead to overfitting (see also the bias–variance dilemma<sup>11</sup>), with a considerable drop in

<sup>1</sup>Department of Clinical Psychology, University of Amsterdam, Amsterdam, the Netherlands. <sup>2</sup>Department of Methodology and Statistics, Tilburg University, Tilburg, the Netherlands. <sup>3</sup>Department of Security and Crime Science, University College London, London, UK. <sup>4</sup>Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, the Netherlands. ✉e-mail: [b.j.verschuere@uva.nl](mailto:b.j.verschuere@uva.nl)

**Table 1 | Overview of the aim and findings of the nine studies**

Study	Finding	Use-the-best heuristic	Control condition
Study 1	The use-the-best heuristic allows people to distinguish lie from truth	Detailedness: $d=+0.89$	Unguided: $d=+0.10$
Studies 2 and 3	Robustness of Study 1 findings in preregistered replications	Detailedness: $d=+1.11$ Verifiability: $d=+1.05$	Unguided: $d=+0.06$
Study 4	The use-the-best heuristic generalizes to novel statements	Detailedness: $d=+0.75$	Unguided: $d=+0.01$
Study 5	Knowing the goal of deception detection does not overrule the use-the-best heuristic	Detailedness: $d=+0.97$ (goal explicit) $d=+1.02$ (goal not explicit)	NA
Study 6	Use-the-best heuristic allows interviewers to make accurate on-the-spot decisions	Detailedness: $d=+1.86$	NA
Study 7	The use-the-best heuristic critically depends on cue diagnosticity	Detailedness: $d=+1.73$	Gaze aversion: $d=+0.21$
Studies 8 and 9	People are better at distinguishing lies from truths when relying on a single cue than when relying on multiple cues	Detailedness: 58.93% (Study 8) 66.41% (Study 9)	Multiple cues (including detailedness): 54.26% (Study 8) 59.14% (Study 9)

Differentiation of truthful versus deceptive statements (Studies 1–7: Cohen's  $d$ ; Studies 8 and 9: percentage accuracy) when guided to use a single cue. Control condition: unguided judgements (using any cue; Studies 1–4), a low diagnostic cue (Study 7) or multiple cues (Studies 8 and 9). NA, not applicable.

accuracy when moving to out-of-sample testing<sup>12</sup>. As a radical alternative to the 'many cues' approach, we reasoned the truth may be found in simplicity and we propose to drop rather than add cues when trying to detect deception<sup>13</sup>.

The need to integrate complex information into a binary judgement is not unique to deception detection. Medical doctors, criminal court judges, human resources consultants and stockbrokers all face a similar challenge: surgery or medication, guilty or innocent, hire or reject, buy or sell. One counterintuitive way of dealing with an information overload is to simply ignore most of the available information<sup>13,14</sup>. For example, using just two criteria—age and criminal record—allowed the prediction of the risk of criminal recidivism with the same accuracy as an algorithm that combined 137 criteria<sup>15</sup>. And a large-scale study on predicting life outcomes showed that complex computational models did not fare better than domain expert judgements based on just four variables<sup>16</sup>. Sometimes, less is more. Would the 'less-is-more' principle also apply to deception detection when lay people and experts seek to distinguish lies from the truth?

The use-the-best (and ignore-the-rest) heuristic is an instance of a one clever-cue heuristic within one-reason decision-making strategies<sup>17</sup>. It guides people to rely only on the best available cue. Here, we examine whether this simple heuristic may allow ordinary people to distinguish lies from the truth. In a series of nine studies, we asked people to evaluate honest and deceptive statements. In our control condition, people either judged statements directly on veracity free to use any cue they like (Studies 1, 2, 3 and 4), or were specifically guided to use multiple cues (Studies 8 and 9). Despite financial incentives for accurate performance, we hypothesized that their performance would be close to the chance level<sup>1</sup>. In the heuristic condition, we guided people to use only a single cue. Specifically, we used detailedness and the verifiability of such details as heuristics cues as these are the best investigated and likely the most valid cues to deception<sup>5,6,18,19</sup>. In the heuristic condition, we therefore guided people to rely on detailedness (Studies 2–9) and the presence of verifiable details (Studies 1–3) (Table 1).

## Results

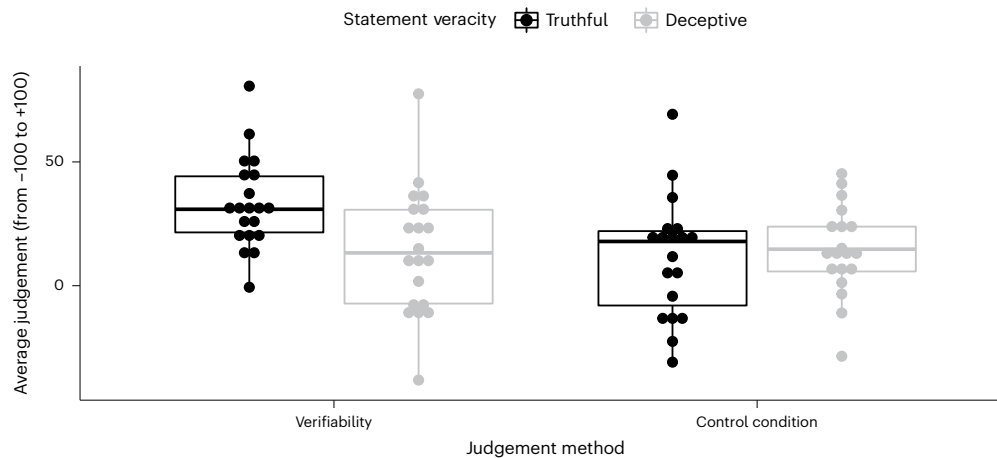
### Study 1: proof of concept

In Study 1, 39 undergraduates judged handwritten truthful and deceptive alibi statements either on deception (using any cue they like; control condition) or only on verifiability (heuristic condition). To test whether participants were better able to discriminate lies from truths in the heuristic condition than in the control condition, we conducted

the judgements to a 2 (judgement method: deception versus verifiability, between-subjects)  $\times$  2 (veracity: truthful versus deceptive, within-subjects) mixed ANOVA. Figure 1 shows the results. The predicted interaction effect between judgement method and veracity was statistically significant,  $F(1, 37) = 8.43, P = 0.006, \eta^2_p = 0.19$ .  $\eta^2_p$  is a measure of effect size expressing the proportion of variance explained by the factor after accounting for variance explained by the other factors in the model. To follow up on the interaction, we conducted a two-tailed paired sample  $t$ -test, contrasting judgements for truthful and deceptive statements within each judgement method. Lie–truth differences when judging deception were small and non-significant, two-tailed  $t(18) = 0.45, P = 0.660, d = 0.10$  (95% confidence interval (95% CI):  $-0.35$ – $0.55$ ), Bayes Factor ( $BF_{01}$ ) = 3.85 ( $\delta = -0.88$ ; 95% CI:  $-0.51$ – $0.33$ ). Cohen's  $d$  (with 95% CI) is the standardized mean lie–truth difference (within-subjects comparison), and is calculated with R for Studies 1 to 4 ( $d_z$ ; ref. <sup>20</sup>, formula 6, p. 4) and with JASP v.0.16.0.0 for Studies 5–9 (formula on <https://forum.cogsci.nl/discussion/3013/what-denominator-does-the-cohens-d-use-on-jasp>). The effect size is  $\delta$  (with 95% credible interval), obtained after updating the prior distribution with the observed data and assuming the alternative hypothesis. In contrast, lie–truth differences when judging verifiability were significant and large,  $t(19) = 4.00, P < 0.001, d = 0.89$ , (95% CI:  $0.36$ – $1.41$ ),  $BF_{10} = 45.65$  ( $\delta = 0.81$ ; 95% CI:  $0.30$ – $1.34$ ). The  $BF_{10}$  expresses how much more likely the data are under the alternative hypothesis of a lie–truth difference than under the null hypothesis of no lie–truth difference. Conversely,  $BF_{01}$  expresses how much more likely the data are under the null hypothesis of no lie–truth difference than under the alternative hypothesis of a lie–truth difference. For all Bayesian tests we relied on the default priors provided by JASP v.0.16.0.0. Default priors are recommended when prior knowledge is not specific or difficult to elicit, and one could argue that it is informed priors that would require stronger justification. For  $t$ -tests, the default prior in JASP v.0.16.0.0 is defined by a Cauchy distribution centred on a zero effect size ( $\delta$ ) and a width of 0.707. For the ANOVA the width (of 0.50) is set so that it mimics the default prior of the  $t$ -test. Given the considerable and surprising size of the obtained effect, we conducted two preregistered follow-up studies to assess the robustness of the heuristics approach to deception detection.

### Studies 2 and 3: registered replication

Studies 2 and 3 (combined  $n = 338$ ) followed the same procedure as Study 1, but (1) in a larger, crowdsourced sample, (2) with preregistration of the hypotheses and analyses and (3) an extra heuristics



**Fig. 1 | Relying on one good cue allows to tell lie from truth.** Average judgement of the truthful and deceptive statements when judging deception without guidance (control condition;  $n = 19$ ) or verifiability ( $n = 20$ ) in Study 1.

The boxplot shows the median (the midline of the box), the interquartile range (the outer borders of the box) and extreme values (that is, values larger than 1.5 times the interquartile range).

condition that judged the statements only on richness in detail. The results replicated and extended the pilot findings. The 3 (judgement method: deception versus verifiability versus detailedness)  $\times$  2 (statement veracity: truthful versus deceptive) mixed ANOVA showed the predicted interaction effect,  $F(2, 335) = 25.92, P < 0.001, \eta^2_p = 0.13$ . We followed up on the ANOVA with a  $t$ -test comparing lies and truths within each condition. The judgements of the control group did not differ between honest (average ( $M$ ) = 27.12, s.d. = 28.33) and deceptive ( $M = 25.43$ , s.d. = 26.93) statements, two-tailed  $t(107) = 0.58, P = 0.560, d = 0.06$  (95% CI:  $-0.25$ – $0.13$ ). Note that this poor deception detection ability is seen despite excluding inattentive participants, and despite financial incentives and high self-reported motivation to accurately judge the statements (Methods). Under the same conditions, but now armed with a simple heuristic, people’s judgements showed significant and large differences between the deceptive and the honest statements. For detailedness, honest ( $M = 36.74$ , s.d. = 26.05) versus deceptive ( $M = 15.27$ , s.d. = 28.47) statements, one-tailed  $t(102) = 11.29, P < 0.001, d = 1.11$  (95% CI:  $0.90$  to  $+\infty$ ), that is, honest statements were indeed judged to be higher in detailedness than lies. For verifiable details, honest ( $M = 38.70$ , s.d. = 21.84) versus deceptive ( $M = 17.00$ , s.d. = 25.58) statements, one-tailed  $t(127) = 11.89, P < 0.001, d = 1.05$  (95% CI:  $0.87$  to  $+\infty$ ). Honest accounts were judged to be more verifiable than lies. Moreover, the simple heuristics did better than a state-of-the-art, resource-intensive deception detection approach by trained coders who coded the statements word-for-word on (verifiable) details<sup>19</sup>.

We translated the heuristics-based judgements into a classification performance by averaging the judgement for each of the 64 statements, for each judgement method. For each judgement method, we used the average statement judgement to predict statement veracity. The receiver operating characteristic (ROC) analysis plots sensitivity against specificity and provides a measure of diagnostic value across all possible cut-off points. The area under the ROC curve varies from 0 to 1 (=perfect classification), with 0.50 denoting the chance level. As shown in Table 2, the ROC a was above chance for the heuristics condition guiding people to rely on a single cue ( $0.71 \leq \text{ROC } a \leq 0.75$ ; with the lower bound of the CIs exceeding 0.50), but at chance level for the control condition that allowed considering any possible cues ( $0.51 \leq \text{ROC } a \leq 0.61$ ; with the 95% CI including 0.50). Using Youden’s  $J^{21}$ , we also identify the optimal cut-off point when equally balancing specificity and sensitivity. We used independent validation, the strictest method to avoid data overfitting. Hence, we used the data of Study 2 to evaluate classification accuracy based on the optimal cut-off derived in Study 3 (and vice versa). Accuracy was above chance for the heuristic

**Table 2 | Accuracy in classifying lies from truths for unguided judgements using any possible cue and for single-cue judgements (verifiability, detailedness) for Studies 2 and 3**

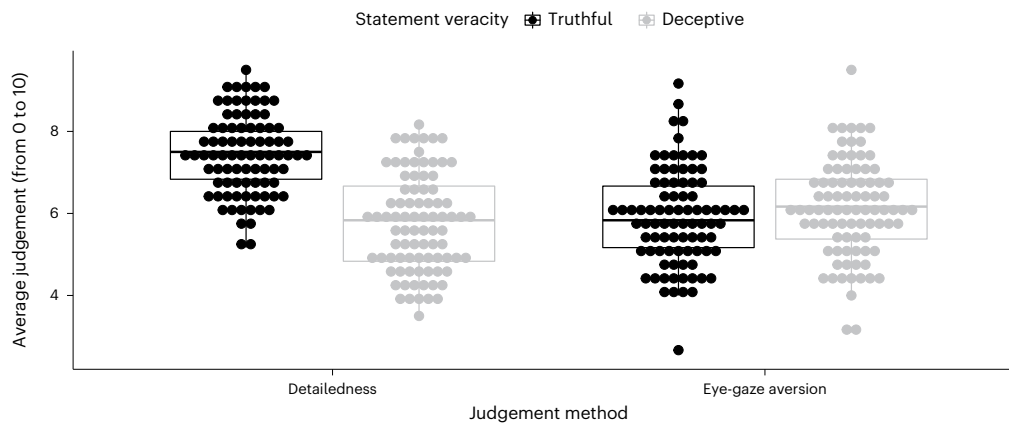
	Area under the curve (with 95% CI)		Accuracy	
	Study 2	Study 3	Cut-off based on Study 2 data, applied to Study 3 data	Cut-off based on Study 3 data, applied to Study 2 data
Unguided (any cue)	0.61 (0.47–0.76)	0.53 (0.44–0.62)	50%	52%
Single cue: verifiability	0.72 (0.61–0.83)	0.75 (0.68–0.82)	70%	69%
Single cue: detailedness	0.72 (0.60–0.83)	0.71 (0.62–0.80)	65%	67%

approach (65–70%) and better than the control condition that allowed judges to incorporate any possible cue (50–52%) (Table 2).

Moving from undergraduates (Study 1) to crowdsourced participants, Studies 2 and 3 indicate that our findings are not restricted to a specific sample. But all three studies relied on the same set of statements written in Dutch. Study 4 sought to assure that our findings generalize beyond the statements used in Studies 1–3.

**Study 4: generalization**

Study 4 participants ( $n = 192$ ) judged interview transcripts in German. The 2 (judgement method: deception versus richness in detail)  $\times$  2 (veracity: truthful versus deceptive) mixed ANOVA again showed the predicted interaction effect,  $F(1, 190) = 20.09, P < 0.001, \eta^2_p = 0.096$ . Lie-truth differences when judging deception were small and non-significant, two-tailed  $t(103) = 0.09, P = 0.929, d = 0.01$  (95% CI:  $-0.18$ – $0.20$ ),  $BF_{01} = 9.17$  ( $\delta = 0.01$ ; 95% CI:  $-0.81$ – $0.20$ ). In contrast, lie-truth differences when judging richness in detail were significant and moderate to large, two-tailed  $t(87) = 7.06, P < 0.001, d = 0.75$  (95% CI:  $0.51$ – $0.99$ ),  $BF_{10} = 2.53 \times 10^7$  ( $\delta = 0.73$ ; 95% CI:  $0.50$ – $0.97$ ). The heuristic judgements took, on average, only a minute per statement. This opens the possibility to apply them in real-life situations (for example, security questioning) where the limited time often prohibits more extensive credibility assessment methods. But before considering real-life applications, we need to ensure they are resistant to strong stereotypes about deception.



**Fig. 2 | The use-the-best heuristic critically depends on cue diagnosticity.** Average judgement of the truthful and deceptive statements when judging a high diagnostic (detailedness;  $n = 85$ ) or a low diagnostic cue (eye-gaze aversion;  $n = 86$ ) in Study 7. The boxplot shows the median (the midline of the box), the interquartile range (the outer borders of the box) and extreme values (that is, values larger than 1.5 times the interquartile range).

$n = 86$ ) in Study 7. The boxplot shows the median (the midline of the box), the interquartile range (the outer borders of the box) and extreme values (that is, values larger than 1.5 times the interquartile range).

### Study 5: making the goal of lie detection explicit

Thus far, participants in the heuristics conditions had not been informed that their judgements served to distinguish lie from truth. We had been concerned that merely knowing the goal of deception detection may have been enough to activate stereotypes about deceptive behaviour (5) thereby overruling the use of the diagnostic cues, and hence diminishing the effectiveness of the heuristics approach. To examine this possibility, Study 5 participants ( $n = 150$ ) had been randomly allocated to the non-explicit condition (mimicking the heuristics condition in Studies 1–4) or to an explicit condition. Only participants in the explicit condition were informed that some statements were deceptive and that their goal was to detect deception.

Using JASP v.0.16.0.0 and its default settings, the 2 (goal of lie detection: explicit versus non-explicit)  $\times$  2 (veracity: truthful versus deceptive) mixed Bayesian ANOVA showed the data were 2.78 times less likely ( $BF_{01}$ ) under the model including the interaction than under the model with only the two main effects. Lie–truth differences when judging richness in detail were significant and large when the goal of lie detection was not explicit (as it was in the when participants relied on heuristics in Studies 1–4), Cohen's  $d = 1.02$  (95% CI: 0.78 to  $\infty$ ) obtained in one-tailed  $t$ -test,  $BF_{10} = 2.78 \times 10^{10}$  ( $M_{\text{truthful}} = 40.56$ , s.d. = 23.40;  $M_{\text{deceptive}} = 22.69$ , s.d. = 25.91;  $\delta = 0.99$ ; 95% CI: 0.71–1.27), but also when the goal of lie detection was made explicit, Cohen's  $d = 0.97$  (95% CI: 0.74 to  $\infty$ ) obtained in one-tailed  $t$ -test,  $BF_{10} = 1.10 \times 10^9$  ( $M_{\text{truthful}} = 39.56$ , s.d. = 24.61;  $M_{\text{deceptive}} = 17.33$ , s.d. = 28.19;  $\delta = 0.95$ ; 95% CI: 0.68–1.22). We conclude that the heuristics approach for deception detection is not easily susceptible to stereotypes about deception. Although our heuristic specified what criterion to rely on, it did not instruct the user how to make a decision about the veracity of the statement. In Study 6, we added an explicit decision rule and explored its accuracy under the most challenging conditions—interviewers making decisions on the spot.

### Study 6: towards real-life application

In Study 6, we explored the applied potential of heuristics for deception detection. In total, 21 deceptive and 23 honest participants were interviewed by four interviewers, who relied on the heuristics approach to make real-time lie–truth decisions. Interviewers judged the statements on detailedness, now on a more user-friendly scale (from 0 = not detailed at all, to 10 = very detailed). The participants that were interviewed received a reward if their statement was deemed credible. The simple decision rule was: 'Consider the statement truthful for detailedness scores of six and more'. In total, 91% of the truthful statements and 67% of the deceptive statements were correctly classified,

with an overall accuracy of 79%. The simple heuristic turned out to be surprisingly accurate.

### Study 7: cue diagnosticity matters

Relying on a single cue avoids cognitive overload. But that does not mean any single cue can be validly used in the heuristic approach. To test whether cue diagnosticity matters<sup>10</sup>, participants ( $n = 171$ ) evaluated truthful and deceptive video statements either on a high diagnostic cue (detailedness) or on a low diagnostic cue (the amount of eye gaze aversion), using a scale from 0 to 10. We found strong support for the idea that the cue diagnosticity determines the success of the heuristic approach. Using JASP v.0.16.0.0 and its default settings, the 2 (cue: richness in detail versus eye gaze aversion)  $\times$  2 (veracity: truthful versus deceptive) mixed Bayesian ANOVA showed that the data were much more likely ( $BF_{10} = 5.73 \times 10^{23}$ ) under the model that included the interaction as compared to the model that only included the two main effects. As is clear from inspecting Fig. 2, cue diagnosticity matters. Lie–truth differences when judging eye-gaze aversion were faint, two-tailed  $t(85) = 1.98$ ,  $P = 0.05$ ,  $d = 0.21$  (95% CI:  $-0.01$ – $0.47$ ),  $BF_{01} = 1.29$  ( $\delta = -0.21$ ; 95% CI:  $-0.42$ – $0.00$ ). In contrast, lie–truth differences when judging richness in detail were significant and large, one-tailed  $t(84) = 15.94$ ,  $P < 0.001$ ,  $d = 1.73$ , (95% CI:  $1.44$  to  $+\infty$ ),  $BF_{10} = 1.79 \times 10^{24}$  ( $\delta = 1.70$ ; 95% CI:  $1.37$ – $2.04$ ). Study 7 hereby also shows that the success of the heuristics approach cannot be attributed to intuitive decision-making but instead hinges on using the best available cue.

### Studies 8 and 9: explicit use of multiple cues

So far, in our control condition people made a deception judgement, leaving them free to use any cue they like. Under such conditions, people report to<sup>2</sup> and actually make use of<sup>6</sup> multiple cues, but we did not explicitly guide them to use multiple cues. In Study 8, we explicitly guided participants ( $n = 146$ ) to rely on either a single cue (detailedness) or on multiple cues (detailedness, affect, unexpected complications and admitting lack of memory) before making a binary lie–truth judgement. A one-tailed independent sample  $t$ -test confirmed that participants' accuracy (that is, the percentage of correctly identified lies and truths) was higher when relying on a single cue (58.93%) than when relying on multiple cues (54.26%),  $t(97) = 2.013$ ,  $P = 0.023$ ,  $d = 0.408$  (95% CI:  $0.07$  to  $+\infty$ ),  $BF_{10} = 2.45$  ( $\delta = 0.37$ ; 95% CI:  $0.05$ – $0.76$ ). Note that this provides for a stringent test of our hypothesis, because both conditions judged detailedness with the sole difference between conditions being that the multiple cue condition coded three additional cues. Because the data provided only anecdotal support for our hypothesis, we repeated the study with some improvements, including

a preregistered stopping rule specifying we would halt data collection when reaching decisive evidence for either the null or the alternative hypothesis. A one-tailed independent sample *t*-test confirmed that participants in Study 9 ( $n = 405$ ) achieved higher accuracy in telling lies from truths when relying on a single cue (66.41%) than when relying on multiple cues (59.14%),  $t(380) = 4.71, P < 0.001, d = 0.482$  (95% CI: 0.311 to  $+\infty$ ),  $BF_{10} = 7,951$  ( $\delta = 0.47$ ; 95% CI: 0.27–0.67).

## Discussion

Although detecting deception is incredibly important, it is also incredibly difficult. We propose a radical alternative to the trend towards ‘many cues’ solutions to increase deception detection accuracy<sup>22,23</sup>. We guided people to only judge the level of detail in the message, and consistently observed it allowed to discriminate lies from truths.

To put the accuracy rates found in the current set of studies, 59–79%, into context, we can compare them to other approaches developed to improve deception detection. First, the cognitive approach advocates active interviewing to increase lie–truth differences by imposing cognitive load, asking unanticipated questions and encouragements to say more<sup>24</sup>. A recent meta-analysis estimated the accuracy of the cognitive approach to be 60%, and, when corrected for publication bias, 55%. Second, approaches advocated for and applied in the field—including the Statement Validity Analysis—show accuracy rates in the same range as those observed in our current studies<sup>25</sup>. As with any tool considered for application in real (legal) contexts, the error margin requires caution. But above all, two central findings persisted through a series of nine experiments: first, compared to other approaches, the heuristic approach is a success in its accuracy and efficiency. Second, this paper revives a deception detection approach that has been thought of as a dead-end: lay decision-making.

Our data show that relying on one good cue can be more beneficial than using many cues. Admittedly, our use-the-best approach is not necessarily restricted to a single cue and could be expanded with other cues. This would, however, require (1) robust evidence for cue validity and (2) clear guidance on how to combine the cues, both of which are lacking from the evidence base to date. Potentially, adding more cues could invalidate the heuristic approach. This risk can be illustrated by a study where people rated statements on 11 cues before making a final lie–truth judgement<sup>26</sup>. Although users correctly scored truthful statements to be richer in detail than deceptive statements, their final veracity judgements were not above the chance level<sup>26</sup>. Similarly, although participants in the multiple cue condition of Studies 8 and 9 judged statements based on several valid cues, including detailedness, their lie–truth judgements were worse than participants relying only on detailedness.

We would like to emphasize three potential avenues for future research. First, an important limitation of the current series of studies is that, to know the ground truth, we instructed our participants to lie. Our findings should be extended to more realistic settings, including self-chosen and/or high-stake lies. Raising the stakes will also increase the odds that liars attempt to alter their message to increase its credibility, for example, by enriching their lies with details. Second, we demonstrated the success of the use-the-best heuristic only in the particular context where statements are about episodic memory and truth tellers were both willing and able to provide specific details. Cues other than detailedness may be more valid in different contexts, and a context-contingent approach<sup>27</sup> may provide guidance on which cues will be most efficient in other contexts such as the detection of fake news. Finally, future research could compare the results of the use-the-best heuristic to the performance of artificial intelligence (AI). The use of AI has also attracted increased attention in deception detection<sup>28</sup>. We would not be surprised if, especially in situations where one cue is dominant, such as deception detection, simple heuristics would outperform complex AI techniques. Moreover, for deep learning—currently the most promising AI approach—it is no longer explainable how

the lie–truth classification came about. This lack of transparency may not be acceptable when making real-life decisions, making human judgements indispensable<sup>29</sup>.

People may not necessarily be poor lie detectors. When judging rich statements about a past event, detailedness provides an easily assessed indicator of truth. The next step is to see whether our findings can be translated to other domains. For now, our findings suggest a simple solution to a complex problem: a rule of thumb may help to find the truth.

## Methods

Our research complies with the guidelines formulated by the Ethics Review Board of the Faculty of Social and Behavioural Sciences, University of Amsterdam. Informed consent was obtained from all participants.

### Study 1 (pilot)

Participants evaluated statements of undergraduates who—honestly or dishonestly—described their recent campus activities. Participants were randomly allocated to one of two conditions. In the control condition, participants judged these statements on veracity (from –100, totally deceitful, to +100, totally truthful). In the heuristics condition, participants judged the statements on verifiability (from –100, totally unverifiable, to +100, totally verifiable). These participants had been explained that (based on ref. 4), and asked them to use this definition when evaluating the statements on verifiability. Ethics approval and materials can be found on <https://osf.io/z26ar/>. No statistical methods were used to predetermine sample sizes: we aimed for  $n \geq 40$  within the available time. This was a single-blinded study (participants were not aware of the different judgement methods), with data analysis not performed blind to the conditions of the experiments.

**Participants.** In total, 51 undergraduates of the University of Amsterdam Psychology Department took part in Study 1. We excluded seven participants who failed the attention check (see below) and five participants who were not Dutch native speakers. Of the 39 remaining participants (31 female, 8 male;  $Age = 19.38$  years,  $s.d. = 1.68$ ),  $n = 19$  judged deception (any cue possible) and  $n = 20$  judged only the single-cue verifiability. Participants received course credits for partaking in the study and the most accurate participant received a 20 euro bonus.

**Procedure.** After providing online informed consent, participants were randomly assigned to the deception judgement or the verifiability judgement condition through Qualtrics. In both conditions, the participants were asked to evaluate 16 alibi statements, presented one by one, in a random order. In the verifiability judgement condition there was no mentioning of deception or lie detection.

Participants in the control condition were asked to evaluate ‘How truthful is this statement?’ on a scale from ‘totally deceitful’ (–100) to ‘totally truthful’ (+100), with the definition of truthfulness provided as ‘a truthful statement is a statement that is true, honest and adheres to the fact of the situation’. Judging verifiability, participants were asked to evaluate ‘How verifiable is this statement?’ on a scale from ‘totally unverifiable’ (–100) to ‘totally verifiable’ (+100), with the definition that ‘verifiable activities are activities that are recorded (for example, a security camera), documented (for example, payment with a debit card or using a smartphone) or an activity with an identifiable witness present’. This definition arose from the verifiability approach (10), but we simplified it to its essence.

An attention check was embedded among the statements. It looked like another alibi statement, but instructed participants to ignore the provided statement and instead answer –47 on the scale. After rating the (real and bogus) statements, a manipulation check asked participants about the basis for the judgements (indicate up to three out of the 11 cues from the list they used most as the basis of

their judgement; with three cues referring to verifiability), a single item asked about motivation to accurately judge the statements (from -100 to +100) and finally participants were asked to provide age, gender (options: male, female, non-binary) and mother tongue.

**Materials.** We used 64 alibi statements (32 truthful, 32 deceptive). To avoid item effects, we created four sets of 16 statements (each containing eight truthful and eight deceptive statements) and participants were randomly assigned to receive one of the sets. The statements were selected from 72 statements obtained in a previous mock crime study, where participants provided a handwritten statement on their whereabouts on campus in the last 15 minutes<sup>30</sup>. Participants either truthfully described their activities, or they lied. The lying participants had just enacted the mock theft of an exam, but pretended to have been on campus as a regular student. These statements were manually pseudonymized (that is, all identifiable information, including names of persons, were changed to plausible alternatives). Content coding by trained coders (11) showed that the 32 truthful statements ( $M = 8.28$ ;  $s.d. = 8.67$ ) contained more verifiable details than the 32 deceptive statements ( $M = 3.47$ ;  $s.d. = 4.65$ ),  $d = 0.69$  (95% CI: 0.18–1.19) (Supplementary Table 1: <https://osf.io/v3kdw/>). Below is the English translation of one example statement (all original (pseudonymized) Dutch statements can be found on <https://osf.io/z26ar/>):

'I quietly walked down until the entrance of G/lab. I was in doubt about what to do (stood still for a moment). Then I walked into the corridor of G, saw a cleaner/guy with a cart and read something about using lockers at the UvA. Then I walked to the outside entrance and walked around (back of G) and looked at the kind of butterflies that are now there for the light festival. So then I walked further around G. Went back inside (second floor lab) and looked for a moment at university pabo, there is a poster next to the door about participating in brain research for money. When I had read that, I walked quietly to this research room.'

### Studies 2 and 3

The predicted interaction did not reach significance in Study 2, but we may have used an underpowered design to uncover the interaction effect. Of note, the lie–truth difference in the Study 2 control condition happened to be larger than anticipated ( $d = 0.39$ ; Supplementary Table 1). We think this is due to sampling error related to the modest sample size<sup>31</sup>. We thus ran the study again with more statistical power (=Study 3). Because of the near-identical design, we merged the Study 2 and 3 data.

The procedure of Studies 2 and 3 followed that of Study 1 with three main differences. First, we preregistered the hypotheses and statistical analyses before the start of the study on 12 March 2021 (Study 2) and 26 March 2021 (Study 3): <https://osf.io/z26ar/>. There are two deviations from the preregistration: we tested  $n = \pm 142$  rather than  $n = \pm 155$  in Study 2 (due to 10 pilot participants being erroneously counted in Prolific) and we merged Study 2 and 3 data. Second, we moved from locally recruited undergraduates to online crowdsourcing. Third, we added a heuristic condition that based their judgements on detailedness, using the following definition: 'Degree to which the message includes details such as descriptions of people, places, actions, objects, events and the timing of events; the degree to which the message seemed complete, concrete, striking or rich in details'<sup>5</sup>. There were a few other, minor changes to the Study 1 procedure (Supplementary Information). Ethics approval and materials can be found at <https://osf.io/z26ar/>. This was a single-blinded study (participants were not aware of the different judgement methods), with data analysis not performed blind to the conditions of the experiments. The sample size for Study 2 was aimed to have sufficient powered for the follow-up  $t$ -tests. To obtain 90% power for a one-tailed paired sample  $t$ -test ( $\alpha = 0.05$ ) for the lie–truth effect observed in our first study ( $d = 0.76$ ) a sample size of  $n = 17$  (in each of the three conditions) was needed. Anticipating that the effect may be smaller than that observed in Study 1 and anticipating

some exclusions, we decided to test  $n = 50$  in each condition, hence we planned for  $n = 150$  in total. Due to simultaneous starting times in prolific we anticipated ending up with slight more participants that start the study ( $n = \pm 155$ ). As Study 2 turned out to be insufficiently powered to pick up the interaction, we ran it again, with twice the sample size (Study 3, planned  $n = \pm 300$ ).

**Participants.** Study 2. In total, 142 participants took part in Study 2. We excluded 34 participants who failed either of the two attention checks. Of the 108 remaining participants (39 female, 69 male;  $M$  age = 29.69 years,  $s.d. = 10.34$ ;  $n = 30$  judging deception,  $n = 39$  judging verifiability and  $n = 39$  judging detailedness), most were Dutch (73%; Belgian: 24%; other: 3%).

**Study 3.** Participants from Study 2 could not partake in Study 3. In total, 303 participants took part in Study 3. We excluded 73 participants who failed either of the two attention checks. Of the 230 remaining participants (107 female, 119 male, four missing;  $M$  age = 30.32 years,  $s.d. = 10.59$ ;  $n = 77$  judging deception,  $n = 89$  judging verifiability and  $n = 64$  judging detailedness), 72% were Dutch (Belgian: 26%; other: 2%).

### Study 4

Study 4 sought to examine whether our findings generalize to other languages and moes of statement production. Hereto, fluent-German crowdsourced participants judged interview transcripts (Materials) either on deception (control condition) or on richness in detail. Ethics approval, data and materials of Study 4 can be found at <https://osf.io/z26ar/>. Hypotheses, analysis plan and predictions were preregistered on 5 July 2021, before the start of data collection and can be found at <https://osf.io/z26ar/>. There were no deviations from the preregistration. This was a single-blinded study (participants were not aware of the different judgement methods), with data analysis not performed blind to the conditions of the experiments. The sample size for Study 4 was determined considering three design aspects. First, we wished to obtain reliable estimates for each judgement method, with simulation research suggesting that it required 1,000+ judgements in each cell of our study design (that is, for each of two judgements methods a minimum of 84 participants judging 12 statements each). Second, to have 95% power to pick up the effect of interest (the interaction between statement veracity and judgement method,  $\eta^2_p = 0.03$  in Study 1), G-POWER showed that a minimum of  $n = 108$  at a significance threshold of 0.05 was needed. Third, we expected up to 24% exclusions. We decided to test  $n = \pm 250$ .

**Participants.** In total, 251 participants took part in Study 4. We excluded 59 participants who failed both attention checks. The 192 remaining participants (92 females;  $M$  age = 27.59 years,  $s.d. = 9.11$ ;  $n = 104$  judging deception,  $n = 88$  judging detailedness) were Polish (26%), German (13%) or were one of 27 other nationalities. Demographics including sex (options: male, female) were obtained from Prolific.

**Procedure.** After providing informed consent, participants were randomly assigned to the deception judgement or richness in detail judgement condition through Qualtrics. In both conditions, participants were asked to evaluate 13 transcripts (including one bogus transcript used as an attention check, but excluded from main analyses), presented one by one, in a random order. The instructions for the judgements were the same as for Studies 1–3.

The first attention check concerned the bogus transcript, which looked like just another transcript but with the instruction to ignore the transcript and instead answer -47 on the scale. The second attention check was a surprise recall test after the last transcript, asking to select a unique utterance (for example, 'forgot the name of the girl I was looking for') in the last transcript among six options. Thereafter, participants indicated their motivation and experienced difficulty and were asked to list, one by one, the cues they had relied on.

**Materials.** We used 72 transcripts (half truthful, half deceptive). To avoid item effects, we created six sets of 12 transcripts, and participants were randomly assigned to receive one of the six sets. Due to a programming error one of the six sets missed one (truthful) statement. The statements were selected from ref. <sup>32</sup>. Participants in that study were native-German speaking undergraduates who were interviewed about the two tasks they claimed to have been doing in the past half an hour. Statements were later transcribed verbatim. We selected the transcripts from participants who had been instructed to consistently lie or tell the truth (that is, the lie–lie and truth–truth conditions), and used only the ‘find Michelle at the bus stop’ task. This task entailed leaving the laboratory, crossing the campus to the bus stop, trying to find a girl named Michelle (of whom they received a photo), making notes of arriving and leaving buses, then returning to the laboratory within 35 minutes. From the structured interview, we selected only the first response to the interviewer’s instruction to describe the task as accurately and in as much detail as possible. Truth tellers described the task they had enacted (trying to find Michelle at the bus stop). Liars also provided a statement about their search to find Michelle, but had not actually enacted that task. We edited the transcripts to correct for spelling errors, but we retained all utterances and filler words (for example, ‘Ehm’).

Trained coders counted the number of perceptual, temporal and spatial details, and we summed these to provide an index of richness in detail (13). Coding was based on the entire interview. This showed that the 36 truthful transcripts ( $M = 38.06$ ;  $s.d. = 15.36$ ) contained more details than the 36 deceptive transcripts ( $M = 25.72$ ;  $s.d. = 9.66$ ),  $d = 0.96$  (95% CI: 0.47–1.45). Below is the English translation of one example statement (all original German transcripts can be found at <https://osf.io/z26ar/>): ‘Okay, so after I finished the task with the café, I went to the stop at the hospital. I didn’t know exactly where it was, so I first meandered through here a bit, asked “uuh I’m doing a task, can you tell me where the stop is?” And I already thought that it was this one and then I went there. Yes, and then I was supposed to look for Michelle. There were two or three people sitting there, three people sitting there, and then I asked them in Dutch if their name was Michelle. Yes, there was no Michelle there, then I sat there for 5 minutes, looked to see if maybe some bus was coming by where a Michelle got off, but no bus came by at all. And then I came back here and, yes, I didn’t complete the task because I didn’t find Michelle.’

### Study 5

Participants judged statements on richness in detail, either being explicitly told or not that their judgements served to tell lie from truth. Participants in the explicit condition were told that some statements were deceptive and that their goal was to detect the deceptive statements. Participants in the non-explicit condition were not given any information about deception or lie detection and merely asked to evaluate the statements. Ethics approval and materials of Study 5 can be found at <https://osf.io/z26ar/>. Hypotheses, analysis plan and predictions were preregistered before the start of data collection on 25 October 2021 and can be found at <https://osf.io/z26ar/>. There were no deviations from the preregistration. This was a single-blinded study (participants were not aware of the different conditions), with data analysis not performed blind to the conditions of the experiments. The sample size for Study 5 was set to assure 1,000+ judgements for each of two judgements methods (thus a minimum of 63 participants each judging 16 statements) and anticipating up to 23% exclusions. We decided to test  $n = \pm 164$ .

**Participants.** In total, 166 fluent Dutch-speaking participants (who had not performed in Studies 2–4) took part in Study 5 on Prolific. We excluded 16 participants who failed both attention checks. The 150 remaining participants (83 females;  $M_{age} = 26.80$  years,  $s.d. = 8.48$ ;  $n = 76$  in the explicit condition and  $n = 74$  in the non-explicit condition)

were Dutch (55.33%), Belgian (31.33%) or another nationality (13.33%). About half of them (52%) were students. Demographics including sex (options: male, female) were obtained from Prolific.

**Procedure.** After providing informed consent, participants were randomly assigned to the explicit versus non-explicit condition through Qualtrics. In both conditions, participants were asked to evaluate 16 statements (and an additional bogus statement used as an attention check, but excluded from main analyses), presented one by one, in a random order. Instructions for the non-explicit condition were similar to those used in Studies 1–4 (judge detailedness), but in the explicit condition participants were informed (1) that some statements were deceptive and (2) that their goal was to detect those lies.

The first attention check concerned the bogus statement, which looked like just another transcript, but with the instruction to ignore the transcript and instead answer –47 on the scale. The second attention check was a surprise multiple-choice question after judging the last statement, asking to indicate the core of the last statement (for example, ‘Search for a book’) from six options. Thereafter, participants rated motivation and difficulty. Finally, there were two (open box) manipulation checks, asking about the goal of the study and what cues they had relied on.

**Materials.** The statements were selected from ref. <sup>30</sup> and are the same as those used in Studies 1–4. We used 64 statements (half truthful, half deceptive). To avoid item effects, we created four sets of 16 statements and participants were randomly assigned to receive one of the four sets.

### Study 6

Undergraduate participants either lied or told the truth in a videotaped interview about their whereabouts at the university campus. Immediately after the interview, the interviewers judged the statement on detailedness (using a 0 to 10 scale), with the statement deemed credible for scores of six and above. We examined the accuracy of these simple, real-time judgements. Study 6 was exploratory and therefore not pre-registered. Ethics approval and materials of Study 6 can be found at <https://osf.io/z26ar/>. No statistical methods was used to predetermine sample sizes: we aimed to interview  $n \geq 50$  within the available time. Interviewers, but not participants, were blind to the condition. Data analysis was not performed blind to the conditions of the experiments.

**Participants.** In total, 47 undergraduate participants from the University of Amsterdam took part in return for course credits. Three participants were excluded, two because they did not complete their mission and one because of suspected intoxication. Of the remaining 44 participants,  $n = 23$  were in the truthful condition ( $M_{age} = 19.87$ ,  $s.d. = 2.70$ ; 43.5% native English speakers; 78% female, 22% male) and  $n = 21$  were in the deceptive condition ( $M_{age} = 19.48$ ,  $s.d. = 1.12$ ; 47.6% native English speakers; 52% female, 43% male, 5% non-binary).

**Procedure.** We recruited participants who were comfortable to provide a video statement. Through a brief screening via email, we tried to balance our sample with about half native English and half non-native English speakers. The entire procedure was conducted in English. Upon arrival to the laboratory, participants were welcomed by a first experimenter. There were four experimenters (undergraduate students) who took the role as Experimenter 2 and each interviewed three to 14 participants. Participants provided written informed consent.

Participants were randomly assigned to the deceptive versus truthful condition and received written instructions for the theft or study location mission, respectively. Participants were asked to paraphrase their mission to the first experimenter to assure it was well understood. In the deceptive condition, participants first went to a building to find a key, then to a another building to open up a mail box with that key and steal an exam and finally to a third building to drop the stolen exam.

In the truthful condition, participants searched for an appropriate study location in several buildings of the campus, taking flyers with them to proof they visited the designated areas. Participants were asked to return in 25–30 minutes.

Upon return to the laboratory, participants were informed that they were suspected of the theft and were briefly informed about the innocent mission (allowing those in the deceptive condition to create a realistic lie). They were informed that they would be interviewed about their whereabouts in the last half an hour by a second experimenter and that their statement would be checked for verifiability (see the information protocol in ref.<sup>19</sup>). A reward in course credits was promised for providing a credible statement and they were given 10 minutes to prepare it.

The participants were then guided to another room, where the second (condition-blind) experimenter conducted the video interview. After a brief explanation and some brief small talk (to build rapport), the interviewer asked the participant to describe their whereabouts of the last half an hour in as much detail as possible. To try and get a rich statement, the interviewers encouraged them to talk for 10 minutes. The experimenter had been instructed not to interrupt the interviewee and to only encourage them to speak (by nodding, 'OK', etc.). When such prompts did not lead the interviewee to say more, a follow-up question was asked (that is, 'What proves to me that you are telling the truth?'). Directly after the interview, the same experimenter scored the interview on detailedness from 0 = not detailed at all, to 10 = very detailed using the DePaulo et al.<sup>5</sup> definition ('Degree to which the message includes details such as descriptions of people, places, actions, objects, events and the timing of events; the degree to which the message seemed complete, concrete, striking or rich in details').

After the interview, participants were guided back to the first experimenter, asked to take an English-language proficiency test, to provide their demographics (including age, native tongue and gender with options male, female, non-binary/third gender, prefer not to say) and to answer a few brief questions about the interview experience (single-scale measures of cognitive demand, emotional arousal, motivation, fatigue and perceived likelihood that statement would be verified; all from -100, not at all, to +100). Participants were thanked and received their credits (with a detailedness score of  $\geq 6$  by the second experimenter leading to the bonus pay).

## Study 7

To show that the heuristic approach critically depends on cue diagnosticity, participants judged truthful and deceptive videotaped interviews either on a high diagnostic cue (richness in detail) or on a low diagnostic cue (eye-gaze aversion). Ethics approval and materials of Study 7 can be found at <https://osf.io/z26ar/>. Hypotheses, analysis plan and predictions were preregistered before the start of data collection on 15 April 2022 and can be found at <https://osf.io/z26ar/>. There were no deviations from the preregistration. This was a single-blind study (participants were not aware of the different cues being judged in each condition). Data analysis was not performed blind to the conditions of the experiments. We planned for  $n = \pm 200$ . This sample size was determined considering (1) having 1,000+ judgements in each condition (that is, a minimum of 83 participants for each of the two conditions, each judging 12 statements; hence a minimum  $n$  of 166), (2) a MORE POWER 6.0 calculation showing that to have 90% power to detect the within-between interaction of  $\eta_p^2 = 0.13$  (based on the interaction we obtained in Studies 2 and 3; detail versus control) in the ANOVA with significance testing requires a minimum total  $n$  of 72 (but note that our confirmatory analyses are Bayesian) and (3) taking into account an expected  $\pm 23\%$  exclusion rate.

**Participants.** In total, 205 participants took part in Study 7. We excluded 34 participants who failed both attention checks. The 171 remaining participants (123 female, 44 male, four other) had a mean age of 22.07

years ( $s.d. = 5.03$ ). In total, 86 participants judged detailedness and 85 judged eye-gaze aversion. They spoke Dutch (31%), English (17.5%) or another (51.5%) language as a mother tongue. Their countries of origin were the Netherlands (30%), Germany (8%) or one of 40 others. Participants were rewarded with course credits or 7.50 euros and the three best-performing participants received a bonus of a 0.50 credit or 5 euros.

**Procedure.** Participants were recruited via the recruitment portal of the University of Amsterdam. The vast majority of this pool consists of undergraduate students, with the remainder consisting of community members. Participants first provided informed consent, which included the explicit agreement not to download, store or share the video statements. Participants were randomly assigned to judge detailedness or eye-gaze behaviour through Qualtrics. In both conditions, participants were asked to evaluate 12 videos, presented one by one, in a random order. Instructions for the detailedness judgements were the same as for Studies 1–6. For eye-gaze aversion, we instructed people to judge 'Looking away', explaining that this 'means the person in the video does not maintain eye contact with the interviewer/camera or looks to the side during the interview'.

One attention check asked about the demeanour of the interviewee (that is, 'Please answer the following question on the content of the last statement. Which is true? The interviewee scratched his hair several times, the interviewee coughed several times, the interviewee had hiccups several times, the interviewee held his nose several times [correct answer], the interviewee laughed several times'), and one attention check asked about the content of the statement (that is, 'Which of the following persons did in the interviewee mention? Boris Johnson, Joe Biden, Angela Merkel [correct answer], Olaf Scholz, Pope Francis'). Thereafter, participants rated motivation and difficulty and were asked to name the cues they had relied on. Finally, we asked age, native language, gender (options: male; female; other, not specified here), country of origin, country of residence, contact details to provide the bonus pay and whether they opted for money or credits. Finally, participants were debriefed and explained the interviewees had been instructed to lie versus tell the truth.

**Materials.** We used 12 video statements obtained in Study 6. These videos are available from the first author after signing a non-disclosure agreement that stipulates the confidential nature of the videos and that they can only be used for research purposes. From the pool of 44 videos, we only used those for which the participants had provided consent to use their video statements in new research, that were below 4 minutes in length after cutting (see below) and where the interviewee was not wearing a face mask. Finally, we selected the videos so that the truthful and deceptive conditions were balanced in native tongue (English versus other). All participants watched the same set of 12 videos (six truthful, six deceptive). From the interview, we cut the initial rapport-building phase and the follow-up question at the end. So we selected the response to the interviewer's instruction to describe what the interviewee had done in the last half an hour in as much detail as possible, trying to fill up the 10 minutes.

Using the detail count by the trained coders of Study 6, the selected six truthful videos were found to contain more details ( $M = 6.17$ ,  $s.d. = 1.17$ ) than the selected six deceptive videos ( $M = 4.17$ ,  $s.d. = 1.17$ ),  $d = 1.71$  (95% CI: 0.30–3.03). Using a stopwatch, one team member (O.K.A.) had measured the time that the interviewee looked away from the interviewer/camera. The coding of a random subset (20%) of the statements by second team member (A.L.) spoke to the reliability of this eye-gaze aversion measurement (intraclass correlation coefficient (ICC) = 0.93). That time was converted to the percentage of the entire interview's duration. Eye-gaze aversion in the six truthful videos ( $M = 59.83\%$ ,  $s.d. = 13.70$ ) did not differ from that of the six deceptive videos ( $M = 61.50\%$ ,  $s.d. = 9.94$ ),  $d = 0.14$  (95% CI: -1.00–1.27). Thus, the

coding confirmed that detailedness, but not eye-gaze aversion, is a diagnostic cue to deception.

### Study 8

In the series of studies presented thus far, although free to use any possible cue, participants in the control conditions were not explicitly guided to use multiple cues. Study 8 addressed this concern. Participants judged a single cue (detailedness) or multiple cues (detailedness, and affect, unexpected complications and admitting lack of memory) before making binary lie–truth judgements. Ethics approval, data and materials of Study 8 can be found at <https://osf.io/z26ar/>. Hypotheses, analysis plan, and predictions were preregistered on 5 July 2022, before the start of data collection, and can be found at <https://osf.io/z26ar/>. There were no deviations from the preregistration. This was a single-blinded study (participants were not aware of the different judgement methods), with data analysis not performed blind to the conditions of the experiments. We based our sample size on ref. <sup>31</sup>. To obtain at least 500 total judgements, avoiding both low numbers of judges and senders, we planned a minimum of 84 participants in each of the two conditions, each judging 12 statements. Accounting for possible exclusions, we started with  $n = 100$ , and we continued testing until  $n \geq 84$  inclusions ( $n \geq 42$  per condition) was reached.

**Participants.** In total, 146 participants took part in Study 8. We excluded 44 participants for failing the attention check and three participants for taking part in similar research. The 99 remaining participants (35 females, 63 males, one non-binary;  $M$  age = 28.21 years,  $s.d.$  = 9.04). Due to a programming mistake, there was a slight imbalance between experimental conditions;  $n = 43$  judging multiple cues, and  $n = 56$  judging a single cue. Participants were Polish (26%), Portuguese (14%), British (10%) or one of 17 other nationalities. Age and gender were surveyed in Qualtrics, nationality was obtained from Prolific.

**Procedure.** After providing informed consent, participants were randomly assigned to the single cue or the multiple cue condition through Qualtrics. In both conditions, participants were asked to evaluate 14 statements, presented in sequential order. This included one example statement in the beginning to illustrate the procedure and one statement serving as an attention check in the end. The two statements were excluded from main analyses. There was one block with six deceptive statements and one block with six truthful statements. Blocks were presented in random order. In each block, the six statements were randomly chosen from a pool of 22 truthful or a pool of 23 deceptive statements.

In the single cue condition, participants judged the statement on detailedness (0 = not at all, to 10 = very), and were then advised to use this judgement to make a binary lie–truth judgement: ‘Truths are typically more detailed than lies. Based on your detailedness judgement, please indicate if you think the statement you have just read is true or false’. In the multiple cue condition, participants judged the statement on detailedness, affect, unexpected complications and admitting lack of memory (all on a scale from 0 = not at all, to 10 = very). Participants in the multiple cue condition were then guided to use their judgements to make a binary lie–truth decision: ‘Truths typically contain more unexpected complications, more details, more admitting lack of memory and more affect than lies. Based on your unexpected complications, detailedness, admitting lack of memory and affect judgement, please indicate if you think the statement you have just read is true or false’.

The attention check concerned the last statement, which looked like just the other statements, but with the instruction to ignore the statement and instead answer three on (each) scale. After the attention check, participants indicated their motivation, task difficulty, the cues they had relied on and their age and gender (options: male, female, non-binary/other, prefer not to say).

**Materials.** Statements were verbatim transcriptions of the video interviews obtained in Study 6. In brief, participants lied or told the truth about their whereabouts on campus. We only selected transcripts from participants who gave permission. We used 45 statements: 22 deceptive and 23 truthful statements.

The statements were coded on cues from two established credibility assessment tools, Criteria-based Content Analysis and Reality Monitoring<sup>33</sup>. Two undergraduates (N.R. and N.J.) who were acquainted with the literature and had been trained by an experienced coder scored each statement independently. We used the average of the two raters as the final cue scoring. For the multiple cue condition, we selected cues that (1) were clearly present in the statements (excluding cues such as ‘raising doubts about one’s own testimony’ and ‘self-deprecation’) and (2) were not clearly associated to detailedness (excluding cues such as ‘temporal info’ and ‘contextual embedding’). The three selected cues were affect ( $M_{\text{truthful}} = 1.32$ ,  $s.d.$  = 1.65 versus  $M_{\text{deceptive}} = 0.48$ ,  $s.d.$  = 0.66, lie–truth difference:  $d = 0.68$ ), admitting lack of memory ( $M_{\text{truthful}} = 0.43$ ,  $s.d.$  = 0.79 versus  $M_{\text{deceptive}} = 0.43$ ,  $s.d.$  = 0.95, lie–truth difference:  $d = 0.00$ ), and unexpected complications ( $M_{\text{truthful}} = 0.52$ ,  $s.d.$  = 0.86 versus  $M_{\text{deceptive}} = 0.36$ ,  $s.d.$  = 1.09, lie–truth difference:  $d = 0.16$ ). Results in full can be found at <https://osf.io/z26ar/>.

### Study 9

Although Study 8 supported our key predictions, it did not deliver decisive evidence. We therefore repeated the study with more power and with fixing some methodological shortcomings (Procedure). Participants judged a single cue (detailedness) or multiple cues (detailedness, affect, unexpected complications and spontaneous corrections) before making binary lie–truth judgements. Ethics approval, data and materials of Study 9 can be found at <https://osf.io/z26ar/>. Hypotheses, analysis plan and predictions were preregistered on 14 December 2022, before the start of data collection, and can be found at <https://osf.io/z26ar/>. There were no deviations from the preregistration. This was a single-blinded study (participants were not aware of the different judgement methods), with data analysis not performed blind to the conditions of the experiments. In consultation with the editors, we used a Bayesian sequential stopping rule to determine our sample size: we planned to start with  $n = 400$  ( $\pm 5$  due to simultaneous starting times on Prolific) and planned to add batches of  $n = 200$  until we obtained decisive evidence ( $BF_{10} > 6$  or  $BF_{10} < 1/6$ ) or the maximal sample size of  $n = 1,000$ . We reached decisive evidence in support of our hypotheses after the first batch.

**Participants.** In total, 405 participants took part in Study 9. We excluded 23 participants for failing both attention checks. The 382 remaining participants (181 females, 193 males, one prefer not to say, seven missing;  $M$  age = 29.04 years,  $s.d.$  = 8.81). There were  $n = 187$  participants in the multiple cue condition and  $n = 195$  participants in the single cue condition. Participants were South-African (28.8%), Polish (15.7%), Portuguese (14.7%) or one of 31 other nationalities. Age, gender and nationality were obtained from Prolific.

**Procedure.** We made the following notable changes to the Study 8 procedure. First, Qualtrics settings now assured that the conditions were equally balanced. Second, statements were now presented in random order. Participants randomly received one of 20 sets of 12 statements (six lies, six truths). Within each set, statements were randomly presented. Third, we changed the attention check so that there were now two multiple-choice questions about the content of a statement presented halfway the survey. Fourth, participants in the multiple cue condition now judged ‘spontaneous corrections’ ( $d = 0.63$ ) rather than ‘admitting lack of memory’ ( $d = 0.00$ ) because of its higher validity. Fifth, we provided participants in the single cue condition with the decision rule that we also used in Study 6. Hence, participants were advised to classify a statement when they had judged it with six or more

on detailedness and classify it as deceptive when they had judged it with a five or less on detailedness. Sixth, to shorten the study length and assure participants attention during the survey, we cut the statements' length by 70%. We presented the first 30% using word count, hence retaining differences in length of the statement, as text ( $M_{\text{length}} = 207.58$  words,  $s.d. = 95.06$ ).

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

### Data availability

All data are publicly available at <https://osf.io/z26ar/>.

### References

- Bond, C. F. & DePaulo, B. M. Accuracy of deception judgments. *Pers. Soc. Psychol. Rev.* **10**, 214–234 (2006).
- Bogaard, G., Meijer, E. H., Vrij, A. & Merckelbach, H. Strong, but wrong: lay people's and police officers' beliefs about verbal and nonverbal cues to deception. *PLoS ONE* **11**, e0156615 (2016).
- Aavik, T. et al. A world of lies. *J. Cross Cult. Psychol.* **37**, 60–74 (2006).
- Hartwig, M. & Bond, C. F. Why do lie-catchers fail? A lens model meta-analysis of human lie judgments. *Psychol. Bull.* **137**, 643–659 (2011).
- DePaulo, B. M. et al. Cues to deception. *Psychol. Bull.* **129**, 74–118 (2003).
- Luke, T. J. Lessons from Pinocchio: cues to deception may be highly exaggerated. *Perspect. Psychol. Sci.* **14**, 646–671 (2018).
- Sporer, S. L., Masip, J. & Cramer, M. Guidance to detect deception with the aberdeen report judgement scales: are verbal content cues useful to detect false accusations? *Am. J. Psychol.* **127**, 43–61 (2014).
- Weinberger, S. Airport security: intent to deceive? *Nature* **465**, 412–415 (2010).
- Hauch, V., Sporer, S. L., Michael, S. W. & Meissner, C. A. Does training improve the detection of deception? A meta-analysis. *Commun. Res.* **43**, 283–343 (2016).
- Street, C. N. H. & Richardson, D. C. The focal account: indirect lie detection need not access unconscious, implicit knowledge. *J. Exp. Psychol. Appl.* **21**, 342–355 (2015).
- Gigerenzer, G. & Brighton, H. Homo heuristicus: why biased minds make better inferences. *Top. Cogn. Sci.* **1**, 107–143 (2009).
- Kleinberg, B., van der Toolen, Y., Vrij, A., Arntz, A. & Verschuere, B. Automated verbal credibility assessment of intentions: the model statement technique and predictive modeling. *Appl. Cogn. Psychol.* **32**, 354–366 (2018).
- Gigerenzer, G., Todd, P. M. & the ABC Research Group. *Simple Heuristics That Make Us Smart* (Oxford Univ. Press, 1999).
- Tversky, A. & Kahneman, D. Judgment under uncertainty: heuristics and biases: biases in judgments reveal some heuristics of thinking under uncertainty. *Science* **185**, 1124–1131 (1974).
- Dressel, J. & Farid, H. The accuracy, fairness, and limits of predicting recidivism. *Sci. Adv.* **4**, eaao5580 (2018).
- Salganik, M. J. et al. Measuring the predictability of life outcomes with a scientific mass collaboration. *Proc. Natl Acad. Sci. USA* **117**, 8398–8403 (2020).
- Gigerenzer, G. & Gaissmaier, W. Heuristic decision making. *Annu. Rev. Psychol.* **62**, 451–482 (2011).
- Johnson, M. K. & Raye, C. L. Reality monitoring. *Psychol. Rev.* **88**, 67–85 (1981).
- Nahari, G., Vrij, A. & Fisher, R. P. Exploiting liars' verbal strategies by examining the verifiability of details. *Leg. Criminol. Psychol.* **19**, 227–239 (2014).
- Lakens, D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Front. Psychol.* **4**, 863 (2013).
- Youden, W. J. Index for rating diagnostic tests. *Cancer* **3**, 32–35 (1950).
- Weinberger, S. Terrorist 'pre-crime' detector field tested in United States. *Nature* <https://doi.org/10.1038/news.2011.323> (2011).
- Boffey, D. EU border 'lie detector' system criticised as pseudoscience. *The Guardian* (2 November 2018).
- Vrij, A., Fisher, R., Mann, S. & Leal, S. Detecting deception by manipulating cognitive load. *Trends Cogn. Sci.* **10**, 141–142 (2006).
- Kleinberg, B., Arntz, A. & Verschuere, B. Being accurate about verbal credibility assessment. Preprint at *PsyArXiv* <https://doi.org/10.31234/OSF.IO/H6PXT> (2019).
- Evans, J. R. & Michael, S. W. Detecting deception in non-native English speakers. *Appl. Cogn. Psychol.* **28**, 226–237 (2014).
- Markowitz, D. M. & Hancock, J. T. in *Handbook of Language Analysis in Psychology* (eds M. Dehghani & R. L. Boyd) 274–284 (Guilford Press, 2022).
- Ott, M., Choi, Y., Cardie, C. & Hancock, J. T. Finding deceptive opinion spam by any stretch of the imagination. In *Proc. 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (eds Lin, D et al.) 309–319 (Association for Computational Linguistics, 2011).
- Rudin, C. Why black box machine learning should be avoided for high-stakes decisions, in brief. *Nat. Rev. Methods Primers* **2**, 81 (2022).
- Verschuere, B., Schutte, M., van Opzeeland, S. & Kool, I. The verifiability approach to deception detection: a preregistered direct replication of the information protocol condition of Nahari, Vrij, and Fisher (2014b). *Appl. Cogn. Psychol.* **35**, 308–316 (2021).
- Levine, T. R., Daiku, Y. & Masip, J. The number of senders and total judgments matter more than sample size in deception-detection experiments. *Perspect. Psychol. Sci.* **17**, 191–204 (2021).
- Verigin, B. L., Meijer, E. H., Vrij, A. & Zauzig, L. The interaction of truthful and deceptive information. *Psychol. Crime. Law* **26**, 367–383 (2020).
- Oberlader, V. A. et al. Validity of content-based techniques to distinguish true and fabricated statements: a meta-analysis. *Law Hum. Behav.* **40**, 440–457 (2016).

### Acknowledgements

We thank A. El Feddali and V. Giannelli for their help with the Pilot Study, E. Wevers, R. Louterse and J. Wong for their help with Study 6, A. Lob for his help with Study 7 and Study 9, N. Jebriel and N. Roijakkers for their help with Study 8 and M. Vestjens and S. Wiechert for their help with Study 9. E.M. is supported by the Israel Institute for Advanced Studies. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the paper.

### Author contributions

B.V. and E.M. conceptualized the study. C.-C.L., S.H., M.W., L.L., T.v.G., E.C., O.K.A., E.M. and B.V. were responsible for the methodology. C.-C.L., S.H., M.W., L.L., T.v.G., E.C. and O.K.A. were responsible for the investigation. B.V. and B.K. conducted the statistical analyses. B.V. and E.M. wrote the original draft. B.V., E.M. and B.K. reviewed and edited the paper.

### Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41562-023-01556-2>.

**Correspondence and requests for materials** should be addressed to Bruno Verschuere.

**Peer review information** *Nature Human Behaviour* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature Limited 2023, corrected publication 2023

## Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- |     |           |
|-----|-----------|
| n/a | Confirmed |
|-----|-----------|
- The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
  - A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
  - The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
  - A description of all covariates tested
  - A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
  - A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
  - For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
  - For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
  - For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
  - Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

- |                 |  |
|-----------------|--|
| Data collection | Data were collected using Qualtrics software (versions 2020, 2021, 2022; Qualtrics, Provo, UT, USA. <a href="https://www.qualtrics.com">https://www.qualtrics.com</a> )  |
| Data analysis   | Studies 1-2-3-4 were analyze with R script (R Core Team, 2022).). R can be downloaded for free from <a href="https://cran.r-project.org/bin/windows/base/">https://cran.r-project.org/bin/windows/base/</a> . Studies 5-6-7-8-9 were analyzed with JASP 0.16.0.0. JASP is based on R and can be downloaded for free from <a href="https://jasp-stats.org/">https://jasp-stats.org/</a> |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

Full data of all 9 studies are publicly available : <https://osf.io/z26ar>

## Human research participants

Policy information about [studies involving human research participants and Sex and Gender in Research.](#)

### Reporting on sex and gender

Sex and Gender are not considered in the study design, and not theorized to be of influence on the study findings, and therefore not a variable of interest. We therefore did not conduct sex or gender based analyses. To describe our samples, participants self-reported their gender (Study1-6-7-8) or sex (Studies 2-3-4-5-9). Participants consented to publicly share their data.

### Population characteristics

Study1: Included in final analyses were 39 participants (31 female, 8 male; M age = 19.38 years, SD = 1.68); Study2: Included in final analyses were 108 participants (39 female, 69 male; M age = 29.69 years, SD = 10.34); Study3: Included in final analyses were 230 participants (107 female, 119 male, 4 missing; M age = 30.32 years, SD = 10.59); Study4: Included in final analyses were 192 participants (92 females; M age = 27.59 years, SD = 9.11). Study5: included in the final analyses were 150 participants (83 females; M age = 26.80 years, SD = 8.48); Study6: Included in final analyses were 44 participants (n = 23 where in the truthful condition: Mage = 19.87, SD = 2.70; 43.5% native English speakers; 78% female, 22% male, and n = 21 where in the deceptive condition (Mage = 19.48, SD = 1.12; 47.6% native English speakers; 52% female, 43% male, 5% non-binary). Study7: included in the final analyses were 171 participants (123 female, 44 male, 4 other) had a mean age of 22.07 years (SD = 5.03); Study8: included in the final analyses were 99 participants (35 females, 63 males, 1 non-binary; M age = 28.21 years, SD = 9.04). Study9: included in the final analyses were 382 participants (181 females, 193 males, 1 prefer not to say, 7 missing; M age = 29.04 years, SD = 8.81).

### Recruitment

Participants in Study 1, 6 and 7 were recruited through an online research portal of the University of Amsterdam (mostly but not restricted to psychology undergraduates). Participants selected themselves for study participation. Participants in Study 2, 3, 4, 5, 8 and 9 were crowdsourced through Prolific, a platform that has access to studies that require high participant engagement. Prolific has participants from most OECD countries (see <https://www.oecd.org/about/members-and-partners/>). There is evidence to suggest that Prolific participants score higher on comprehension, attention, and honesty, than other crowdsourcing platforms. Participants selected themselves for study participation on a first come first serve basis.

### Ethics oversight

Our research complies with the guidelines formulated by the Ethics Review Board of the Faculty of Social and Behavioral Sciences, University of Amsterdam (Amsterdam, The Netherlands). Informed consent was obtained by all participants.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences  Behavioural & social sciences  Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

### Study description

For all 9 studies: Quantitative, experimental. More specifically:

Study1. Experimental Design. Participants were randomly assigned to one of two Judgment Methods (Judge Deception vs Judge Verifiability, between-subjects). All participants judged truthful and deceptive statements (Veracity: Truthful vs Deceptive, within-subjects).

Study2-3. Experimental Design. Participants were randomly assigned to one of three Judgment Methods (Judge Deception vs Judge Verifiability vs Judge detailedness, between-subjects). All participants judged truthful and deceptive statements (Veracity: Truthful vs Deceptive, within-subjects).

Study4. Experimental Design. Participants were randomly assigned to one of two Judgment Methods (Judge Deception vs Judge Detailedness, between-subjects). All participants judged truthful and deceptive statements (Veracity: Truthful vs Deceptive, within-subjects).

Study5. Experimental Design. Participants were randomly assigned to one of two goal awareness condition (Lie detection goal explicit vs lie detection goal implicit, between-subjects). All participants judged truthful and deceptive statements (Veracity: Truthful vs Deceptive, within-subjects), on detailedness.

Study6. Experimental Design. Participants were randomly assigned to one of two guilt conditions (Veracity: Truthful vs Deceptive, between-subjects) before being interviewed

Study7. Experimental Design. Participants were randomly assigned to one of two Judgment Methods (Judge Richness in detail vs Judge Eye Gaze Aversion). All participants

judged truthful and deceptive statements (Veracity: Truthful vs Deceptive, within-subjects).

Study8-9. Experimental Design. Participants were randomly assigned to one of two Judgment Methods (Judge Only Richness in detail vs Judge 4 cues including Richness in detail). All participants judged truthful and deceptive statements (Veracity: Truthful vs Deceptive, within-subjects).

## Research sample

Study1 Undergraduate participants (University of Amsterdam, The Netherlands). N=39 inclusions (31 female, 8 male; M age = 19.38 years, SD =1.68). Non-representative sample. Rationale: The Study1 statements that participants were asked to judge were in Dutch and about the UvA campus, we therefore selected native Dutch participants likely familiar with the UvA campus.

Study2-3 sought to replicate Study1 findings. We moved from locally recruited undergraduates to online crowdsourcing using Prolific.co participants with Dutch as first language. Rationale: This allowed to more efficiently collect data (studies of this size can be run within the day on Prolific) and get first signs of generalizability (e.g., UvA>Prolific; native Dutch>first language Dutch). Being fluent in Dutch was a requirement as to-be-judged statements were in Dutch. Study2 had N=108 inclusions (39 female, 69 male; M age = 29.69 years, SD = 10.34; n = 30 judging deception, n = 39 judging verifiability, and n = 39 judging detailedness) mostly had the Dutch nationality (73%; Belgian: 24%; Other: 3%). Study3 had N= 230 inclusions (107 female, 119 male, 4 missing; M age = 30.32 years, SD = 10.59; n = 77 judging deception, n = 89 judging verifiability, and n = 64 judging detailedness) 72% had the Dutch nationality (Belgian: 26%; Other: 2%).

Study4 Fluent-German crowdsourced participants. N= 192 inclusions (92 females; M age = 27.59 years, SD = 9.11; n = 104 judging deception, n =88 judging detailedness) were Polish (26%), German (13%) or had one of 27 other nationalities. Rationale: The to-be-judged Study4 statements were in German so we selected Fluent German participants.

Study5. Fluent Dutch-speaking participants from Prolific. N= 150 inclusions (83 females; M age = 26.80 years, SD = 8.48; n = 76 in the explicit condition and n = 74 in the non-explicit condition) were Dutch (55.33%), Belgian (31.33%) or had another nationality (13.33%). About half of them (52%) were students. Rationale: Being fluent in Dutch was a requirement as to-be-judged statements were in Dutch.

Study6. Undergraduate participants from the University of Amsterdam. N=44 inclusions with n = 23 in the truthful condition (M age =19.87, SD = 2.70; 43.5% native English speakers; 78% female, 22% male), and n = 21 in the deceptive condition (M age = 19.48, SD = 1.12; 47.6% native English speakers; 52% female, 43% male, 5% non-binary). Rationale: Convenience sample UvA (mock crime study). Study6 interviews were conducted in English so that they could be more widely used in follow-up research (including Study7).

Study7. Convenience sample of University of Amsterdam participants. N=171 inclusions (123 female, 44 male, 4 other) had a mean age of 22.07 years (SD = 5.03). Eighty-six participants judged detailedness, and 85 judged eye gaze aversion. They had Dutch (31%), English (17.5%) or another (51.5%) language as mother tongue. Their country of origin was The Netherlands (30%), Germany (8%) or one of 40 other nationalities. Rationale: Convenience sample UvA

Study8. Participants from Prolific. 99 inclusions (35 females, 63 males, 1 non-binary; M age = 28.21 years, SD = 9.04). Due to a programming mistake, there was a slight imbalance between experimental conditions; n = 43 judging multiple cues, and n = 56 judging a single cue. Participants were Polish (26%), Portuguese (14%), British (10%) or had one of 17 other nationalities. Rationale: Convenience sample Prolific allows for efficient data collection. As statements were now in English, no language restrictions were deemed necessary and the entire participant pool could subscribe.

Study9. Participants from Prolific. N=382 inclusions (181 females, 193 males, 1 prefer not to say, 7 missing; M age = 29.04 years, SD = 8.81). There were n = 187 participants in the multiple cue condition, and n = 195 participants in the single cue condition. Participants were South-African (28.8%), Polish (15.7%), Portuguese (14.7%) or had one of 31 other nationalities. Rationale: Convenience sample Prolific allows for efficient data collection. As statements were now in English, no language restrictions were deemed necessary and the entire participant pool could subscribe.

## Sampling strategy

Study1 was a non-preregistered pilot study. We used a convenience sample of the UvA.

Study2 was preregistered and used a convenience sample of Prolific participants. We aimed for N=150 Prolific participants (N=50 for each of the 3 conditions). We based our sample size justification on the planned follow up t tests to have 90% power for a one-tailed paired sample t test ( $\alpha = .05$ ) to be able to pick up the lie-truth effect of the size seen in Study1 ( $d = .76$ ). This required N=51. Anticipating that the effect could be smaller than that observed in Study1 and anticipating exclusions we decided to test N=150. Due to simultaneous starting times in Prolific we may end up with slight more participants that start the study (+155, based on our experience with Prolific).

Study3 was preregistered, and is identical in design as Study2 (the sole difference is that it has a new, larger sample), also using a convenience sample of Prolific participants. Study2 showed the expected, significant, and large lie-truth difference when people relied on the Use-the-best Heuristic. Yet, it was not significantly higher than the lie-truth difference in the control condition – which happened to be larger than we anticipated ( $d = 0.39$ ). We reasoned this was sampling error due to modest sample size. To be able to obtain the predicted interaction, we therefore decided (1) to run Study2 again, now with double the sample size (Aimed for N=300, Prolific participants due to simultaneous starting times we obtained N=303), and (2) merge the data from Study2 and Study3.

Study4 was preregistered, and used a convenience sample of Prolific participants. We opened n=250 spots on Prolific (due to simultaneous starting times, we ended up with N=251). This sample size was determined considering three design aspects: First, we wish to obtain reliable estimates for each judgement method. Recent simulation studies simulations that came available at the time of research (April 2021 see <https://www.youtube.com/watch?v=1KUX8CuXAgM>) show that at least 1000 judgements (arising from number of statements x number of judges) are needed for stable estimates. To collect 1000+ judgements in each design cell of each study, we need a minimum of 84 participants (each judging 12 statements) for each of two judgements methods. Second, we wish to have high power (95%) to pick up the effect of interest, which is the interaction between Statement Veracity and Judgement Method. In our initial work the observed effect size was  $\eta^2 p = .03$  (Study1) and  $\eta^2 p = .20$  (Study2). To provide for a conservative estimate of the effect (given we examine generalization to a novel context), we rely on the smallest obtained effect size. G-POWER

shows that a minimum of  $n=108$  is needed to pick up that effect with 95% power and a significance threshold of .05. Third, in our previous studies exclusion criteria led to +24% exclusions (though we now set exclusion more liberal; i.e. only excluding participants who fail BOTH attention checks).

Study5 was preregistered, and used a convenience sample of Prolific participants. We opened  $N=164$  spots on Prolific and obtained  $N=166$  due to simultaneous starting times. This sample size was determined considering two design aspects: 1./ We wish to obtain reliable estimates within each condition. Recent simulation studies simulations show that at least 1000 judgements (arising from number of statements  $\times$  number of judges) are needed for stable estimates. To collect 1000+ judgements in each condition, we need a minimum of 63 participants (each judging 16 statements), totaling a minimum  $N$  of 126. 2./ Using the exclusion criteria described above, we previously excluded +23% of participants in similar research.

Study6 was exploratory. We aimed to test as many participants at UvA as possible within the time frame of the bachelor thesis project (test period: Dec 2021). Given the large lie-truth differences with the Use-the-best heuristic obtained in studies 1 to 5 ( $d \geq .75$ ),  $N \geq 17$  would have 90% power to pick up such a large lie-truth difference. We included  $N=44$ .

Study7 was preregistered, and used a convenience sample from the UvA. We open spots for  $N=200$  on the UvA participant recruitment tool lab.uva.nl and would collect data for maximally a month. We obtained  $N=205$ . This sample size was determined considering (1) The desire to have reliable estimates within each condition, with simulations (Levine et al 2021) showing that 1000+ judgements (arising from number of statements  $\times$  number of judges) are preferable to obtain for stable estimates. To collect 1000+ judgements in each condition, we need a minimum of 83 participants per condition (each judging 12 statements), totaling a minimum  $N$  of 166, (2) a MORE POWER 6.0 calculation showing that to have 90% power to detect the within-between interaction of partial eta squared = 0.13 (based on the interaction we obtained in Study2-3; detail vs control) in the ANOVA with significance testing requires a minimum total  $N$  of 72 (but note that our confirmatory analyses are Bayesian), and (3) taking into account an expected +23% exclusion rate.

Study8 was preregistered, and used a convenience sample of Prolific participants. We based our sample size on Levine et al. (2022). To obtain at least 500 total judgements, avoiding both low numbers of judges and senders, we planned a minimum of 84 participants in each of the two conditions each judging 12 statements. Accounting for possible exclusions, we started with  $n = 100$ , and we continued testing until  $n \geq 84$  inclusions ( $n \geq 42$  per condition) was reached.

Study9 was preregistered, and used a convenience sample of Prolific participants. In consultation with the editors, we used a Bayesian sequential stopping rule to determine our sample size: We started with  $n=400$  ( $\pm 5$  due to simultaneous starting times on Prolific) and planned to add batches of  $n=200$  until we obtained decisive evidence ( $BF_{10} > 6$  or  $BF_{10} < 1/6$ ) or the maximal sample size of  $n=1,000$ . We reached decisive evidence in support for our hypotheses after the first batch.

#### Data collection

Data were collected online using Qualtrics software. We did not track whether others were present while participants performed the study (it is possible). Participants were assigned to conditions by Qualtrics software without interaction with the experimenter (Studies 1-2-3-4-5-7-8-9). In Study6 participants were interviewed by a condition-blind experimenter. Researchers were not blind to study hypotheses.

#### Timing

Study1: December, 2020; Study2: March, 2021; Study3: March, 2021; Study4: July, 2021; Study5: October, 2021; Study6: December, 2021; Study7: April-May, 2022; Study8- July 2022; Study9: December 2022

#### Data exclusions

Study1: From the 51 participants we excluded 7 participants who failed an attention check and 5 participants were not Dutch native speakers. Exclusion criteria were discussed and planned with the research team prior to data analysis, but not preregistered. Findings do not hinge on the exclusion criteria as results are similar when not excluding anyone.

Study2: From the 142 participants we excluded 34 participants who failed either of the two attention checks. Exclusion criteria were preregistered. Findings do not hinge on the exclusion criteria as results are similar when not excluding anyone.

Study3: From the 303 participants we excluded 73 participants who failed either of the two attention checks. Exclusion criteria were preregistered. Findings do not hinge on the exclusion criteria as results are similar when not excluding anyone.

Study4: From the 251 participants we excluded 59 participants who failed both attention checks (note that our exclusion criterion is more liberal than that used in Studies 1-2-3 as we noted that our findings in Studies1-2-3 did not depend on the more conservative exclusion criteria, hence it may be too strict and unnecessarily excluding participants). Exclusion criteria were preregistered. Findings do not hinge on the exclusion criteria as results are similar when not excluding anyone.

Study5: From the 166 participants we excluded 16 participants who failed both attention checks. Exclusion criteria were preregistered. Findings do not hinge on the exclusion criteria as results are similar when not excluding anyone.

Study6: From the 47 participants, we excluded 3 participants (2 because they did not complete their mission, and 1 because of suspected intoxication). Exclusion criteria were discussed and planned with the research team prior to data analysis, but not preregistered. Findings do not hinge on the exclusion criteria as results are similar when not excluding anyone.

Study7: From the 205 participants we excluded 34 participants who failed both attention checks. Exclusion criteria were preregistered. Findings do not hinge on the exclusion criteria as results are similar when not excluding anyone.

Study8: From the 146 participants, we excluded 44 participants for failing the attention check, and 3 participants for taking part in similar research. Exclusion criteria were preregistered. Findings do not hinge on the exclusion criteria as results are similar when not excluding anyone.

Study9: From the 405 participants, we excluded 23 participants for failing both attention checks. Exclusion criteria were preregistered. Findings do not hinge on the exclusion criteria as results are similar when not excluding anyone.

#### Non-participation

Participants self-selected to participate. Non-participation is not tracked by lab.uva.nl nor Prolific.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involvement in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

### Methods

n/a	Involvement in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging