WORKING PAPERS

N° 155

Lie-detection algorithms attract few users but vastly increase accusation rates

Alicia von Schenk, Victor Klockmann, Jean-François Bonnefon, Iyad Rahwan and Nils Köbis

Institute for Advanced Study in Toulouse

Lie-detection algorithms attract few users but vastly increase accusation rates

[1,2]* Alicia von Schenk, [1,2]*Victor Klockmann, [3,4]Jean-François Bonnefon, [2]Iyad Rahwan & [2]Nils Köbis

* shared first authorship

1 - Department of Economics, University of Würzburg

2 - Max Planck Institute for Human Development, Center for Humans and Machines

3 - Institute for Advanced Study Toulouse

4 - Toulouse School of Economics

Correspondence to: Nils Köbis, koebis@mpib-berlin.mpg.de

Acknowledgments: JFB acknowledges support from grant ANR-19-PI3A-0004, grant ANR-17-EURE-0010, and the research foundation TSE-Partnership.

Abstract

People are not very good at detecting lies, which may explain why they refrain from accusing others of lying, given the social costs attached to false accusations — both for the accuser and the accused. Here we consider how this social balance might be disrupted by the availability of lie-detection algorithms powered by Artificial Intelligence (AI). Will people elect to use lie-detection AI that outperforms humans, and if so, will they show less restraint in their accusations? To find out, we built a machine learning classifier whose accuracy (66.86%) was significantly better than human accuracy (46.47%) lie-detection task. We conducted an incentivized lie-detection experiment (N = 2040) in which we measured participants' propensity to use the algorithm, as well as the impact of that use on accusation rates and accuracy. Our results reveal that (a) requesting predictions from the lie-detection AI and especially (b) receiving AI predictions that accuse others of lying increase accusation rates. Due to the low uptake of the algorithm (31.76% requests), we do not see an improvement in accuracy when the AI prediction becomes available for purchase.

Introduction

People lie a lot (DePaulo et al. 1996, Pascual-Ezama et al. 2020, Serota et al. 2010, Tergiman and Villeval 2022). In many contexts, it would be advantageous to detect lies and call them out (van den Assem et al. 2012, Köbis et al. 2022, Turmunkh et al. 2019, Warren and Schweitzer 2018). While some methods help with lie-detection (Nahari et al. 2014, Verschuere et al. 2023), the time, effort, and skill they require place them beyond the reach of ordinary people. Accordingly, recent studies (Pascual-Ezama et al. 2021) and large-scale meta-analyses indicate that people do not perform much better than chance when trying to detect lies (Hartwig and Bond 2011, Hauch et al. 2016). This general poor performance in lie-detection may explain why people typically refrain from accusing others of lying (Gilbert 1991, Levine et al. 1999). Indeed, not being able to discern truth from lies increases the risk of making false accusations, which are harmful both to the accused and to the accuser. False accusations can harm the accused because of the social stigma of being called a liar, and they can, in turn, harm the accuser, who is held accountable for unjustly tarnishing the reputation of the accused. Since people are generally bad at detecting lies, it may be a safer strategy to refrain from lying accusations that can hurt both the accuser and the accused if they are unfounded.

As a corollary, anything that would reduce either the harm to the accused or the accountability of the accuser may upend our current social balance and increase the rate at which people accuse each other of lying. For example, the harm of false accusations to the accused can be reduced by systematic fact-checking. Currently, this time-consuming process is mostly reserved for high-stakes accusations (e.g., in judicial or political contexts) and is unlikely to be available in all accusation contexts. Technology may change that, if fact-checking can be automated and scaled up, but the real technological game-changer

may consist of automatic lie-detection that decreases the accountability of the accuser rather than automated fact-checking that reduces harm to the accused.

Indeed, progress in Artificial Intelligence (AI) is opening a new chapter in the long history of lie-detecting machines. While older machines such as the polygraph have questionable accuracy (Saxe et al. 1985), current Natural Language Processing algorithms can detect fake reviews (Pérez-Rosas et al. 2017) and achieve higher-than-chance accuracy for lie-detection (Kleinberg and Verschuere 2021). If this AI technology continues to improve and becomes massively available, it may disrupt the current social balance in which people largely refrain from accusing each other of lying.

Imagine a world in which everyone has access to a superhuman lie-detection technology, such as Internet browsers that screen social media posts for lies; algorithms that check CVs for deception; or video conferencing platforms that give real-time warnings when one's interlocutor or negotiation partner seems to be insincere, as is not unusual in negotiations (Gaspar and Schweitzer 2013). Consulting a lie-detection algorithm, or delegating accusations to the algorithm, could reduce accusers' sense of accountability, increase the psychological distance from the accused, and blur questions of liability (Hohenstein and Jung 2020, Köbis et al. 2021), resulting in higher accusation rates.

This assumes, however, that people do elect to use such AI tools for lie-detection. We know that people are often reluctant to use algorithms, especially when the algorithms are not error-proof (Dietvorst et al. 2018), and that this aversion is especially high in emotion-laden domains (Castelo et al. 2019). As a result, the disruptive potential of lie detection algorithms may be neutralized or delayed by low adoption. In this work, we develop a lie-detection algorithm whose accuracy is better than that of humans, and we conduct an incentivized lie-detection experiment in which we measure participants'

propensity to use the algorithm, as well as the impact of that use on accusation rates and accuracy. We manipulate, on the individual level, the choice and availability of the lie-detection algorithm and show that both have an impact on accusation rates.

Methods

Overview of Studies & Open Science Statement. As preparation for our main study, in the Statement Collection Study, we collected a dataset of true and false statements to be used for algorithm training and our lie-detection task. This dataset was collected in January 2022. We then conducted a first pilot study on the use of lie-detection algorithms in April 2022. The main Judgement study took place in May 2023. All data collections were approved by the Ethics Committee of the *blinded for review.* Participants provided informed consent at the start of the study. Data sets and STATA analysis scripts for the analyses for the pilot study and the results reported below, as well as the pre-registrations, are available on the Open Science Framework (https://osf.io/eb59s/?view_only=bf8c8f966c084941a59b117126e4aea8).

Statement Writing Study. We recruited 986 participants via Prolific.co and asked them to describe something they intended to do during the next weekend (a neutral and not politically loaded context). While some studies let people decide whether to lie or not (Erat and Gneezy 2012, Gneezy 2005, Leib et al. 2021), we adopted standard procedures in research on lie-detection and elicited true and false statements from each participant (Kleinberg and Verschuere 2021, Verschuere et al. 2018). Participants were first asked to write a true statement together with a supporting text that their statement was indeed truthful. Afterward, they saw the activities of four other participants and were asked to indicate which of them they were not going to carry out. One of the selected activities was then picked at random, and participants then wrote a false statement with incentives to

write convincingly (they earned a bonus of £2 if a future participant judged their statement to be true, see details below). They were not informed beforehand that they would have to write a false statement after the truthful one.

This approach has the advantage of obtaining better training data for the lie-detection algorithm because (a) we avoid selection bias of lies stemming from the endogenous choice by participants, and (b) true and false statements are perfectly balanced in the training dataset.

Two research assistants coded the quality of these statements. First, they checked whether the participant followed the instructions and wrote meaningful sentences with at least 150 characters. Second, for truthful statements, they checked whether the author's supporting text fitted the statement (see Supplemental Material). Participants were excluded from the dataset if either of their statements did not meet either of these criteria. After this quality check, our dataset contained 1536 statements from 768 authors.

Lie-Detection Algorithm. To generate a lie-detection algorithm that performed better than humans, we relied on the open-source BERT language model by Google (Devlin et al. 2018). To obtain an out-of-sample prediction of truthfulness for each statement, we repeatedly performed an 80:20 split of the data. Specifically, we first split the 1536 statements into five equally sized subsets. We then trained the model with 80% of the data and tested it with the remaining 20%, and repeated this step five times. Accuracy was then measured by the total performance of the algorithm across all five folds. The algorithm reached 66.86% accuracy (=correctly identifying lies as lies and truthful statements as truth), which was deemed sufficient to proceed to the next stage, given the expectation that humans would not perform much better than chance level at this task.

Judgment Study. For the main Judgment Study, we selected 510 statements at random while preserving the algorithm's confusion matrix (see Table S1A & S1B in the Supplementary Material). We recruited 2040 participants via Prolific (M_{Age} = 36.93, SD_{Age} = 10.63, 38.38% female). In all experimental treatments, each participant read one statement and decided whether to judge this statement as truthful or whether to accuse the author of lying. They were incentivized for accuracy (bonus of £0.50 for a correct guess) and were informed beforehand that half of all statements in the underlying dataset were truthful and the other half were lies. They also learned that the statement they would see was picked randomly. To understand the consequences of their decisions, participants learned that when they accused an author of lying, the payoff of this author would decrease by £2.

Participants were randomly assigned on an individual level to one of four treatments that orthogonally manipulated the choice and availability of the lie-detection algorithm. As a result of a 2 (choice vs. forced) x 2 (AI available vs. not) between-subjects design, four treatments exist. Namely, in the *Baseline* treatment, participants judged the statements without any lie-detection AI available, and they did not learn about the existence of a lie-detection algorithm until after their judgment. In the *Forced* treatment, all participants receive the prediction of the lie-detection algorithm. In the *Blocked* and *Choice* treatments, participants could request the prediction of the lie-detection algorithm ("true statement" or "false statement") for a small cost of £0.05. They were informed that there was a chance that the prediction would be blocked. In all cases in the *Blocked* treatment, the prediction was blocked. Participants were simply informed that the prediction was unavailable (without further elaboration), and their purchase was refunded. Hence, in this treatment, some participants requested but did not receive the AI prediction. In the *Choice* treatment, all participants who requested an AI prediction received it and were charged the £0.05 cost (see Figure 1).

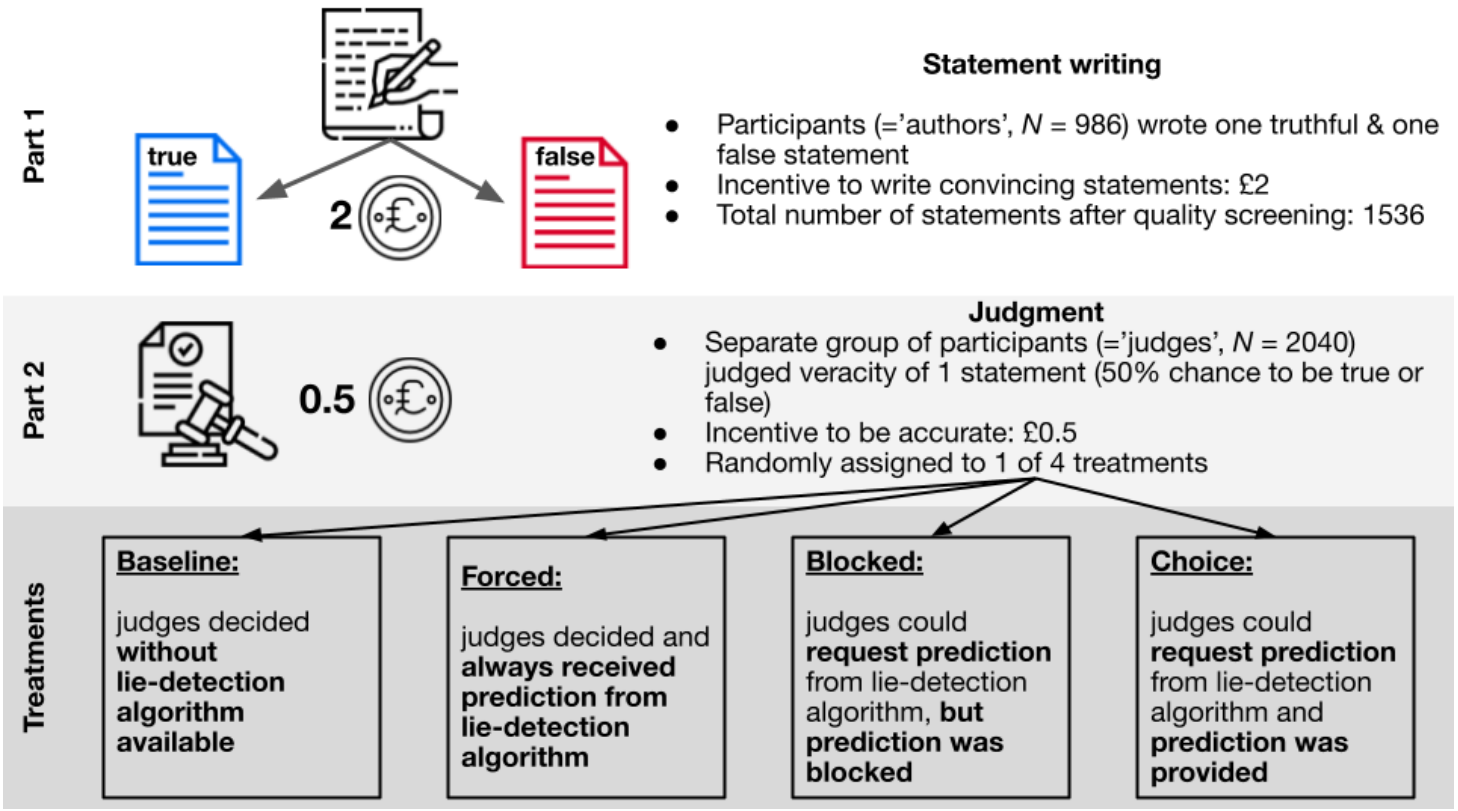


Figure 1 | Overview of the study design. The study consisted of two parts. In Part 1, participants (=authors) wrote one true and one false statement; In Part 2, a separate sample of participants (=judges) judged one randomly drawn statement in four different treatments: In the *Baseline*, judges decided by themselves, without any lie-detection algorithm available; in the *Forced* treatment, all judges received a prediction from the lie-detection algorithm; in the *Blocked* treatment, judges could request a prediction from a lie-detection algorithm, but that prediction was blocked; in the *Choice* treatment judges could request a prediction from a lie-detection algorithm, and that prediction was provided.

At the end of the study, all participants answered a series of questions measuring their beliefs about the accuracy of the algorithm (percentage of mistakes, percentage of false accusations, accuracy compared to the average human, accuracy compared to themselves) and feelings of guilt when accusing somebody of lying in this study. We informed those participants who did not have access to the lie-detection algorithm about the existence of an intelligent algorithm that was designed to predict the truthfulness of statements. This elicitation allowed us to assess whether the decision to use the algorithm correlated with subjective expectations about its performance.

Results

Human and Algorithmic performance. Participants achieved a 46.47% accuracy rate, in line with previous findings documenting people's inability to discern truthful statements from lies (Verschuere et al. 2018). The accuracy of their accusations was even lower than chance, albeit not significantly so (40.82%, t = -1.84, p = 0.07).

The lie-detection algorithm achieved an overall 66.86% accuracy rate which is comparable to previously developed lie-detection algorithms (Kleinberg and Verschuere 2021). It significantly exceeds both random guessing (t = 8.08, $p < 0.001$) and human performance (t = 6.16, $p < 0.001$). As can be seen in the confusion matrix in the Supplementary Material, Table S1A & S1B, we observe higher accuracy for untruthful statements (80.78%) than truthful statements (52.94%).

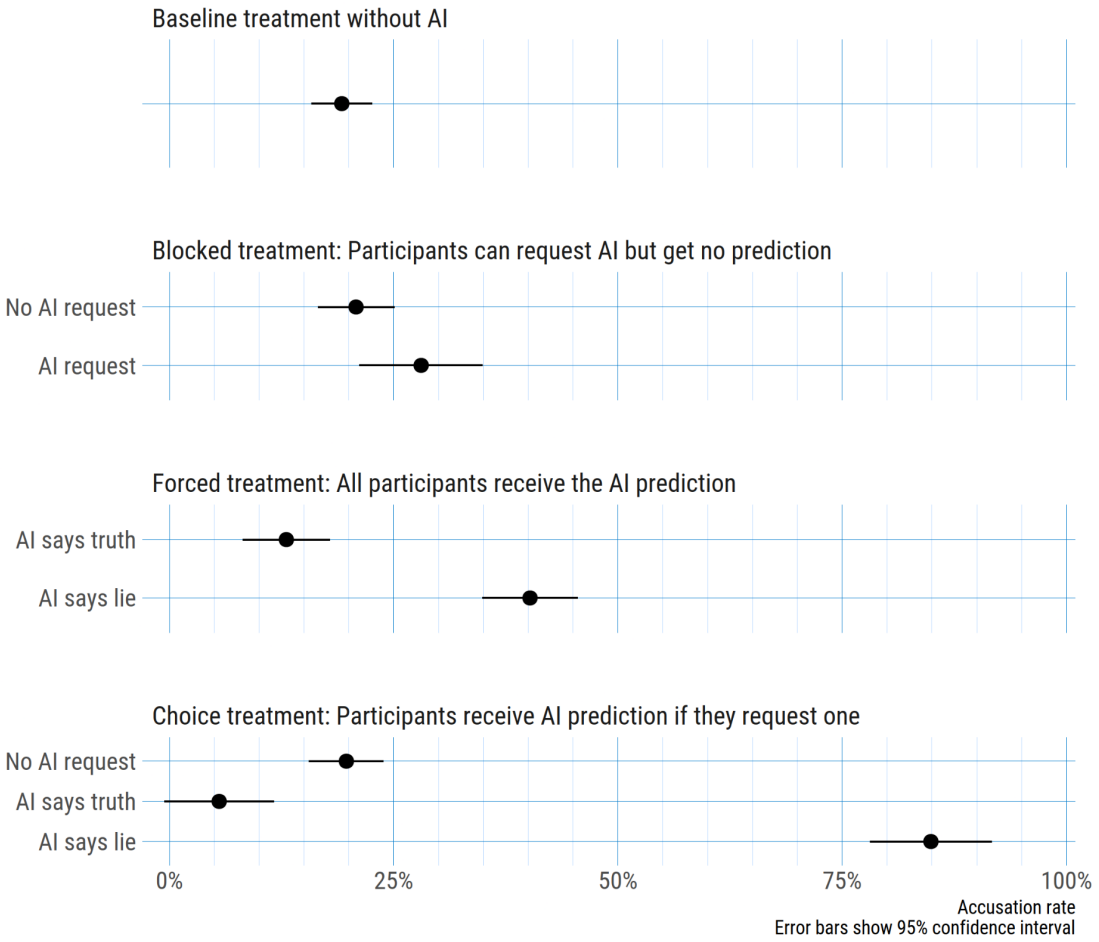


Figure 2 | Accusation rates across treatments. The figure plots the estimate (and 95% confidence intervals) of accusation rates across treatments, across request status AI predictions (request vs. not), and AI prediction type (truth vs. lie).

Lie Accusations. In the *Baseline* treatment (no lie-detection algorithm available), the accusation rate was 19.22%, even though participants knew that 50% of all statements in the underlying database were lies. This finding confirms the assumption that people typically refrain from accusing others of lying (Gilbert 1991).

In the *Blocked* treatment, the overall accusation rate was 23.14% (see also Figure 2), and the accuracy of these accusations was again not different from chance level (48.43%, t = -0.71, p = 0.48). We see that 32.16% of participants requested a prediction (but did not

obtain it). Accusation rates only slightly differed between those who did (28.05%) and those who did not request a prediction from AI (20.81%) ($\chi^2 = 3.28$, $p = .07$). Accuracy rates and false accusation rates between those who requested (and were denied) AI predictions and those who did not request do not differ significantly (accuracy 48.78% vs. 48.27%, $\chi^2 = 0.01$, $p = 0.91$; false accusations 30.68% vs. 21.56%, $\chi^2 = 2.58$, $p = 0.11$). There is thus no strong evidence for a systematic difference between algorithm adopters and non-adopters regarding their accusation behavior. Further, there is no significant difference in accusations between the *Blocked* treatment and the *Baseline* treatment ($\chi^2 = 2.35$, $p = 0.13$). The mere information on the existence of an algorithm thus does not affect choices.

In the *Forced* treatment, where all participants passively received AI predictions, the overall accusation rate was 30.39%, which significantly exceeds the accusation rate in the *Baseline* treatment ($\chi^2 = 17.08$, $p < 0.001$). Overall accuracy in this treatment (56.47%) significantly exceeds chance level ($t = 2.94$, $p = 0.003$), and the accuracy of accusations increased even further to 60.65%, also significantly higher than chance level ($t = 2.70$, $p = 0.008$). We observed a strong asymmetry in the degree participants adopted the prediction of the AI, depending on when the AI prediction says “lie” or “truth”. When the AI predicted that the statement is true, 86.96% of participants adopted this prediction, yet when the AI predicted a lie, only 40.18% of participants adopted this prediction ($\chi^2 = 105.02$, $p < 0.001$).

As a consequence, accusation rates significantly differ too: when the algorithm predicted “truth”, the accusation rate was merely 13.04%, while when it predicted “lie”, it ballooned to 40.18% ($\chi^2 = 40.95$ with $p < 0.001$). This finding suggests that when by default, people receive recommendations from a lie-detection algorithm, they tend to follow them more when such compliance does not require accusation. Our last treatment provides an

opportunity to see whether this asymmetry holds for participants who actively request a prediction from AI instead of passively receiving it.

In the *Choice* treatment, where participants received a prediction from AI only if they actively requested and purchased it, around 31.37% of participants made that choice, which replicates the 32.16% uptake observed in the *Blocked* treatment ($\chi^2 = 0.07$ with $p = 0.79$). The overall accusation rate was 31.76%. This accusation rate was significantly higher than in the *Baseline* treatment ($\chi^2 = 21.14$, $p < 0.001$) but not higher than in the *Forced* treatment ($\chi^2 = 0.22$, $p = 0.64$). The overall accuracy in this treatment and the accuracy of accusations was not different from chance level (overall accuracy = 50.78%, $t = 0.35$, $p = 0.72$; accuracy of accusations = 51.23%, $t = 0.31$, $p = 0.75$).

Requesting and receiving the AI prediction affected accusation rates. Namely, the accusation rate among those who did not request AI predictions was 19.71%. This accusation rate does not differ from the accusation rates in the *Baseline* treatment ($\chi^2 = 0.03$, $p = 0.86$) or among those who did not request AI predictions in the *Blocked* treatment ($\chi^2 = 0.13$, $p = 0.72$). However, when participants requested and received an AI prediction, their accusation rate significantly increased to 58.13% ($\chi^2 = 74.74$, $p < .001$). In contrast to the *Forced* treatment, participants in the *Choice* treatment did not display a significant asymmetry in compliance with the different types of predictions of AI. Namely, when the AI predicted the statement was true, 94.44% of participants adopted that prediction; and 84.91% of participants adopted the prediction of the AI when it predicted a lie ($\chi^2 = 3.11$, $p = 0.08$).

As a consequence, accusation rates are strongly shaped by the content of the prediction. When the lie-detection algorithm says “truth”, accusations rates drop to 5.56%, while after AI predictions of “lie” accusation rates shoot up to 84.91% ($\chi^2 = 92.55$, $p <$

0.001). It shows that people are even more willing to follow AI predictions when they had previously requested them (88.13%), compared to when they just received them as in the *Forced* treatment (57.06%) ($\chi^2 = 51.32$ with $p < 0.001$). In particular, when the lie-detection predicts a statement to be untruthful in the *Choice* treatment, it has a strong effect on people's decisions. Namely, lying accusations among those who receive "lie" predictions in the *Choice* treatment (84.91%) are significantly higher than in the *Forced* treatment (40.18%, $\chi^2 = 64.03$, $p < 0.001$).

These results thus provide additional perspective on the results in the *Forced* treatment. In the *Forced* treatment, we observed a 30.39% accusation rate overall and a 40.18% accusation rate when the AI predicted a lie. Assume that about 30% of participants would have requested a prediction from AI (cf. results in the *Forced* and *Choice* treatment) and that these participants make about 60% accusations overall, 85% when the AI predicts a lie (cf. results in the *Choice* treatment). With these numbers, the results of the *Forced* treatments are in line with the interpretation that the 70% of participants who received a prediction but would not have asked for it disregarded it completely and accused at the baseline rate of 20% regardless of what the AI predicted. These findings show that *choosing* and *receiving* AI predictions substantially increases accusations of lying.

Predictors of Algorithmic Uptake. Across both treatments, we observe AI uptake levels of around 32%. Arguably such low uptake levels undermine the effect that algorithms have on social interactions. We analyze several predictors of uptake to find out which correlates with the decision to use lie-detection algorithms and to anticipate the magnitude of the social changes they might provoke.

At the end of the experiment, we asked participants five exploratory questions to gain a better understanding of who decides to request a prediction from AI. Two questions

asked for a subjective estimation of the absolute performance of the AI, namely, its accuracy (from 0% to 100%) and the probability that its accusations are wrong (from 0% to 100%). Two questions asked for a subjective estimation of the comparative performance of the AI: whether it would be better than the average human (on a scale from -5, Human's performance is better, to +5, Algorithm's performance is better), and whether it would be better than the participant himself or herself (on a scale from -5, My performance is better, to +5, Algorithm's performance is better). Finally, we asked how much guilt the participant would feel about making a wrong accusation (on a scale from -5, not guilty, to +5, very guilty).

Table 1 displays the results of logistic regression models predicting the frequency at which participants elect to use the algorithm as a function of their four judgments about its expected performance and feelings of guilt (restricted to the *Blocked* and *Choice* treatments). Because the responses to the four performance questions were all correlated (in the 0.20 to 0.65 range), we do not include them simultaneously in a single regression model. First, we see no significant link between the belief about the algorithm's general predictive accuracy and guilt. Second, the other three measures all significantly correlate with algorithmic uptake in the intuitive direction. Namely, people are more willing to request AI predictions when they believe (a) it outperforms an average human, (b) it outperforms themselves, and (c) the probability of false accusations is low.

To assess their economic significance, we calculate the marginal effect at the mean when the respective belief increases by one standard deviation. For the belief in the algorithm's performance relative to the average human (SD = 2.30), the probability increases by 6.71pp; for the belief in the algorithm's performance relative to oneself (SD = 2.42), the probability increases by 4.81pp; and for the belief on the algorithm's false accusation rate

(*SD* = 20.70), the probability of using the algorithm decreases by 3.24pp. Taken together, we find that beliefs about the relative performance of the lie-detection algorithm and of its error rate, not the beliefs about the algorithm's performance in general or subjective feelings of guilt, predicted adoption rates.

Table 1 | Predictors of Algorithm Usage. Logistic Regression of the frequency of requesting a hint in the *Blocked* and *Choice* treatments on the beliefs about the general accuracy of the algorithm (in %), its false accusation rate (in %), its performance compared to an average human (from -5, the human is better, to +5, the algorithm is better), its performance compared to the participant (from -5, oneself the algorithm is better, to +5, the algorithm oneself is better), and feelings of guilt when accusing somebody of lying in the study (from -5, not guilty, to +5, very guilty). Standard errors are reported in parentheses. Significance coding: * $p < .05$, ** $p < .01$, *** $p < .001$

	(1)	(2)	(3)	(4)	(5)
Belief Accuracy	-0.0022 (0.0034)				
Belief False Accusation		-0.0073* (0.0033)			
Belief Average vs. Algo			0.1318*** (0.0303)		
Belief Own vs. Algo				0.0906** (0.0284)	
Guilt					-0.0054 (0.0204)
Constant	-0.6329** (0.2174)	-0.4453** (0.1575)	-0.7924** (0.0687)	-0.7873** (0.0683)	-0.7642** (0.0673)
N	1020	1020	1020	1020	1020
Log-Likelihood	-637.38	-635.11	-627.82	-632.39	-637.55

Simulating the Future of AI-based Lie Detection. Our experimental design allowed us to estimate the proportion of participants who elected to use AI, the behavior of participants who did not use the AI and the degree to which participants followed the AI prediction when they obtained one. However, we must be careful when reporting population-level results, such as aggregated accusation rates and aggregated accuracy rates. For example, we reported that the accusation rate in the *Choice* treatment was 32%, but this global accusation rate is partly driven by the specific behavior of the algorithm we used in the experiment (in particular, the likelihood that it classifies a statement as a lie). The same logic applies to the global accuracy of participants' decisions, which is partly driven by the performance of the specific algorithm we used in the experiment.

In Figure 3 (left Panel), we simulate what the global accusation rate would have been if we had used other versions of the algorithm (In Figure S1, we adopt the same strategy to simulate what the aggregate accuracy of participants would have been, had we used different versions of the algorithm). These other versions vary along two dimensions, the probability that they classify a true statement as a lie and the probability that they classify a lie as true. For each version, we can simulate the aggregate accusation rate we would have observed in the experiment based on our estimation of the proportion of participants who elected to use AI, the behavior of participants who did not use the AI and the degree to which participants followed the AI prediction when they obtained one. As shown by this simulation, the specific behavior of the algorithm does not have a large impact on accusation rates, mostly because a 30% uptake is not enough to lead to large-scale changes at the population level. Higher levels of uptake (shown in Figure 3, middle and left panels) lead to more dramatic changes and greater variation in population-level outcomes as a function of the performance of the algorithm. Hence, higher uptake levels will drastically increase the downstream social effects of lie-detection algorithms. For instance,

with the current accuracy levels, overall accusation rates will increase to more than 50% when uptake reaches 90%.

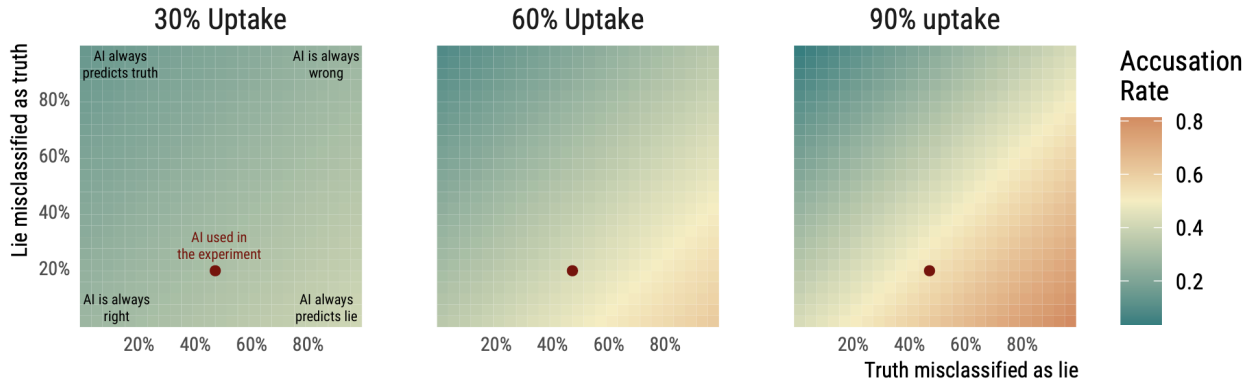


Figure 3 | Simulated accusation rates for varying levels of algorithm performance and uptake. The left panel shows simulations of false positive and false negative rates with different accusation rates for an uptake level of 30% (observed in the current study), the middle panel for an uptake level of 60%, and the left panel for an uptake level of 90%. For accusation rates, we use the color coding of green indicating lower, while red indicating higher accusation rates.

Discussion

Today, we live in a world where false accusations of lying come with costs, both to the accuser and the accused. One important feature of this current social equilibrium is that people generally refrain from accusing others of lying. We are interested in pre-emptively understanding the potential impact of novel AI-driven lie-detection technologies on this social equilibrium.

Specifically, we hypothesized that technologies that lower the probability or the cost of false accusations might disrupt the current social equilibrium and increase the rate at which people accuse each other of lying. We introduce a version of such future technologies into an incentivized experiment, one where AI significantly outperforms humans in detecting lies in written statements. To understand how such technology affects

the social dynamics of lie-detection and accusations, we manipulated the choice and availability of AI predictions.

The aggregate results across treatments suggest the following conclusions: (1) In the absence of AI, people are reluctant to make accusations; (2) When AI is available, a minority of people want to obtain its prediction; (3) The minority that does almost always goes for the AI prediction, even if it means to make an accusation; (4) People who request AI predictions are not endogenously more likely to accuse, as out treatment in which they request but do not obtain AI predictions suggest; (5) Tose who would not actively request the AI prediction do not change their behavior when they passively receive one; (6) Beliefs about the relative performance of the lie-detection algorithm predicted adoption rates. We discuss each of these findings before turning to the results of our simulations that sketch a potential way forward.

We find that people are overall reluctant to accuse others of lying, especially when no lie-detection algorithms are available. This finding replicates commonly observed findings in the lie-detection literature, documenting that people typically refrain from accusing others of lying (Gilbert 1991, Levine et al. 1999). One potential reason is that they are simply not very good at it and want to reduce the risks of paying the costs of false accusations for themselves and the accused. Supporting this notion, also in our study, people did not succeed at reliably discerning true from false statements.

The machine-learning algorithm we trained, however, did manage to exceed chance levels at this task. Intuitively, it would thus make sense for people to use it. However, we find that only approximately one-third of the participants decided to do so. This reluctance to use AI predictions, even when they can improve human decisions, is in line with a rich literature on algorithm aversion that has revealed such reluctance across various domains

(review: Burton et al. 2020). We add to this literature by documenting algorithm aversion in the context of lie-detection.

Moreover, we find an indication that people who are less averse to AI predictions are not more likely to accuse others of lying. Namely, accusation rates in the *Blocked* treatment did not differ between those who requested but did not receive AI predictions and those who did not request AI predictions. We thus find no evidence that potential AI adopters are endogenously more likely to make accusations. It appears instead that the availability of predictions shifts people's willingness to accuse others of lying.

Namely, if a lie-detection algorithm provides a prediction per default, people incorporate this prediction into their judgment, albeit less strongly than when they request the prediction. Interestingly, the *Forced* provision of the algorithm is the only treatment in which the overall accuracy levels succeed at chance levels, again underlining that algorithm aversion in the *Choice* treatment reduces the social effects of lie-detection AI. While mandatory AI predictions increase overall accuracy, requested AI predictions have a stronger influence on people's willingness to follow its prediction.

This downstream effect of requested AI predictions becomes particularly apparent for the accusation rate. When people sought AI predictions, and the algorithm flagged a statement as a lie, accusation rates climbed to almost 85%. One plausible explanation is that a lie-detection algorithm available offers the opportunity to transfer the accountability for accusations from oneself to the AI system (Hohenstein and Jung 2020, Köbis et al. 2021). However, when participants can use an algorithm for lie-detection, they only rely on its recommendations when they believe it makes accurate predictions. This finding suggests that in a morally controversial domain such as lie-detection, algorithmic uptake is not purely driven by blame-shifting motives but also by the desire to rely on algorithmic

support to make more accurate and fair judgments. Delegating a decision involving as much as calling someone a liar to an algorithm without a secure fact-checking process at least invokes considerations about the predictive power and reliability of such systems.

The path forward

Taken together, lie detection algorithms could have a strong disruptive potential for our current social equilibrium. Yet, in our experiment, low uptake of the algorithm weakened the disruptive impact — only 30% of participants elected to use the algorithm. But if algorithm uptake increases, our society might undergo significant transformations, for better or worse. Indeed, our simulations show that with higher algorithm uptake, the lie accusation rates, but also the overall accuracy, would increase, potentially leading to a lasting shift from the current social equilibrium of people's widespread reluctance to accuse others of lying. How this shift would play out is unclear at this stage.

One possibility is that high accusation rates may strain our social fabric by fostering generalized distrust and further increasing polarization between groups that already find it difficult to trust one another. However, making accusations easier, especially if these accusations are reasonably accurate, may also lead to beneficial effects by discouraging insincerity and promoting truthfulness in personal and organizational communications. Accuracy is an important factor here: we know that individuals can easily get false confidence in their ability to detect lies. Such is the case when they are exposed to pseudo-scientific methods of spotting liars, such as after learning the techniques of the TV show "Lie to me" (Levine et al. 2010). An advantage of lie-detection algorithms is that they can be properly tested and certified for above-human accuracy in a specific domain (Guszcza et al. 2018).

Limitations

Estimating the positive and negative social effects of lie-detection algorithms is not easy in a lab experiment since these effects may unfold slowly, in a cumulative manner, over a long period. Lab experiments are not the best tool for estimating these long-term cumulative effects, which is one limitation of our current work. But even if we cannot fully assess the magnitude and probability of these social changes, it seems reasonable to accept that to maintain a positive balance between benefits and costs, we will need to be mindful of the performance of lie-detection algorithms before making them massively available, and to use them responsibly as individuals and organizations, taking into account their limitations.

Our findings provide at least an encouraging signal in that direction: algorithmic uptake depends on the perceived accuracy of algorithms. This finding suggests that individuals may be mindful of the performance of lie-detection algorithms and use them somewhat responsibly to make accusations.

Organizations, on the other hand, may not always be so careful. Some managerial domains, such as negotiations with suppliers or clients, might be early adopters of lie-detection algorithms and pressure other domains, such as human resources, to do the same. Since suspicion about out-groups may be more socially acceptable than suspicion within the in-group, using lie-detection algorithms when dealing with other organizations may pave the way for their use within an organization. Behavioral science has a crucial role to play in anticipating these dynamics and carefully managing the transition to a high-accusation social world.

References

van den Assem MJ, van Dolder D, Thaler RH (2012) Split or steal? Cooperative behavior when the stakes are large. *Manage. Sci.* 58(1):2–20.

Burton JW, Stein M, Jensen TB (2020) A systematic review of algorithm aversion in augmented decision making. *J. Behav. Decis. Mak.* 33(2):220–239.

Castelo N, Bos MW, Lehmann DR (2019) Task-Dependent Algorithm Aversion. *J. Mark. Res.* 56(5):809–825.

DePaulo BM, Kashy DA, Kirkendol SE, Wyer MM, Epstein JA (1996) Lying in everyday life. *J. Pers. Soc. Psychol.* 70(5):979–995.

Devlin J, Chang MW, Lee K, Toutanova K (2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv [cs.CL]*.

Dietvorst BJ, Simmons JP, Massey C (2018) Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. *Manage. Sci.* 64(3):1155–1170.

Erat S, Gneezy U (2012) White Lies. *Manage. Sci.* 58(4):723–733.

Gaspar JP, Schweitzer ME (2013) The emotion deception model: A review of deception in negotiation and the role of emotion in deception. *Negot. Confl. Manag. Res.* 6(3):160–179.

Gilbert DT (1991) How mental systems believe. *Am. Psychol.* 46(2):107–119.

Gneezy U (2005) Deception: The role of consequences. *Am. Econ. Rev.* 95(1):384–394.

Guszcza J, Rahwan I, Bible W, Cebrian M, Katyal V (2018) Why We Need to Audit Algorithms. *Harvard Business Review* (November 28) https://hbr.org/2018/11/why-we-need-to-audit-algorithms.

Hartwig M, Bond CF (2011) Why do lie-catchers fail? A lens model meta-analysis of human lie judgments. *Psychol. Bull.* 137(4):643–659.

Hauch V, Sporer SL, Michael SW, Meissner CA (2016) Does Training Improve the Detection of Deception? A Meta-Analysis. *Communic. Res.* 43(3):283–343.

Hohenstein J, Jung M (2020) AI as a moral crumple zone: The effects of AI-mediated communication on attribution and trust. *Comput. Human Behav.* 106:106190.

Kleinberg B, Verschuere B (2021) How humans impair automated deception detection performance. *Acta Psychol.* 213:103250.

Köbis NC, Bonnefon JF, Rahwan I (2021) Bad machines corrupt good morals. *Nat Hum*

Behav 5(6):679–685.

Köbis NC, Starke C, Rahwan I (2022) The promise and perils of using artificial intelligence to fight corruption. *Nat Mach Intell* 4:418–424.

Leib M, Köbis NC, Soraperra I, Weisel O, Shalvi S (2021) Collaborative dishonesty: A meta-analytic review. *Psychol. Bull.* 147(12):1241–1268.

Levine TR, Park HS, McCornack SA (1999) Accuracy in detecting truths and lies: Documenting the "veracity effect." *Commun. Monogr.* 66(2):125–144.

Levine TR, Serota KB, Shulman HC (2010) The impact of Lie to Me on viewers' actual ability to detect deception. *Communic. Res.* 37(6):847–856.

Nahari G, Vrij A, Fisher RP (2014) Exploiting liars' verbal strategies by examining the verifiability of details. *Legal Criminol. Psychol.* 19(2):227–239.

Pascual-Ezama D, Muñoz A, Prelec D (2021) Do Not Tell Me More; You Are Honest: A Preconceived Honesty Bias. *Front. Psychol.* 12:693942.

Pascual-Ezama D, Prelec D, Muñoz A, Gil-Gómez de Liaño B (2020) Cheaters, Liars, or Both? A New Classification of Dishonesty Profiles. *Psychol. Sci.* 31(9):1097–1106.

Pérez-Rosas V, Kleinberg B, Lefevre A, Mihalcea R (2017) Automatic Detection of Fake News. *arXiv [cs.CL]*.

Saxe L, Dougherty D, Cross T (1985) The validity of polygraph testing: Scientific analysis and public controversy. *Am. Psychol.* 40(3):355–366.

Serota KB, Levine TR, Boster FJ (2010) The Prevalence of Lying in America: Three Studies of Self-Reported Lies. *Hum. Commun. Res.* 36(1):2–25.

Tergiman C, Villeval MC (2022) The Way People Lie in Markets: Detectable Vs. Deniable Lies. *Manage. Sci.*

Turmunkh U, van den Assem MJ, van Dolder D (2019) Malleable Lies: Communication and Cooperation in a High Stakes TV Game Show. *Manage. Sci.* 65(10):4795–4812.

Verschuere B, Köbis NC, Bereby-Meyer Y, Rand D, Shalvi S (2018) Taxing the Brain to Uncover Lying? Meta-analyzing the Effect of Imposing Cognitive Load on the Reaction-Time Costs of Lying. *J. Appl. Res. Mem. Cogn.* 7(3):462–469.

Verschuere B, Lin CC, Huismann S, Kleinberg B, Willemse M, Mei ECJ, van Goor T, Löwy LHS, Appiah OK, Meijer E (2023) The use-the-best heuristic facilitates deception detection. *Nat Hum Behav* 7(5):718–728.

Warren DE, Schweitzer ME (2018) When Lying Does Not Pay: How Experts Detect Insurance Fraud. *J. Bus. Ethics* 150(3):711–726.

Supplementary Material

Inclusion Criteria for Statements. The research assistants checked whether the authors wrote a meaningful statement about their activities (or, for the false statements, as if they were going to carry it out) as intended. For the truthful statements, they further verified that the additional question asking for supportive information fitted and reinforced the participant's entry. The third criterion was automatically applied and flagged all statements with less than 150 characters. If at least one statement of the authors failed at least one verification, we took out this author completely and did not use any of his/her statements for Part 2.

Additional Results on Algorithmic Performance. We illustrate the performance of the lie-detection algorithm used in this task for truthful and untruthful statements with a confusion matrix with the absolute numbers (Table S1A) as well as relative frequencies (Table S1B).

Table S1A | Confusion matrix in absolute numbers:

	Statement is untruthful	Statement is truthful
Prediction = untruthful	206	120
Prediction = truthful	49	135

Table S1B | Confusion matrix in relative frequencies:

	Statement is untruthful	Statement is truthful
Prediction = untruthful	40.39%	23.53%
Prediction = truthful	9.61%	26.47%

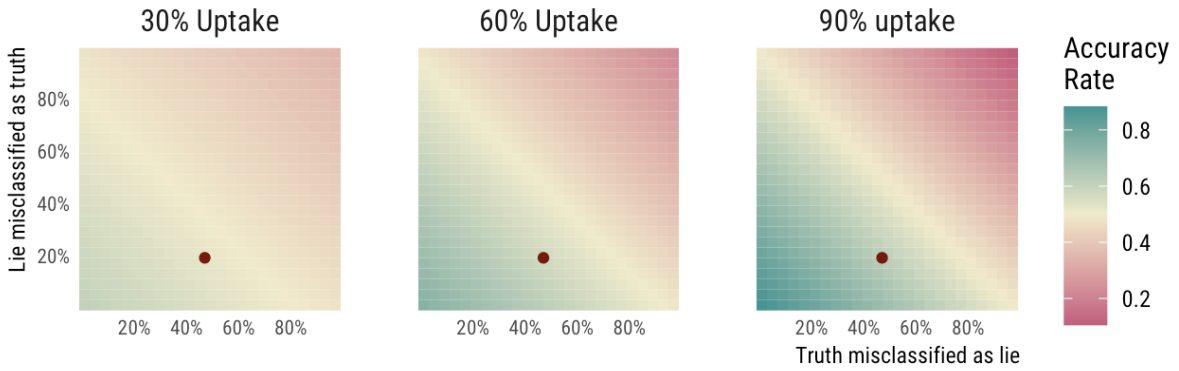


Figure S1 | Simulated accuracy rates for varying levels of algorithm performance and use

Experimental Instructions - Authors

Welcome

Welcome to this online experiment and thank you for your participation!

We will now give you detailed instructions. **Please read them carefully.**

You receive a base payment of £1.00 that will be paid out today.
In addition, you can earn extra money in the experiment depending on your choices.
You will receive this additional payment at a later date.

Introduction

Overview

In this experiment, you will write a **short statement about your most significant non-work-related activity** in the next seven days.
This statement must be **truthful** and must describe an activity you truly will carry out.

We will show your statement to a future participant.
This participant will **guess whether it is truthful or false.**
This future participant receives a financial bonus for correctly classifying lies as lies and truthful statements as truthful.

Payment

For your participation in this study, you receive a base payment of £1.00 that will be paid out today.
On top of that you can earn an additional payment of up to £4.00.
The additional payment depends on the judgment of a future participant.

Namely, if a future participant believes your statement to be true, you earn an additional £2.00.
If a future participant believes your statement to be false, you earn an additional £2.00 only with 50% probability.

You will receive the additional payment (if any) after the future participant has taken part in the study.

Your Activity

Please describe your most significant, non-work-related activity in the next seven days.
Your statement must be **truthful** and must describe an activity you truly will carry out.
This activity should be specific, have a clear start and an end time, and not be a continuous or daily activity.
Do not reveal any personally identifiable information such as names or addresses.
Furthermore, do not include any potentially harmful or illegal content (incl., hate speech,

etc.).
Your statement should be about 300-400 characters.

Name for your activity:

Please describe your activity as specific as possible:

Which information can you give us to reassure us that you are telling the truth? (1-2 sentences)

The experimenters will check whether you have written a meaningful, formulated text.
If not, you will not receive any payment.

Activities of Other Participants

We further would like to know which activities you will NOT carry out.
Below you find three activities of other participants of this study.
Please indicate which of these activities do not apply to you and you will not carry out.

⚪ [Other Activity 1]

⚪ [Other Activity 2]

⚪ [Other Activity 3]

Activities of Other Participants

Second Statement

You indicated that you will not carry out the activity "[Other Activity 1]".
For this activity, we ask you to make another statement.
This other statement must be false.
Please pretend that this would be your most important activity for the next week.
Provide a made-up statement similar to the one about your truthfully reported activity.

Payment

For this statement you can earn £4.00, as well.
The additional payment again depends on the judgments of a future participant.

Namely, if a future participant believes your statement to be true, you earn an additional £2.00.
If a future participant detects your statement as a lie, you receive no additional payment.

You will receive the additional payment (if any) after the future participant has taken part in the study.

Your Made-up Statement

You indicated that you will not carry out the activity "[Other Activity 1]".
Please pretend that this would be your most important activity for the next week.
Provide a made-up statement similar to the one about your truthfully reported activity.
This statement should be about 300-400 characters.

Name for the activity:
[Other Activity 1]

Made-up description of your activity:

The experimenters will check whether you have written a meaningful, formulated text.
If not, you will not receive any payment.

Lie Detection Algorithm *[shown to half of all authors]*

Recall that future participants will judge the truthfulness of your statements.
For each statement a future participant believes to be true, you earn an additional £2.00.
The future participants will be able to use a state-of-the-art **artificially intelligent lie detection algorithm** for their judgments.
This algorithm can analyze text and make predictions about the truthfulness of the content.
The future participants can pay a small fee to obtain an algorithmic prediction of the truthfulness of your statement.
You can prevent this use by paying a small fee of £0.30.
This will block the application of the lie detection algorithm to *both* of your statements.

Do you want to prevent the use of the lie detection algorithm for your statements?

⚪ Yes
⚪ No

Results

The main part of the experiment is now finished.
We will transfer your base payment of £1.00 as soon as possible.

You receive an additional £2.00 for each of your statements that a future participant believes to be true.
We will pay out the additional payment (if any) after the future participants have taken part in the study.
Note that this may take some time.

Assessments

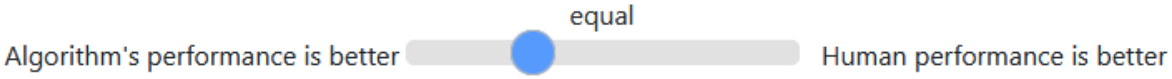
Besides the judgment of future participants, a state-of-the-art intelligent algorithm was designed to predict truthfulness.
This algorithm can analyze text and make predictions about the truthfulness of the content.
Before the end of the experiment, we would like to ask for your opinion in some short questions.
These questions concern the performance of this algorithm.

Question 1

How frequently do you think the state-of-the-art intelligent algorithm **correctly predicts** whether a statement is true or false?

Question 2

How good do you think the **average human performance** is compared to the performance of the intelligent lie detection algorithm in predicting whether a statement is true or false?

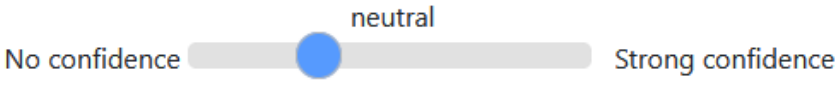


Question 3

How frequently do you think the intelligent algorithm **incorrectly predicts a lie** although it is actually a true statement?

Question 4

How much confidence do you have in your assessment of the performance of the algorithm?



Survey

Please fill out this final survey before finishing the experiment.

How old are you?

What is your gender?

⚪ female
⚪ male
⚪ other/non-binary

What is your highest educational degree?

⚪ No degree
⚪ High school
⚪ Bachelor
⚪ Master
⚪ PhD

If you go/went to university, what is/was your major?

⚪ Not applicable
⚪ Economics
⚪ Law
⚪ Psychology
⚪ Political sciences
⚪ Medicine
⚪ Natural sciences
⚪ Engineering
⚪ Other social sciences
⚪ Other

What is your employment status?

⚪ Unemployed
⚪ Part-time
⚪ Full-time

How familiar are you with new technologies such as machine learning?

⚪ Not familiar at all
⚪ Rather not familiar
⚪ Neutral
⚪ A little familiar
⚪ Very familiar

End of Experiment

The experiment is now finished.
Please click on the button below to return to Prolific.
You can only receive payment after being redirected to Prolific.
You will receive your payment as soon as possible.

Back to Prolific

Experimental Instructions - Judges

Welcome

Welcome to this online experiment and thank you for your participation!
We will now give you detailed instructions. **Please read them carefully.**
You receive a base payment of £1.20 that will be paid out today.
In addition, you can earn extra money in the experiment depending on your choices.
You will receive this additional payment at a later date.

Introduction

Overview

In this experiment, you will read a short statement about a non-work-related activity that is either truthful or a lie.
A past participant of this study made this statement.
We refer to this participant as *the author* in what follows.
Your task is to guess whether it is truthful or a lie.
Half of *all* statements are truthful and half of them are lies.
The statement we show you will be randomly selected.

Payment

For your participation, you receive a base payment of £1.20.
On top of that you receive a bonus for a correct guess.
That is, if you judge a truthful statement to be truthful, you receive £0.50.
Likewise, if you judge a lie to be a lie, you receive £0.50.
Otherwise, you receive no bonus.

Lie detection algorithm *[Lie-detection treatments]*

You can use a state-of-the-art **artificially intelligent lie detection algorithm** for your judgments.
This algorithm shows moderately better performance in distinguishing truth from lies than the average human.
You can pay a small fee of £0.05 to obtain an algorithmic prediction of the truthfulness of the statement.
For some statements, the prediction of the algorithm is not available.
In this case, you would not be charged the £0.05 for purchasing a prediction.

Consequences of lie accusations for the author of the statement

Whenever you accuse the author of the statement of lying, this author is punished.
In this case, this author loses £2.00 of his/her total achievable payoff!
This happens *regardless* of whether you are correct in accusing the author of lying or not.

Example 1 / Example 2

This screen shows an example of how your main tasks will look like.
The texts for the activity and description are just placeholders.
Here you will see the statement of the past participant you need to judge.
After reading the statement, you make your judgment by clicking on either Truth or Lie.
Go on and try one button then the other, this is an example so your choices have no consequences for now.

Name of the activity:
Lorem ipsum

Statement and description of the activity:
Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr.

[Lie-detection treatments]
Do you want to purchase a prediction of the state-of-the-art algorithm for this statement for £0.05?
For some statements, the prediction of the algorithm is not available.
In this case, you would not be charged the £0.05 for purchasing a prediction.

Purchase for £0.05

The algorithm predicted this statement to be ** *here it will show whether the algorithm predicted a truth or a lie* **.

What do you think:
Is this statement truthful or a lie?

Truth Lie

Judgment

Name of the activity:
[Activity 1]

Statement and description of the activity:
[Description of activity 1]

Do you want to purchase a prediction of the state-of-the-art algorithm for this statement for £0.05?
For some statements, the prediction of the algorithm is not available.
In this case, you would not be charged the £0.05 for purchasing a prediction.

Purchase for £0.05

The algorithm predicted this statement to be **a lie / truthful**.

What do you think:
Is this statement truthful or a lie?

Truth Lie

Results

[Treatments without lie-detection algorithm]
You rated the statement correctly.
You therefore earned £0.50 on top of your base payment of £1.20.

Your total payoff is thus **£1.70**.
We will transfer your total payoff as soon as possible.

[Lie-detection treatments, algorithm available]
You rated the statement correctly.
You therefore earned £0.50 on top of your base payment of £1.20.

You requested a prediction from the algorithm for £0.05.
This prediction was available.
This hint costs you £0.05.

Therefore, you receive a payoff of £0.50 – £0.05 = £0.45 on top of your base payment of £1.20.
Your total payoff is thus **£1.65**.
We will transfer your total payoff as soon as possible.

Assessments

[Treatments without lie-detection algorithm]
Besides your judgment of the statements, a state-of-the-art intelligent algorithm was designed to predict truthfulness.
This algorithm can analyze text and make predictions about the truthfulness of the content.
Before the end of the experiment, we would like to ask for your opinion in some short questions.
These questions concern the performance of this algorithm.

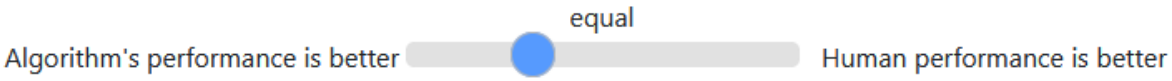
[Lie-detection treatments]
Before the end of the experiment, we would like to ask for your opinion in some short questions.

Question 1

How frequently do you think the state-of-the-art intelligent algorithm **correctly predicts** whether a statement is true or false?

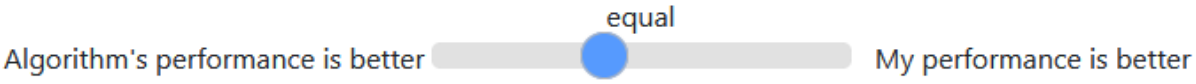
Question 2

How good do you think the **average human performance** is compared to the performance of the intelligent lie detection algorithm in predicting whether a statement is true or false?



Question 3

How good do you think **your performance** is compared to the performance of the intelligent lie detection algorithm in distinguishing truth from lies?

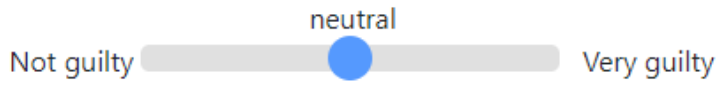


Question 4

How frequently do you think the intelligent algorithm **incorrectly predicts a lie** although it is actually a true statement?

Question 5

In this study, how guilty would you feel if authors lost some of their payment because you clicked on "Lie" while their statements were true ?



Survey

Please fill out this final survey before finishing the experiment.

How old are you?

What is your gender?

⚪ female
⚪ male
⚪ other/non-binary

What is your highest educational degree?

⚪ No degree
⚪ High school
⚪ Bachelor
⚪ Master
⚪ PhD

If you go/went to university, what is/was your major?

⚪ Not applicable
⚪ Economics
⚪ Law
⚪ Psychology
⚪ Political sciences
⚪ Medicine
⚪ Natural sciences
⚪ Engineering
⚪ Other social sciences
⚪ Other

What is your employment status?

⚪ Unemployed
⚪ Part-time
⚪ Full-time

How familiar are you with new technologies such as machine learning?

⚪ Not familiar at all
⚪ Rather not familiar
⚪ Neutral
⚪ A little familiar
⚪ Very familiar

End of Experiment

The experiment is now finished.
Please click on the button below to return to Prolific.
You can only receive payment after being redirected to Prolific.
You will receive your payment as soon as possible.

Back to Prolific