Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Improving Human Deception Detection Using Algorithmic Feedback

Marta Serra-Garcia; , Uri Gneezy

Published online in Articles in Advance 25 Apr 2025

. https://doi.org/10.1287/mnsc.2023.02792

informs
https://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–19
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Improving Human Deception Detection Using Algorithmic Feedback

Marta Serra-Garcia,[a,b,*] Uri Gneezy[a]

[a] Rady School of Management, University of California San Diego, La Jolla, California 92093; [b] CESifo, 81679 München, Germany
*Corresponding author
Contact: mserragarcia@ucsd.edu, https://orcid.org/0000-0002-5875-4986 (MS-G); ugneezy@ucsd.edu (UG)

Received: September 1, 2023
Revised: January 18, 2024; June 18, 2024
Accepted: August 7, 2024
Published Online in Articles in Advance: April 25, 2025

https://doi.org/10.1287/mnsc.2023.02792

Abstract. Can algorithms help people detect deception in high-stakes strategic interactions? Participants watching the preplay communication of contestants in the TV show *Golden Balls* display a limited ability to predict contestants' behavior, whereas algorithms do significantly better. To increase participants' accuracy, we provide them with algorithmic advice by flagging videos for which an algorithm predicts a high likelihood of cooperation or defection. We test how the effectiveness of flags depends on their timing. We show that participants rely significantly more on flags shown *before* they watch the videos than flags shown *after* they watch them. These findings show that the timing of algorithmic feedback is key for its adoption.

History: Accepted by Marie-Claire Villeval, behavioral economics and decision analysis.
Funding: Funding provided by an Innovation Grant for Inclusive Research Excellence at UC San Diego.
Supplemental Material: The online appendix and data files are available at https://doi.org/10.1287/mnsc.2023.02792.

Keywords: detecting lies • machine learning • cooperation • experiment

1. Introduction

Machine learning (ML) algorithms are often developed to assist people in making predictions. Do people use the advice algorithms provide? The existing empirical evidence provides contradictory answers to this question. Some studies find that individuals are likely to adopt algorithmic advice (e.g., Bundorf et al. 2019), whereas others find that they largely ignore it (e.g., Kim et al. 2024). Given the mixed results, understanding under what conditions people will use algorithmic advice is important in making such tools useful and in advancing the science of human-machine interactions.

We propose that the mixed results in the literature may be reconciled by considering the timing of the advice. Timing is an important choice in information design (Kamenica 2019) that can influence how algorithms are perceived and adopted. When individuals receive algorithmic advice *before* forming their own belief, they may use it significantly more than when they receive it *after* forming their initial belief. The reason is that when algorithmic advice is presented first, individuals can look for evidence that supports the advice and weigh it strongly when forming their own beliefs. By contrast, when people first form their own beliefs, they may compare the algorithmic advice against their own assessment, giving the algorithmic advice less weight.

The potential differential impact of algorithmic advice based on timing is motivated by confirmation bias, a tendency to actively search for and interpret information that matches their beliefs (Nickerson 1998). In psychology, the confirmation bias literature has demonstrated that people tend to select information that supports their views, often putting less weight on contradicting information. People also tend to interpret ambiguous evidence as supporting their existing beliefs (e.g., Mynatt et al. 1977, Baron 2000). In economics, the literature has focused on how the timing of signals affects updating, finding evidence of confirmation bias in some cases (e.g., Charness and Dave 2017), but not always (e.g., Eil and Rao 2011, Möbius et al. 2022; see Benjamin 2019 for a review).

To the best of our knowledge, the impact of timing for algorithmic advice has not been studied thus far, and the existing related literature does not provide evidence for a clear effect. The mixed evidence on confirmation bias could indicate that timing may not affect the adoption of algorithmic advice. However, our review of the literature on algorithmic advice suggests that timing could be important. We found that when algorithmic advice is presented early in the judgment process, several studies document that individuals tend to follow the algorithmic advice (as discussed in Section 2). By contrast, when individuals first form a judgment in their mind, we found several studies showing hesitancy to follow algorithmic advice. A better understanding of how timing impacts the adoption of algorithmic advice can therefore help explain the mixed patterns in algorithmic adoption in previous studies and provide a new design choice that can be leveraged to increase adoption.

To test how timing affects the adoption of algorithmic advice, we design an experiment in which participants need to predict whether people in video clips are being truthful or deceptive about their intentions. Predicting behavior when there is an incentive to mislead others is important in many situations. Consider a politician who posts a video online with campaign promises or a salesperson who promotes a product with a video featuring its qualities. Because the politician's and salesperson's interests might not be aligned with those of viewers, it can be difficult for viewers to know whether to trust such promises.

Research suggests that people display a limited ability to detect deception, and, when observing others' interactions, their predictions are often not much better than chance (see, e.g., Ockenfels and Selten 2000, Belot et al. 2012, Konrad et al. 2014, Belot and van de Ven 2017, Dwenger and Lohse 2019, Serra-Garcia and Gneezy 2021; for a meta-analysis, see Bond and DePaulo 2006). Machine learning algorithms often predict behavior more accurately using text (e.g., Perez-Rosas et al. 2015), audio (e.g., Chen et al. 2020), and visual features (Hu and Ma 2021) of their conversations. When potentially deceptive videos are posted online, the websites hosting them could leverage algorithms to flag suspicious content to help people form more accurate beliefs, either before or after they watch the videos. An open question is how much individuals' beliefs will respond to algorithmic advice and how timing may affect its effectiveness.

We study potentially deceptive videos from the high-stakes prisoner's dilemma game played within the TV show *Golden Balls*, with an average prize of over GBP 13,000 (or $26,000 in 2007 terms). In this show, contestants play a version of the prisoner's dilemma: They first have a brief conversation with each other and then simultaneously decide "steal" or "split." If both choose split, they share the prize. If one chooses split and the other chooses steal, the player who chooses steal wins the entire prize, and the other wins nothing. And if both choose steal, neither wins money.

Although contestants in the TV show elect steal 46% of the time, they almost always make nonbinding preplay statements in their conversations, declaring their intention to choose split. The challenge is to predict which contestants will nevertheless choose steal. Forming accurate beliefs in this setup is complicated by the heterogeneity in lying costs and preferences to cooperate (e.g., Gneezy 2005, Fischbacher and Föllmi-Heusi 2013, Abeler et al. 2019).[1]

In the first experiment, we show that participants display a limited ability to predict contestant behavior, and a simple ML algorithm provides a significantly higher accuracy than our participants. Hence, in our setting, algorithmic advice has the potential to improve participant predictions if it is adopted. A potentially important design choice is when the advice is presented. We leverage this design dimension to address our main research question: how can we provide individuals with algorithmic advice that they will use?

The algorithmic advice we study is based on flagging extreme predictions of the algorithm, informing participants that an algorithm predicted the person in the clip they are watching is very likely or very unlikely to choose steal (with more than a 70% chance). We chose to use feedback in the form of flagging for three main reasons. First, flagging is a relatively simple way of providing feedback, and it is easy to understand. Second, ML predictions are more reliable when the algorithm predicts steal with a very high or a very low probability. Third, such flags are easy to apply in practice, and similar real-world flagging procedures already exist.

We test whether the effectiveness of algorithmic advice depends on *when* the advice is presented to participants in the second experiment. Participants watch 20 video clips, and in the treatments with flagging, four of them are flagged. Two of the videos are flagged as contestants who are "very likely to steal" and two as "very likely to split." Participants see the flag prior to watching the video (*Flag-Before* treatment) or after watching the video (*Flag-After* treatment) or do not see flags (*Control*). If individuals are rational and unbiased, the predictions should not depend on the timing of the flags, as in both treatments with flags, participants face the same information: their assessment of the video and the ML flag.

The data show that the timing of flags significantly affects participant predictions. Flags shown *before* the participant watches the video lead to more than a five-fold increase in the difference in participants' beliefs between videos flagged as very high and those flagged as very low chance of choosing steal. Participants also exhibit a significant increase in the accuracy of their predictions. By contrast, when flags are shown *after* the participant watches the video, the effect on predictions is significantly weaker, and the impact on accuracy is not significant.

As we discussed above, confirmation bias would predict participants rely more on the first piece of information they receive, leading to a primacy effect. Consistent with this explanation, participants report overwhelmingly relying on their own beliefs when the flag is shown after they watch the video (in *Flag-After*). Though timing does not affect the perceived accuracy of the algorithm, participants' confidence in their own ability increases, leading them to believe they are as accurate or significantly more accurate than the algorithm. By contrast, when participants see the flag first (in *Flag-Before*), they are more likely to report that they trust the algorithm and combine its advice with their own beliefs.

The timing of flags may also affect participants' behavior if they wish to minimize the time spent in making predictions (e.g., Dykstra et al. 2022). We do not find evidence consistent with this behavior. The timing of advice does not affect the time spent on each video, and participants spend significantly more time watching the video than required, double the average length of contestant conversations.

Our findings provide an explanation for the contradictory results in the literature, suggesting that the timing of algorithmic advice could be an important factor to help explain when it will influence individual decision making. Understanding the importance of timing of algorithmic advice could help organizations improve the adoption of predictive modeling and algorithmic aids in a wide range of applications. Our findings indicate that the effectiveness of algorithmic advice will be lower when decision makers are already familiar with the situations and have formed an initial belief.

In the context of deception detection, real-world flagging procedures already exist. Some websites, including YouTube and TikTok, have algorithms that analyze videos and teams that check whether videos are in line with their guidelines. These websites could choose to flag potentially deceptive content before the video is viewed or wait until after viewers have watched a video. In some cases, delays may naturally occur, and flags may only appear after concerns have been raised by users, which is often referred to as "debunking" (e.g., Chan et al. 2017). Posting flags as early as possible is a choice websites can make to reduce the spread of deceptive content, and our results suggest that this choice regarding the timing of flags is crucially important.

2. Related Literature

A growing body of research has studied the adoption of algorithms in a variety of domains. There are many factors that affect individuals' willingness to adopt algorithmic advice (e.g., Dietvorst et al. 2018, Burton et al. 2020). Our paper focuses on the potential importance of the timing of algorithmic advice, and we hence briefly review the existing literature focusing on when advice was presented to its users: whether the advice was presented *before* or *after* individuals formed their own judgments. We focus on the papers reviewed in Chugunova and Sele (2022) and extend them with recent work on deception detection with the use of algorithmic advice, as summarized in Table 1.

In studies in which individuals are exposed to the ML advice *before* they form their own judgments, ML advice often has statistically and economically significant effects on their decisions (Dijkstra 1999, Promberger and Baron 2006, Bundorf et al. 2019, Wismüller and Stockmaster 2020, Allen and Choudhury 2022, Dell'Acqua 2022, Agarwal et al. 2023, Wang et al. 2023, Leib et al. 2024). For example, in the health domain, ML (or AI) affects the decisions of both individuals and professionals. In Bundorf et al. (2019), when algorithms were used early to determine how prescription drug insurance plans were ranked when presented to individuals, they significantly influenced their choices. When algorithms are used to aid professionals such as radiologists, they also affect their workflow. For example, Wismüller and Stockmaster (2020) measured the impact of algorithmic flags for head CT scans on turnaround time (TAT). They found that their early flagging procedure reduced TATs substantially and affected radiologists' decision making.

In technology-centered service jobs, such as coding information from medical charts, AI also affects individuals' work processes when it is presented early (e.g., Allen and Choudhury 2022, Wang et al. 2023), though the effects are moderated by worker experience. These heterogeneous effects are suggestive of experience working as an anchor against taking the algorithmic advice, consistent with the timing differences we study and document experimentally.

By contrast, when experts are exposed to ML advice *after* they have formed their own judgment, the extent of adoption is limited (Longoni et al. 2018, Stevenson and Doleac 2023, Kim et al. 2024). For example, in Kim et al. (2024), restaurant inspectors first decided on their own ranking of restaurants to inspect. After doing so, they received the algorithm-based ranking. The data show that inspectors largely ignored the algorithmic advice.

In the legal system, risk assessment algorithms are used by judges as part of the decision-making process, assisting them in predicting recidivism. Judges receive this risk assessment after conviction when the risk assessment and sentence guidelines worksheets are provided for sentencing, parole, and related decisions. Stevenson and Doleac (2023) find that judges changed sentencing practices in response to the risk assessment but that discretion played a large role in mediating its impact. As a result, risk assessment algorithms did not bring any detectable benefit in terms of public safety or reduced incarceration (see also Kleinberg et al. 2018). An important feature of these algorithmic assessments is that judges receive them when they are

Table 1. Literature Review: Timing Effects of Algorithmic Advice

	Paper	Domain	Timing of advice	Effect?	Effect description
1	Dijkstra (1999)	Legal	Before	Yes	79% of participants agreed with the incorrect opinion of an expert system.
2	Promberger and Baron (2006)	Medical	Before	Yes	Participants followed the computer recommendation to operate (mean = 0.35, on a scale of −1 to 1, where −1 indicated not to operate and 1 to operate), and its recommendation not to operate (mean = −0.10).
3	Bundorf et al. (2019)	Medical	Before	Yes	The switching probability of shoppers increased significantly (eight percentage points) when exposed to advice from algorithm.
4	Wismüller and Stockmaster (2020)	Medical	Before	Yes	Clinicians' prioritization of cases was significantly influenced by an algorithm that flagged case priority, reducing treatment time significantly by 59 minutes.
5	Allen and Choudhury (2022)	IT	Before	Yes	Participants with more years of IT experience were 16–20 percentage points more likely to accurately resolve tickets with AI recommendations relative to solving them manually.
6	Dell'Acqua (2022)	HR	Before	Yes	Recruiters followed the AI advice between 53% and 85% of the time, depending on its quality.
7	Agarwal et al. (2023)	Medical	Before	Yes	Radiologists follow AI assistance, reducing their deviation from AI by 16%–18% relative to when they are not exposed to it.
8	Wang et al. (2023)	Medical	Before	Yes	AI reduces coding time by 11% per medical chart.
9	Leib et al. (2024)	Deception	Before	Yes	AI advice that promotes dishonesty increases dishonesty, but AI advice that promotes honesty does not affect participants.
10	Longoni et al. (2018), study 9	Medical	After	No	Participants were resistant to using medical AI as the decision maker (mean = 4.88, where 1 is very unlikely and 7 very likely, relative to mean = 5.57 with human provider).
11	Kim et al. (2024)	Health	After	No	Inspectors were only two-thirds as likely to inspect restaurants based on the algorithms relative to their own rankings.
12	Stevenson and Doleac (2023)	Legal	After	No	The net effect of AI on both incarceration rates and sentence lengths was a precisely estimated null, and it did not affect racial disparities in sentencing either.
13	Kleinberg and Verschuere (2021)	Deception	After	No	Although the AI was significantly better than chance, participants deviated from recommendations, and their accuracy was not significantly different from chance.

Notes. This table summarizes existing studies with different timings of algorithmic advice. The column "Timing of advice" describes whether algorithmic advice was presented *before* or *after* individuals had made their own judgments, within the decision making process. The column "Effect?" summarizes whether the algorithmic advice affected decision making. The column "Effect description" describes the context in which the effect was measured and, if available, a measure of its size.

already familiar with the case. This timing might explain why the risk assessment tools have a small impact relative to the clinical examples above.

Within the context of deception detection, multiple studies have developed algorithms that are more accurate than humans (for a review, see Constancio et al. 2023, Prome et al. 2024). The literature shows that individuals' ability to detect deception is limited when presented with text-based (e.g., Kleinberg and Verschuere 2021), audio (Wu et al. 2015), or video-based information (e.g., Serra-Garcia and Gneezy 2021). An open question is how willing individuals are to take algorithmic advice in the context of deception detection.[2]

We argue that timing is important to the success of AI input in influencing behavior. For example, Kleinberg and Verschuere (2021) examined how ML advice influences deception detection in the context of text-based autobiographical statements. They presented the ML prediction after the individual read the text available. Consistent with findings in other domains, individuals frequently overruled the ML prediction. As a result, although the algorithm was significantly more accurate than individuals, human accuracy did not increase.[3]

3. The Setting: The *Golden Balls* TV Show

Golden Balls is a TV show that was broadcast in the United Kingdom from 2007 to 2009. The show consists of four rounds of play. In the first three rounds, contestants make claims about the cash value of golden balls they have been allocated, after which they discuss and vote against each other. The prize ("jackpot") in the last round depends on the balls drawn over play in these first three rounds.

In this paper, we focus on contestant behavior in the fourth and last round of the show, which is a high-stakes prisoner's dilemma with an average prize of over GBP 13,000. Two contestants simultaneously and privately choose split or steal. If they both choose split, they share the prize equally. If they both choose steal, they each receive zero. If one contestant chooses split while the other chooses steal, the former receives nothing, while the latter receives the entire prize. Therefore, choosing steal is a weakly dominant strategy. The payoff matrix is presented in Table 2.

Prior research has documented several interesting behavioral regularities in this gameshow. Fifty-four percent of contestants choose split (e.g., van den Assem et al. 2012). Women choose split more often than men, particularly young male contestants, and attractiveness increases cooperation in mixed-gender pairs (van den Assem et al. 2012, Darai and Gratz 2013). Similar results are found in a Dutch variation of the show by Belot et al. (2009, 2012) and in studies of the American TV show *Friend or Foe* (List 2006, Oberholzer-Gee et al. 2010).

Prior to making the split or steal decision, contestants engage in a brief conversation in which they discuss their intentions with each other. During this preplay communication, which lasts approximately 20 seconds, contestants typically talk about their intention to choose split or try to get assurances that the other contestant will choose split. Turmunkh et al. (2019) find that over 83% of conversations feature a statement involving the intention to choose split. Among these statements, when contestants tell malleable lies (statements that are malleable to ex post interpretation as truths), they are more likely to choose steal.

The sample of *Golden Balls* videos that we study consists of 430 contestants from four of the show's six seasons. These videos were shared for research purposes by Donja Darai, who obtained them directly from the show producer. Our analysis focuses on the final conversation prior to the split or steal decision. Participants in our experiments were shown these conversations without any edits and were asked to predict one of the contestants' decisions.

Table 2. Payoff Matrix in the Prisoner's Dilemma of *Golden Balls*

		Contestant B: Split	Contestant B: Steal
Contestant A	Split	50%, 50%	0%, 100%
	Steal	100%, 0%	0%, 0%

3.1. Preplay Communication

The behavior and conversation of a contestant prior to the cooperation decision can be captured by nonverbal as well as verbal features. By nonverbal features, we refer to facial movements and expressions, which can reflect emotions. By verbal features, we refer to what contestants said and how they said it.

People's choices may be linked to their emotions. For example, people who lie may feel fear or guilt and overall have fewer positive emotions than those who tell the truth (Ekman 2009). Facial expressions have been used recently in experimental games to measure how players strategically display emotions, for example, in the ultimatum game (e.g., van Leeuwen et al. 2018, Chen et al. 2019) or to test how their smiles relate to behavior in the trust game (e.g., Centorrino et al. 2015a, b). Serra-Garcia and Gneezy (2021) use simple probit models to relate facial expressions to truth-telling by experimental participants. In that study, participants were recorded in 30-second videos making either true or false statements. Several nonverbal features were associated with the sender's truthfulness. Hu and Ma (2021) use nonverbal and verbal features to estimate the positiveness in videos of startup pitches and relate these emotions to funding decisions.

3.1.1. Nonverbal Features. We use the software FaceReader to analyze the facial expressions of each contestant in our sample of *Golden Balls* videos during their conversation prior to the split or steal decision. FaceReader is a facial analysis software that measures, over time, the six basic ("universal") emotions described by Ekman (1970): happy, sad, angry, surprised, scared, disgusted, as well as neutral (Bijlstra and Dotsch 2011). The software also analyzes arousal, which measures the level of activity on the face. Values are between zero and one.

FaceReader additionally reports several facial movements. In each frame, it measures whether the contestant's mouth and eyes are open, the position of eyebrows, and the direction of gaze. The software also measures the orientation of the head along three axes (in degrees) and provides a measure of the quality of the video, which is between zero and one. Figure 1 shows an example of the software analyzing a participant in the TV show.

Our main algorithm uses the average strength of each contestant's emotions and arousal and the average rate of facial movements during their conversation prior to the split of steal decision. Additional analyses using the standard deviation, minimum, and maximum of each feature during the conversation do not increase predictive accuracy (and are described in Online Appendix C). To facilitate FaceReader's analysis, these videos were edited to remove all shots that did not display the contestants. The removed shots included shots of the audience and shots including the host. Details are provided in Online Appendix B.

3.1.2. Verbal Features. Contestants' verbal behavior, measured by speech (what they say) and voice (how they say it), could provide cues on their final choice between split and steal. We include several aspects of speech. First, we focus on two simple features: word count and sentiment score, using the transcripts in Turmunkh et al. (2019). Past research shows these easy-to-interpret features correlate with lying (Serra-Garcia and Gneezy 2021). Second, we also include whether a contestant makes explicit or implicit promises, based on the classification provided in the data set of Turmunkh et al. (2019).

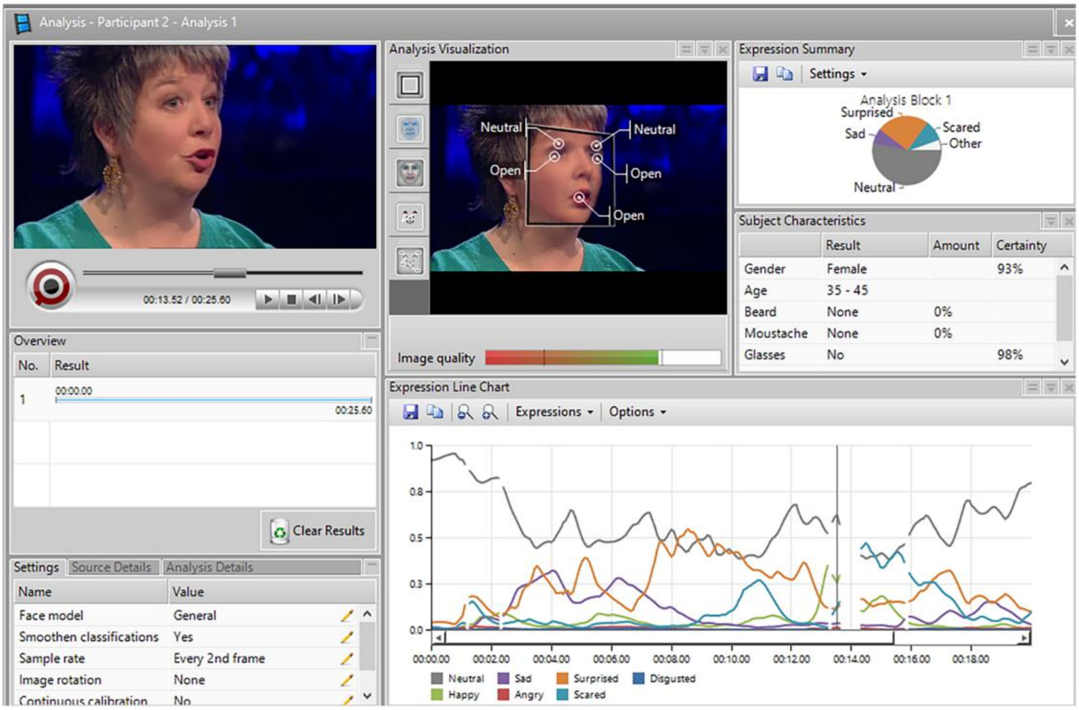
For voice, we include two simple features that describe the contestants' voices: intensity and pitch. The voices of each contestant were obtained by sequentially muting their opponent's voice in the preplay conversation. The resulting voices were analyzed using Praat by Boersma and Weenink (2020), a standard phonetics software that analyzes sound waves. The intensity of a sound is the power per unit area carried by the wave and is an approximate measure of the loudness of a contestant's voice. Pitch captures how high or low a sound is. It is defined as the fundamental frequency of each sound wave and is measured in hertz. A detailed description of the meaning of these features is included in Online Appendix B.

3.1.3. Additional Features. In addition to verbal and nonverbal features of the conversation, we include demographic characteristics and game information as potential predictors of behavior (e.g., van den Assem et al. 2012). We include the participants' age and gender and the prize at stake, as coded in Turmunkh et al. (2019). For the sample of videos we study, the average age of contestants was 36 years, and 54.0% were women. The average prize was GBP 13,445, and 46% of contestants chose steal.[4]

3.2. Descriptive Statistics

FaceReader is best able to analyze facial expressions on straight-ahead faces with proper lighting. As such, the *Golden Balls* videos are not the optimal settings in which FaceReader can be run. For several participants, some frames could not be analyzed by the software, and in some cases, the software captured no frames at all. Our sample for analysis focuses on all the contestants for whom FaceReader analyses of emotional states could be conducted for at least one frame, resulting in 430 contestants across 217 episodes. On average, 56.3% of the frames for each of these contestants could be read, and the emotions were analyzed by FaceReader. We find this feature important in

Figure 1. (Color online) Example of FaceReader Analysis

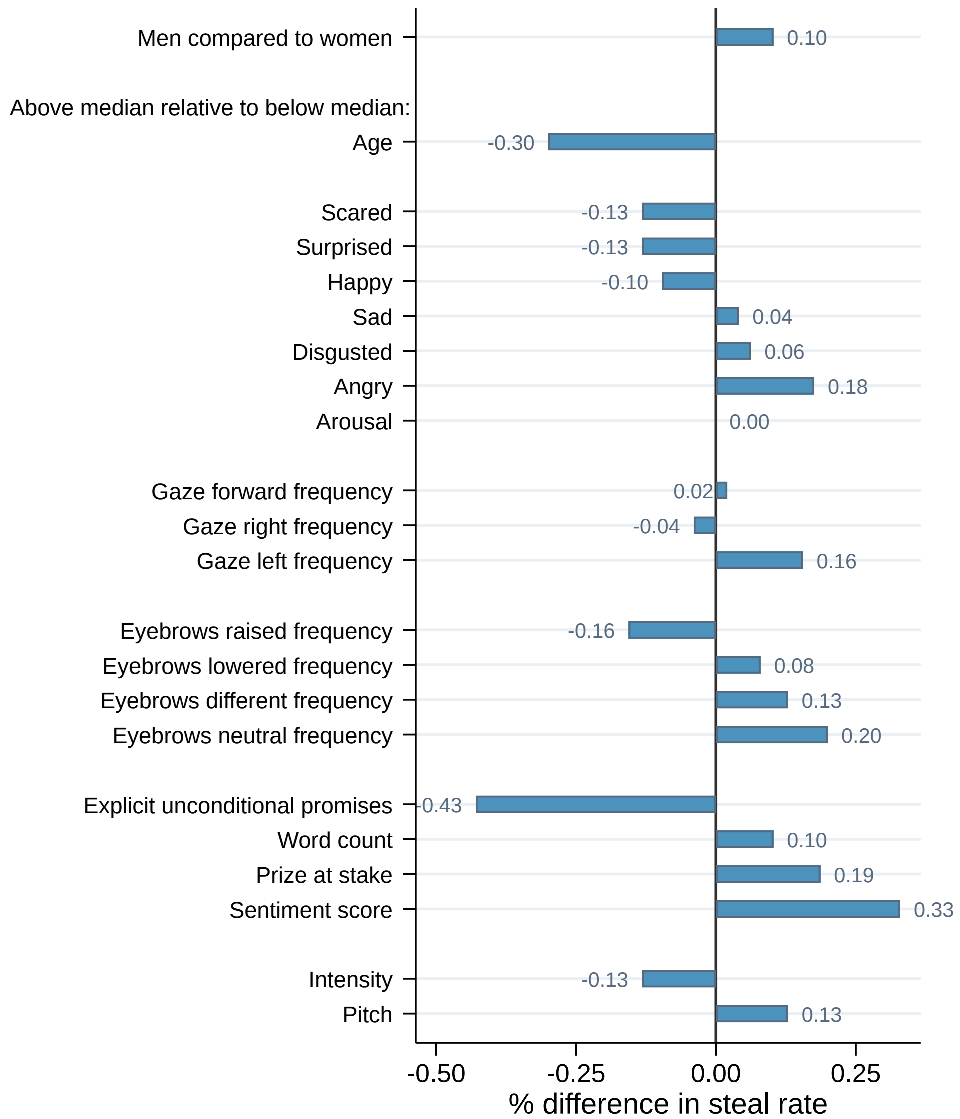


showing the applicability of our methods to real-world settings that are not created just for the use of such software. A documented (Loijens et al. 2016) problem with the software is that it is less accurate when reading non-Caucasian faces, children's faces, or faces over the age of 65. A large majority of the *Golden Balls* participants are Caucasian and between the ages of 18 and 65. More details on FaceReader are provided in Online Appendix B.

Figure 2 below shows summary statistics for contestants in the entire data set, comparing those who chose split to those who chose steal. We provide a detailed comparison in Online Appendix D.1. Descriptively, male contestants were 10% more likely to choose steal, and contestants whose age was above the median were 30% less likely to choose steal. In their conversations with their opponents, contestants who chose steal expressed different emotions. Contestants who were relatively more angry, sad, and disgusted were more likely to choose steal. Those who were more happy, surprised, or scared were less likely to choose steal.

There was also a difference in verbal communication between contestants who chose split and those who chose steal. Contestants who made explicit and unconditional promises were less likely to choose steal. Contestants who said more words and expressed more positive sentiment in their words were more likely to choose steal. Those who had a higher pitch in their voice and lower intensity, which implies their voice was quieter, were also more likely to choose steal.

Figure 2. (Color online) Difference in Steal Likelihood by Contestant Characteristics and Behavior



Notes. Percent difference in the likelihood of steal by covariate. For the only binary covariate, gender, the figure displays the percent difference between men and women. For all other covariates, the figure shows the percent difference in the likelihood of steal between those above the median value of the covariate and those below the median value of the covariate ($N = 430$).

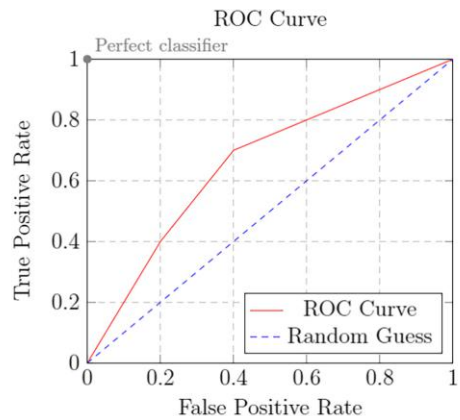
3.3. Measuring Predictive Accuracy

To measure predictive accuracy, we use the predicted probability of steal choice, either by participants in the experiments or by the ML algorithm, and compare it with the actual decision of the contestant. We use two measures of accuracy. The first and simplest measure captures whether the prediction is correct, using a 0.50 threshold. A prediction is correct if the contestant chose split (steal) and the predicted likelihood of split (steal) is above 0.5, and zero otherwise.

Second, we can change the threshold to be any value between zero and one, in addition to the 0.50 threshold. As we vary the threshold, the rate of true positives and false positives will change. The true positive rate (TPR) is the fraction of times that a contestant who chooses steal is correctly predicted to steal. The false positive rate (FPR) is the fraction of times that a contestant who chooses steal is incorrectly predicted to choose split. The AUC is the area under the receiver operating curve (ROC), which presents the false positive rate (FPR) on the x-axis against the true positive rate (TPR) on the y-axis at different thresholds.

As shown in Figure 3, a perfect classifier has all true positive rates (TPR = 1) and no false positives (FPR = 0). Because the AUC is the area under the ROC, the AUC would be one. A "random" classifier would have an ROC on the 45-degree line (TPR = FPR) and thus an AUC of 0.5. The higher the accuracy of the model, the higher the AUC.

Figure 3. (Color online) ROC Curve



4. Individual and Algorithmic Predictive Ability

4.1. Individual Ability: Experimental Design

We conducted two main experiments, summarized in Table 3. In each experiment, we elicited the beliefs of participants regarding whether contestants would choose split or steal. Each participant saw 20 randomly drawn videos and made 20 guesses.

4.1.1. Experiment 1. Experiment 1 allows us to examine whether, in this context, individuals have a limited ability to predict contestant decisions and may benefit from algorithmic feedback. We conducted experiment 1 in two waves. In the first wave of experiment 1 (for all instructions, see Online Appendix A), participants were assigned to the *No Learning* treatment, which meant that they provided their prediction without learning the actual decision contestants made ($N = 258$). In this wave, the videos presented to participants were drawn from a sample of 128 videos, which was the test sample in predictive modeling, as explained in Section 4.2.

Because participants' ability to predict behavior was limited, we added a second wave, which included both the *No Learning* ($N = 52$) and the *Learning* treatment ($N = 159$). In the *Learning* treatment, after providing their prediction, participants learned the contestant's actual choice of split or steal and therefore were afforded the opportunity to learn. We found that learning the actual decision did not increase the predictive accuracy of participants, both measured as the AUC (χ^2 test, $p = 0.659$) and in terms of the fraction of correct guesses (t-test, $p = 0.618$). In this wave, we also expanded the set of videos presented to participants to include all 430 videos in our data. The fraction of correct predictions and the AUC in the *No Learning* treatment did not vary significantly across waves (t-test, $p = 0.844$ for correct guesses and χ^2 test, $p = 0.451$ for the AUC).[5] Hence, we pool all data together to present the main results of experiment 1 and present detailed results in Online Appendix D.2. The experiment was preregistered (preregistration numbers 39504 and 73632 on aspredicted.org) and conducted online (on Prolific Academic).

To examine the robustness of results and whether accuracy would increase when participants' behavior is closely monitored, we ran an additional wave of the experiment in a laboratory at UC San Diego. Participants were assigned to the *No Learning* treatment ($N = 146$). Their accuracy did not significantly differ from that in the online experiment (t-test, $p = 0.967$ for correct guesses and χ^2 test, $p = 0.616$ for the AUC). Hence, we also pool these participants with the online sample in all analyses and present detailed results in Online Appendix D.2.

4.1.2. Experiment 2. Given the limited ability of participants to predict contestant decisions, experiment 2

Table 3. Overview of Experiments

Experiment	Description and treatments	N
1	Human predictive ability	615
2	Timing of algorithmic feedback	
	Control	204
	Flag-Before	202
	Flag-After	191

tests how algorithmic feedback can be designed to increase predictive accuracy. This experiment examines the effect of "flagging" videos that an ML algorithm predicts have a high or low chance of steal choice (preregistration number 107116 on aspredicted.org). The experiment was conducted online only and consisted of three treatments: *Control*, *Flag-Before*, and *Flag-After*. The control group did not provide participants with any information about the ML algorithm's prediction.[6]

In the *Flag-Before* and *Flag-After* treatments, participants were told that the researchers had used the contestants' facial expressions and speech to train a standard ML algorithm to predict choices. They were then told, "We used this algorithm to 'flag' four out of the 20 videos for which the algorithm either predicted that the contestant chose to split [or steal] with very high chance or very low." The flags were symmetric, flagging both high likelihood of cooperation and defection. This approach differs from the use of flags in other contexts with deception, which focus on flagging lies or deceptive content (e.g., Pennycook et al. 2020). In the treatments with flags, participants did not know the accuracy of the algorithm. We chose not to tell participants the accuracy because in most cases in which individuals see flags in naturally occurring environments (e.g., flagged content online), the accuracy is unknown. We elicited participants' beliefs regarding accuracy at the end of the experiment.

In the *Flag-Before* treatment, if the video was flagged, the flag was shown on the screen prior to the screen showing the video. Then, participants saw the video and submitted their prediction on the same screen. In the *Flag-After* treatment, participants watched the video first and were asked to think about their guess, and on the next screen, they submitted their prediction. If the video was flagged, they saw the flag on the screen in which they submitted their prediction. In all cases, participants had to spend at least 20 seconds on the screen that displayed the video, which is the average length of the preplay conversations.

Because there is a separation between the screen in which participants watch the video and the screen in which they submit their guesses in the *Flag-After* treatment, this separation may have an effect on choice. To test for such an effect, we run two versions of the *Control* treatment. In one version, as in *Flag-Before*, participants submitted their guesses on the same screen as they watched the video. In the other version, as in *Flag-After*, participants submitted their guesses on the screen following that in which they watched the video. There is no difference in predictions or accuracy across the two versions of the *Control* treatment (t-test, $p = 0.683$ for predictions, and t-test, $p = 0.369$ for accuracy). We hence present both versions of the *Control* treatment pooled together.

In contrast to experiment 1 in which we show participants different sets of 20 videos drawn from all the available videos from the TV show, in this experiment, participants always watch the same 20 videos to cleanly identify the effects of flags, keeping everything else equal.

In the set of 20 videos, four videos had an ML prediction of over a 70% likelihood of steal (high chance of steal) and four videos for which it had predicted over a 70% likelihood of split (high chance of split). We created two groups of videos that only varied the flags shown (i.e., which videos were flagged to participants). In each group, four videos were flagged. Two flags indicated a very high chance of steal, and two indicated a very low chance of steal. This design implies that participants always saw the same eight videos that could have been flagged, but only four were flagged. Which four were flagged varied by group. Following the actual accuracy of the algorithm, three out of four flags were correct. Each participant was randomly allocated to one of the groups; such that we can measure, at the video level, participants' predictions for each video both when it is flagged and when it is not.

4.1.3. Experimental Procedures. At the beginning of the experiment, participants were asked to read a description of *Golden Balls* and the decision contestants faced. They were shown a video of the presenter of the show explaining the rules governing the split or steal decision and were asked three questions about the rules. As preregistered, if they failed to answer any one of the questions correctly, they were disqualified from participation. Following standard procedures, only participants who completed the study, which required watching all 20 videos, are included in the sample. Overall, the decisions of 1,212 participants in experiments 1 and 2 constitute our main analysis sample.

Participants then received the instructions for the guessing task. We used a stochastic scoring rule based on Holt and Smith (2009) and Karni (2009) to incentivize guesses. Specifically, we asked participants to guess the likelihood that the contestant chose steal or split (balanced) on a scale from 0 to 100. We refer to this guess as G. The receiver's potential bonus payment was \$5. One randomly selected guess counted toward payment. For this guess, the computer randomly drew a number R from 1 to 100. If R was smaller than or equal to the participant's G, they received \$5 if their guess was correct, and \$0 otherwise. If G was greater than R, the participant received \$5 with chance R. We provided participants with an example and asked two comprehension questions that the participants had to answer correctly to be able to proceed. We wrote in the instructions that reporting their true guess regarding the choice of the contestant will maximize the chance of earning their bonus payment.

This stochastic rule is an adaptation of the Becker-DeGroot-Marschak method (Becker et al. 1964) used to elicit probabilities instead of willingness to pay. It is simpler than the binarized scoring rule and adequate for eliciting binary probabilities of events as in the case of *Golden Balls*. By using this rule, we were able to obtain a precise prediction of the chance the contestant chose split or steal (for reviews, see Schotter and Trevino 2014, Charness et al. 2021).[7] ML models provide such a prediction, which allows for a comparison between the participants and ML on their predictions. Although this rule is more complex than an elicitation rule that elicits a binary decision (e.g., the participant will split or steal), it results in similar participant accuracy as the one observed in experiments on lie detection (e.g., Serra-Garcia and Gneezy 2021; for a review, see Bond and DePaulo 2006).

After completing their 20 predictions, we elicited three beliefs from participants. First, in the *No Learning* treatment of experiment 1 and in experiment 2, we elicited their estimate of how many of their predictions were correct, using a 0.5 threshold. Second, we elicited their estimated performance relative to other participants by selecting the quartile of performance to which they thought they belonged. Third, in experiment 2, we elicited participants' beliefs regarding the accuracy of the flags generated with the algorithm separately for when they flagged a video as having a high likelihood of steal or a high likelihood of split. In each case, the participant was asked to consider 10 videos flagged by the algorithm and to indicate how many of those flags (between 0 and 10) they believed would be correct. For each question, participants received a $1 bonus if their guess was correct. Participants concluded the study by reporting their gender, age, and whether they had seen the TV show before.

We recruited online participants through Prolific Academic (Peer et al. 2022), restricting the sample to people residing in the United States. Participants needed to have a previous approval rate of over 95% for studies completed on the platform. All participants (online and in the laboratory) had to answer a question checking that they were not a robot. To check that participants could listen to videos, they had to transcribe one sentence that was said in an audio file. They could not participate in the study without correctly completing these checks, which were presented at the very beginning (after they consented to participation).[8]

Across the two experiments, 51.8% of participants were female (51.6% in experiment 1 online sample, 59.6% in experiment 1 laboratory sample, and 50.1% in experiment 2), the average age was 33.0 (30.0 in the experiment 1 online sample, 20.7 in the experiment 1 laboratory sample, and 38.3 in experiment 2), and 89.9% (90.6% in the experiment 1 online sample, 87.0% in the laboratory sample, and 89.9% in experiment 2) reported never having seen the TV show before the study.

4.2. Machine Learning Algorithm

Because the behavior of a contestant prior to the cooperation decision contains many different nonverbal as well as verbal features, we use ML for predictive modeling of contestant behavior. One can apply a variety of ML (or statistical learning) approaches, including unsupervised and supervised learning. We focus on a supervised learning approach: generalized boosted regression trees (GBM) (see Friedman 2002). Regression trees consist of splitting the space of predictions into regions based on covariates. Rather than estimate only one tree, we estimate GBMs, which combine multiple trees with random subsampling to improve prediction accuracy (see Online Appendix C for a detailed description).

Existing prediction models often present a trade-off between interpretability and flexibility (e.g., James et al. 2013). We focus on GBMs because they are flexible, allow for nonlinearity, and have been previously found to have high predictive accuracy (for example, in ecology to predict a species' presence, as reviewed in Hastie et al. 2008, chapter 10). We also estimate regularized logistic regression models with rigorous penalization (rigorous logistic lasso). This approach assumes linearity in the predictors but is easier to interpret than GBM. The predictive accuracy of both methods is similar. We focus on GBM in the main text and report results for rigorous logistic lasso in Online Appendix C. Both prediction methods are widely used and available as standard tools in existing software, which allows for easy replication and extension in future predictive work.

In line with standard methods in the ML literature on prediction models, our analysis is based on two main steps. First, we train an algorithm to predict the likelihood that a contestant will choose steal. Then, we evaluate the algorithm's ability to predict out of sample. For that purpose, we use all videos in the sample (430 videos), and we randomly split the sample into a training data set (302 videos) and a testing data set (128 videos). We analyze whether ML models can reliably predict split or steal decisions out of sample, only on the testing data set. We use the REFORMS checklist (Kapoor et al. 2024) to provide a detailed report of the ML method used (Online Appendix F). Descriptive statistics on the training and the testing data set are also presented in Online Appendix D.1.

4.3. Hypotheses

Our hypotheses are based on the preregistration plans. The first hypothesis is that ML algorithms can detect features that correlate with choices across a large set of contestants. Therefore, they could achieve a predictive accuracy that is significantly better than chance. By contrast, existing work discussed above on lie detection indicates that individuals would display a limited ability to detect lies. This leads to Hypothesis 1:

Hypothesis 1. *Algorithms will outperform participants in predicting behavior. Participants will not be better than chance at predicting whether a contestant chooses split or steal, whereas the ML algorithm will predict better than chance.*

Motivated by existing work on lie detection (Serra-Garcia and Gneezy 2021), we also hypothesized that participants would believe they are better at predicting behavior than they actually are, both in absolute and in relative terms.

Hypothesis 2. *Participants will be overconfident in their ability to predict behavior, both in relative and in absolute terms.*

Our third and main hypothesis regards the impact of ML feedback on behavior. Flags provide information to participants based on the ML algorithm, and we therefore hypothesized that flags will affect participants' beliefs.

We further hypothesized that the timing of flags will be important. Flags will have a stronger effect on participants who have not yet formed a belief about the likelihood of steal or split (*Flag-Before*) relative to participants who first see the video and form a belief prior to seeing the flag (*Flag-After*). The importance of timing, or primacy effects, has been shown with human advice (e.g., Gneezy et al. 2020, Saccardo and Serra-Garcia 2023). Although human advice can be more persuasive than algorithmic advice (e.g., Burton et al. 2020), we hypothesized that the effects of timing would also be present for algorithmic advice and may be an important determinant of individuals' willingness to follow advice generated by algorithms. By changing the order, we expected to change the influence of flags on beliefs and thereby observe a stronger increase in accuracy in *Flag-Before* than in *Flag-After*, given that flags are accurate 75% of the time, whereas individuals display an accuracy that is only slightly above chance.

Hypothesis 3. *Participants' predictions will be affected by the flags based on the algorithm's prediction. Flags will be more effective when shown before rather than after the participant watches the video.*

The extent to which participants follow the algorithm depends on how accurate they believe the algorithm is. Participants knew the algorithm was fed features from a facial analysis software and speech analysis, which could miss certain features of communication (e.g., body movements of the participants or handshakes) that could be observed by them in the video. Hence, although learning about the algorithm's prediction could be valuable, we hypothesized that participants believe the algorithm is only somewhat better than them in predicting. We hypothesized that the timing of flags would not significantly affect beliefs about the algorithm's accuracy because participants are not (fully) aware that they rely on the algorithm differently in *Flag-Before* or *Flag-After*.

Hypothesis 4. *Participants believe that ML algorithms are better than them at predicting contestant behavior but can still make mistakes. Participants' beliefs about the accuracy of ML algorithms do not depend on the timing of flags.*

5. Results

We start by presenting the accuracy of participants and algorithms, focusing on experiment 1. We then examine how participants use flags based on the algorithmic predictions, focusing on experiment 2.

5.1. Experiment 1: Predictive Performance of Participants and ML Models

Participants correctly guess the decisions of contestants 53.5% of the time (using the 50% threshold). Their rate of correct guesses is significantly better than chance (t-test, $p < 0.001$), though by less than four percentage points. Participants make more type I errors (incorrectly believing the contestant will split), 24.7% of the time, than type II errors (incorrectly believing the contestant will steal), 21.8% of the time. Focusing on the AUC, participants achieve an average AUC of 0.53 (95% confidence interval (CI): 0.52–0.54), which is significantly different from chance (Z-statistic = 6.2, $p < 0.001$), by three percentage points.

The ML model correctly predicts 65.6% of the contestants' decisions and achieves an AUC of 0.71 when predicting out of sample. The ML model makes type I errors less frequently than type II errors in 14.8% of the cases compared with 19.5%, respectively. The ML model is more accurate than participants (t-test, $p < 0.001$ for correct guesses, and χ^2 test, $p < 0.001$ for the AUC). This difference in accuracy is also obtained using lasso instead of GBMs (see Online Appendix C.4). Considering the distribution of participants' performance measured using individual-level AUC, the ML model is more accurate than 90.9% of participants.

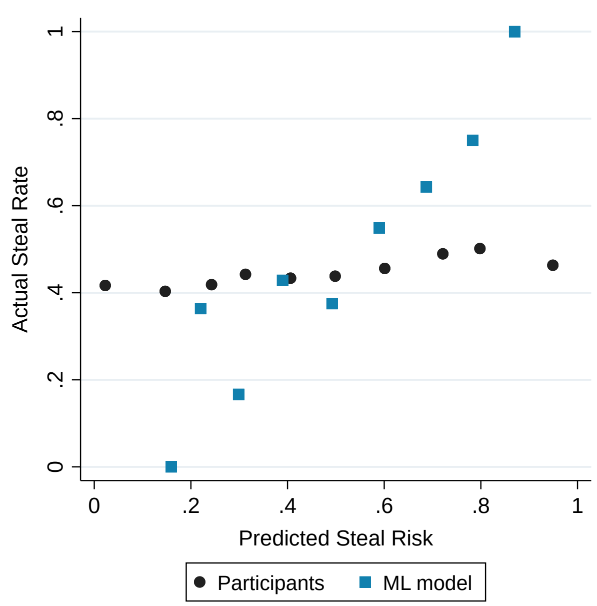
Result 1. Machine learning models predict significantly better than chance (AUC = 0.71) and than participants, who are only better than chance by a small margin (AUC = 0.53).

The relationship between the predictions regarding steal risk and the actual steal rate is shown in Figure 4. The relationship between actual steal risk and the ML model's predictions is stronger than for participants' predictions. This observation is confirmed by the marginal effects of regressions in which the actual steal decision of the contestant is the dependent variable, and the predicted steal likelihood is the independent variable. For participants, the relationship between the predicted steal risk and the actual steal risk is weak (detailed regression results are shown in Online Appendix D.2). A one-percentage-point increase in participants' predicted steal risk is associated with a 0.096-percentage-point increase in the actual steal rate. This relationship is positive and significant. By contrast, for the ML model, a one-percentage-point increase in the predicted steal risk is associated with a 0.964-percentage-point increase in the actual steal rate. The coefficient is not significantly different from one (χ^2 test, $p = 0.84$).

Participants' beliefs suggest they are aware that the task is difficult. They believe they only correctly predict 56.4% of the decisions, but this belief overestimates their actual success rate (t-test, $p < 0.001$). Their beliefs are uncorrelated with actual performance (Spearman correlation coefficient, $\rho = -0.009$, $p = 0.84$), suggesting that participants were not able to predict how accurate their guesses were.

Despite the task difficulty, participants believe they are better than others in relative terms, consistent with previous findings (e.g., Serra-Garcia and Gneezy 2021). Less than 3.6% of participants place themselves in the fourth (and last) quartile of the distribution of performance. By contrast, 50.2% believe their performance is in the second quartile of the distribution, above the median. But few believe they are in the top quartile, as only 12.2% place themselves in the first quartile.[9]

Figure 4. (Color online) Predicted Steal Risk and Actual Stealing, Using Participant and ML Predictions



Notes. Relationship between each decile of predicted steal risk and the actual steal rates of the contestant. The relationships for participants' predictions are shown in black circles. The relationship for the ML model is shown in squares.

Result 2. Participants are overconfident about their ability to predict behavior, in absolute and relative terms, though not by a large magnitude.

5.2. Correlates of Participant and Algorithmic Predictions

The advantage of the ML model is that its predictions correlate with the correct cues participants' beliefs do not. For example, consistent with the descriptive statistics shown in Figure 2, age and the prize at stake are two important features that the model consistently uses to make predictions about steal choice. In addition, several emotions are used to predict the contestants' decisions: sadness, disgust, happiness, and anger. How often the contestant gazes left is another important facial cue for behavior used by the algorithm. In addition, two features of their speech matter: sentiment score and whether the contestant makes an explicit, unconditional promise to choose split. Voice intensity matters as well. We provide details on the relative influence of each covariate and the relationship between features and predictions in Online Appendix D.3.[10]

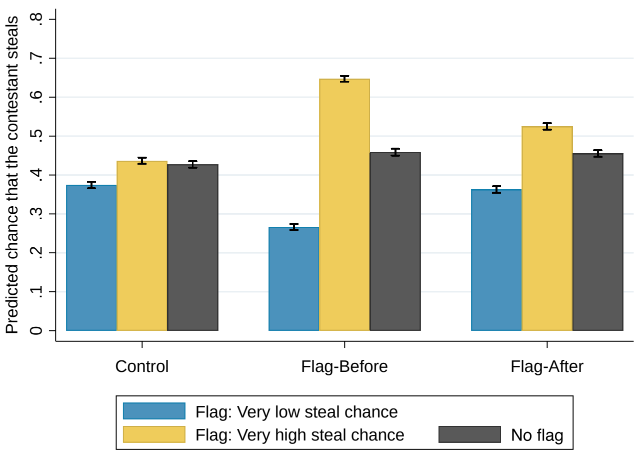
5.3. Experiment 2: Flagging Predictions of the Algorithm

Because the ML algorithm is more accurate than participants, it can be used to provide algorithmic feedback to participants and potentially shape their beliefs. Participants may be open to receiving predictions from the model. Yet they may also believe that because videos are a rich source of information (and multidimensional), the human eye is able to capture many subtle cues that are not easily coded into features upon which the algorithm is trained. This potential difference between the perception of the human eye and the features considered by the algorithm provides participants with ambiguity regarding the extent to which they should rely on the algorithm rather than their own beliefs.

In Figure 5, we compare the average prediction of participants when exposed to different flags and no flags, by treatment. Without flags, participants believe the chance of steal choice is 37.4% for videos that the ML algorithm would have flagged as "very low chance of stealing" and 43.7% for those that would be flagged as very high. The difference between videos that would be flagged as "very low" or "very high chance of stealing" is significant (t-test, $p < 0.001$) by only six percentage points.

Flags significantly change predictions. Participants in the *Flag-Before* treatment significantly reduce their predicted chance of steal choice to 26.7% when the video is flagged as "low chance of stealing" and increase it to 64.7% when the video is flagged as "high chance of stealing." Participants in *Flag-After* also

Figure 5. (Color online) The Effects of Flags on Beliefs



Notes. Average predicted chance that the contestant chose steal, by treatment and (potential) presence of a flag in experiment 2. Whisked bars denote 95% confidence intervals.

change their predictions, but by a smaller magnitude. Their predicted chance of steal choice is 36.3% when the video is flagged as low chance of stealing, and it is 52.5% when the video is flagged as high chance of stealing.

Table 4 examines the effects of the treatments and the flags on beliefs and the accuracy of predictions, including contestant fixed effects. Column (1) reveals that the introduction of flags leads to a small increase in the belief of steal choice. In both *Flag-Before* and *Flag-After*, the predicted chance that a contestant chooses steal increases by three percentage points. Columns (2) and (3) show that significantly larger changes in beliefs occur for flagged videos, especially in *Flag-Before*.

In *Flag-Before*, when the flag is very low, participants follow the flag and decrease their belief about the chance of steal choice by 17 percentage points relative to the average prediction for all videos. Similarly, when the flag is very high, participants follow it and increase their belief about the chance of steal choice by 16.9 percentage points. Relative to the same flagged videos in *Control*, observing a very low flag in *Flag-Before* leads to a 14-percentage-point-lower guess to steal choice (-0.17 for very low flag $+ 0.03$ overall, F-test, $p < 0.001$). For very high flags, we observe a larger change in beliefs about stealing compared with *Control*, leading to a 20-percentage-point-higher estimated chance of steal choice in *Flag-Before* (0.17 for very high flag plus 0.03 overall, F-test, $p < 0.001$). Participants who are exposed to flags react to them, displaying an average difference in beliefs after being exposed to videos flagged as very low compared with videos flagged as very high chance of steal choice of 34 percentage points, over five times larger than the difference observed in the *Control* treatment (F-test, $p < 0.001$).

In *Flag-After*, participants react to a very low flag by decreasing the predicted chance of steal choice by

Table 4. Effects of Flagging on Beliefs and Accuracy

	Predicted chance contestant steals			Correct prediction		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment effects						
Flag-Before	0.035***	0.036**	0.034**	0.048***	0.028**	0.027**
	(0.013)	(0.014)	(0.014)	(0.010)	(0.012)	(0.012)
Flag-After	0.030**	0.032**	0.031**	0.010	0.003	0.003
	(0.014)	(0.014)	(0.014)	(0.011)	(0.012)	(0.012)
Treatment × Flag effects						
Flag-Before × Very low flag		−0.170***	−0.170***		0.061***	0.061**
		(0.014)	(0.014)		(0.024)	(0.024)
Flag-Before × Very high flag		0.169***	0.169***		0.144***	0.144***
		(0.014)	(0.014)		(0.028)	(0.028)
Flag-After × Very low flag		−0.071***	−0.071***		0.042	0.042
		(0.014)	(0.014)		(0.028)	(0.028)
Flag-After × Very high flag		0.051***	0.051***		0.030	0.030
		(0.015)	(0.015)		(0.029)	(0.029)
Constant	0.510***	0.510***	0.540***	0.510***	0.519***	0.503***
	(0.013)	(0.013)	(0.021)	(0.022)	(0.022)	(0.026)
Demographic controls	No	No	Yes	No	No	Yes
Observations	11,940	11,940	11,940	11,940	11,940	11,940
R^2	0.092	0.113	0.118	0.085	0.087	0.088

Notes. Coefficients and standard errors from linear regression models of participants' beliefs and correctness of predictions (columns (1)–(3)), using the 50% threshold (columns (4)–(6)) in experiment 2. All specifications include video (contestant) fixed effects and an indicator showing to which group of flags the participant was assigned. Demographic controls include gender, age, and familiarity with the TV show. Robust standard errors clustered at the participant level are presented throughout.

*, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

seven percentage points. They also react to a very high flag, increasing their predicted chance of steal choice by five percentage points. Relative to the same flagged videos in *Control*, the effect of a very low flag in *Flag-After* is a small, four percentage points (−0.07 for very low flag, +0.03 overall), but statistically significant (*F*-test, $p = 0.04$). A very high flag has a stronger effect, leading to beliefs about the chance of steal choice that are eight percentage points higher than in *Control* (*F*-test, $p < 0.001$). The effects of both types of flags (very high and very low) are significantly smaller in *Flag-After* compared with *Flag-Before* (*F*-test, $p < 0.001$ in both cases). As a result, participants' beliefs when exposed to videos flagged as very low compared with videos flagged as very high chance of steal choice differ by 12 percentage points in *Flag-After*, which is larger than in *Control*, but significantly smaller than in *Flag-Before* (*F*-test, $p < 0.001$).

Because flags affect beliefs and they are correct 75% of the time, the accuracy of beliefs in *Flag-Before* increases significantly. Table 5 shows that in the *Flag-Before* treatment, the fraction of correct guesses is 66.6% for videos flagged as low chance of steal choice, which is directionally but not significantly higher than in *Control* (*t*-test, $p = 0.25$). For videos flagged as high chance of steal choice, *Flag-Before* leads to a significantly larger increase in accuracy: 60.9% for videos flagged compared with 40.7% in the *Control* treatment (*t*-test, $p < 0.001$). In *Flag-After*, the fraction of correct guesses (62.6%) does not change for videos flagged as low chance relative to *Control* (62.7%, *t*-test, $p = 0.96$). The fraction of correct guesses increases directionally by six percentage points—to 46.9%—for videos flagged as high chance (*t*-test, $p = 0.08$).

Interestingly, participants are more accurate when videos would be flagged as low chance of steal choice, even in *Control* (62.7% versus 40.7%). This difference in accuracy is also observed for the ML algorithm, which is more accurate with very low flags relative to very high flags (79% versus 64% of correct predictions, respectively), and in experiment 1 among participants, who are more accurate when they predict a low steal chance, below 30%, compared with when they predict a high steal chance, above 70% (58.1% versus 46.1% of correct predictions, respectively). These findings suggest that there may be features of contestants and their conversations that enable the detection of those who will split, by both ML and participants, but there are fewer features predictive of stealing, which makes such predictions relatively less accurate for both ML and humans.

Overall, considering all flagged videos, flags lead to a significant increase in the AUC from 0.54 to 0.66 for *Flag-Before* (χ^2 test, $p < 0.001$), but a smaller and only directional increase in the AUC to 0.58 for *Flag-After* (χ^2 test, $p = 0.14$).

Considering nonflagged videos, the accuracy of participants in *Flag-Before* was three percentage points higher. There are two potential reasons for this difference. The first reason is that participants could learn from videos that are flagged, and this effect would lead to higher accuracy in videos that are not flagged. Such an explanation is unlikely because in experiment 1, participants did not improve in the *Learning*

Table 5. Accuracy and Beliefs about Accuracy in Experiment 2

	Treatment		
	Control	*Flag-Before*	*Flag-After*
	(1)	(2)	(3)
Panel A: Fraction of correct guesses			
Videos flagged as low chance (%)	62.7	66.6	62.6
Videos flagged as high chance (%)	40.7	60.9	46.9
Not-flagged videos (%)	55.0	58.1	55.6
Overall (%)	54.4	59.2	55.4
Panel B: AUC			
Flagged videos	0.54	0.66	0.58
Not-flagged videos	0.60	0.63	0.61
Overall	0.59	0.64	0.60
Panel C: Beliefs			
Absolute ability (%)	58.9	57.7	60.6
Relative ability	2.17	2.24	2.17
Accuracy of flags indicating low chance (%)	—	61.9	62.3
Accuracy of flags indicating high chance (%)	—	59.5	57.1

Note. This table shows the fraction of correct guesses (50% threshold), AUC, and participants' beliefs about ability, by treatment.

treatment, when they were told the actual choice of each contestant after their prediction. Additionally, as reported in Online Appendix D.2, we observe that even before the first flagged video is shown to participants, the accuracy of participants in *Flag-Before* is three percentage points higher. There is also no evidence of increased accuracy with the number of flagged videos observed. The second reason is that, by chance, participants in *Flag-Before* were more accurate. The data suggest that this slightly higher accuracy explains the difference between the groups.

Columns (4)–(6) of Table 4 examine the effects of flags on the fraction of correct predictions. After controlling for baseline differences in accuracy, participants in *Flag-Before* are six percentage points more accurate in their predictions when observing a very low flag and 14 percentage points more accurate when observing a very high flag. This increase in accuracy is consistent with the stronger impact of flags on predictions for very high flags compared with very low flags (F-test, $p = 0.02$). There is no significant increase in accuracy in *Flag-After*, either in response to very high or very low flags, though the effects are directionally positive. The effect of flags is similar in *Flag-Before* and *Flag-After* when the flag is very low (F-test, $p = 0.56$), whereas the effect of very high flags is significantly stronger in *Flag-Before* than *Flag-After* (F-test, $p = 0.002$).

Result 3. Participants' guesses are significantly affected by the ML flags, with significantly stronger effects in *Flag-Before* than in *Flag-After*. These effects lead to a significant increase in predictive accuracy in *Flag-Before* but not in *Flag-After*.

Participants' beliefs about their own absolute ability to correctly predict contestant behavior are lower in *Flag-Before* (57.7%) than in *Flag-After* (60.6%), leading them to be directionally more confident about their absolute ability in *Flag-After* than in *Flag-Before* (t-test, $p = 0.05$) without affecting their believed relative ability.

Participants' beliefs about the accuracy of the ML algorithm do not vary significantly, depending on the timing of the flags. In *Flag-Before* and *Flag-After*, participants believe the algorithm to be correct 62% of the time when the flag is very low. When the flag is very high, participants believe the algorithm is correct 59.5% of the time in *Flag-Before* and 57.1% in *Flag-After* (t-test, $p = 0.21$). In both treatments, beliefs are below the actual accuracy of the ML algorithm. Considering all videos for which the algorithm made extreme predictions (out of sample), the algorithm is correct 79.3% of the time when the flag is very low and 64.3% when the flag is very high.

When videos are flagged, beliefs about the accuracy of flags matter. As shown in Figure 6(a), when a video is flagged as very low chance of steal choice, participants who believe these flags are highly accurate believe the chance the contestant steals is lower. Similarly, Figure 6(b) shows that when a video is flagged as very high chance of steal, participants who believe those flags are more accurate believe the chance of steal is higher. Irrespectively, in both cases, the stronger impact of flags shown in *Flag-Before*, compared with flags shown in *Flag-After*, emerges across the distribution of beliefs about flag accuracy. In addition, the strength of the relationship between beliefs about flag accuracy and beliefs about stealing does not change depending on the timing of flags (t-test, $p = 0.947$ and $p = 0.610$, respectively). Detailed regression results are shown in Online Appendix D.2.

Result 4. Participants are more confident in their ability to predict in *Flag-After* than in *Flag-Before*. Participants who believe the flags are more accurate are more likely to change their beliefs.

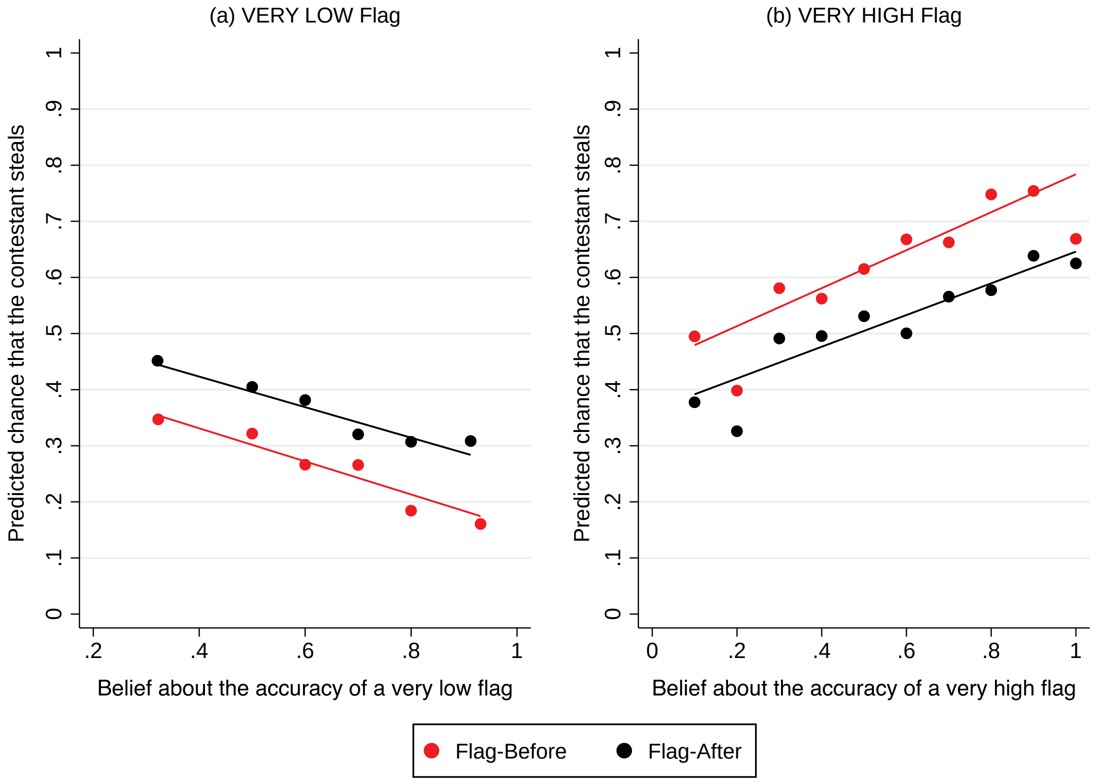
5.4. Understanding the Effects of Timing of Algorithmic Feedback

We conduct two additional analyses to better understand why the effect of algorithmic feedback depends on its timing. First, we examine participants' reports of how they made predictions when the videos were flagged. Second, we examine the time spent watching the videos across different conditions.

Using open-ended questions at the end of the experiment, we asked participants to explain how they made their predictions. After providing their 20 guesses and their beliefs about their own accuracy and the accuracy of the algorithm, they were asked how they made predictions for all videos and, on the same screen, how they made predictions for flagged videos. Two independent coders, blind to treatment assignment, classified their answers into several categories. The four most prevalent categories were participants reporting to follow the algorithm ("Trust ML"), combine their own belief with the algorithm ("ML & own belief"), only follow their own beliefs ("Own belief"), and not trust the algorithm ("Not trust ML"). The agreement between coders was high (Cohen's kappa $\kappa = 0.82$).

Participants report to react differently to flags shown before and after they watch the video (χ^2 test, $p < 0.001$). Figure 7 shows the distribution of categories for the *Flag-Before* and *Flag-After* treatments, based on the cases in which coders agreed. In *Flag-Before*, 31% of participants report to trust ML, whereas 40% combine their own belief with ML. Hence, 71% report following the algorithm. By contrast, in *Flag-After*, 40% of participants report following their own belief (despite the flag). They are less likely to trust ML (13%) or combine it with their own belief (22%). Using word frequency analyses to compare how participants report using flags in *Flag-Before* and *Flag-After*, we find a qualitatively similar

Figure 6. (Color online) Effects of Flags and Beliefs About Their Accuracy



Notes. (a) Average participant's belief when exposed to a very low flag by decile of the distribution of beliefs about accuracy of the algorithm in these flags. (b) Average participant's belief when exposed to a very high flag by decile of the distribution of beliefs about accuracy of the algorithm in these flags. Each dot is a decile in the distribution of the corresponding belief. Beliefs are only shown for *Flag-Before* and *Flag-After* for flagged videos.

pattern of differences in their explanations across treatments (as reported in Online Appendix D.2).

These findings are consistent with confirmation bias and the primacy effect because participants' reports indicate that they rely on the information they receive first. Participants report relying on the flag more often when it is shown first (*Flag-Before*). They report relying on their own belief when the first piece of information is the video (*Flag-After*).

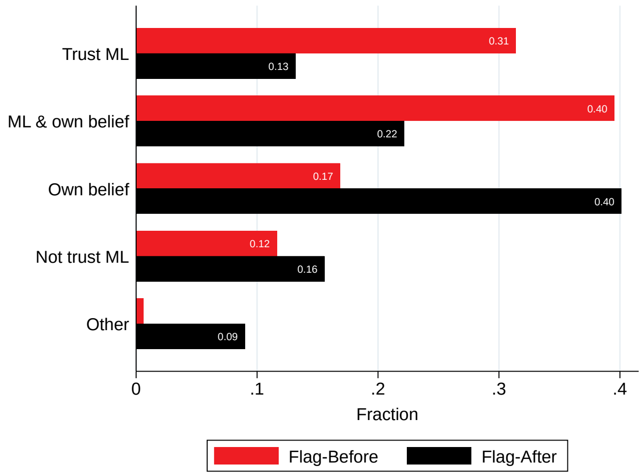
Another reason why flags can affect individuals' predictions differently if they are shown first is that individuals may avoid the effort and time spent in forming a belief by going with the flag (e.g., Dykstra et al. 2022). To explore this explanation, we use time spent on each video. On average, participants spend 39.7 seconds watching each video, significantly more time than the minimum of 20 seconds, which is also the typical duration of the contestants' conversation. We do not find a difference in time spent depending on whether the flag was shown before or after the video was watched. They spend 36.5 seconds (standard deviation (SD) = 19.2) watching when flags are shown before and 38.2 (SD = 47.4) when flags are shown after the video was watched (t-test, $p = 0.24$). Relatedly, time spent on a video is uncorrelated with participants' confidence in their accuracy (Spearman's $\rho = 0.03$, $p = 0.58$), which suggests that spending less time on a video does not explain lower perceptions of ability. Hence, in the context of our experiment, avoidance of the task does not seem to explain the differences in beliefs.

6. Conclusion

We test a potential mechanism to reduce mistakes in evaluating the truthfulness of videos using machine learning algorithms that can flag suspicious videos. Our focus is on how the timing of algorithmic feedback can be designed to affect individual updating and improve belief accuracy.

Our main finding is that the effectiveness of algorithmic feedback depends on when the feedback is provided. Even when standard theoretical frameworks would deem it irrelevant, the timing of feedback is important. After people form their initial beliefs, they are less likely to update using the ML feedback. The effect of timing we find is consistent with confirmation bias. Individuals tend to rely on their own signals and judgment (e.g., Conlon et al. 2022). We leverage

Figure 7. (Color online) How Participants Report to Make Predictions with Flags



Notes. Distribution of participants' reports of how they made predictions when the video was flagged, as coded by two independent raters. The sample includes 86% of the responses for which the coders agreed on a category ($N = 339$ out of 393). "Trust ML" means that the participant reported following the algorithm and relying on or trusting it. "ML & own belief" includes participants who report combining their own belief with the algorithm (and sometimes trusting the flags). "Own belief" includes participants' explanations that are only based on their own beliefs and not the algorithm. "Not trust ML" includes participants who report not trusting the algorithm.

findings from the literature on confirmation bias and prior-based updating, which has shown that initial signals can affect belief formation in some contexts, but not all (e.g., Eil and Rao 2011, Möbius et al. 2022).

In our context, we ask how policymakers could leverage automation and algorithmic models to improve belief accuracy. In the online context, text-based information and video content have the potential to contain deceptive claims. Some companies (e.g., ReviewMeta) offer ML assistance for text-based reviews of online purchases. As in our paper, their algorithm is used to flag reviews in the form of the most and least trusted reviews. Similarly, video platforms, such as TikTok, employ algorithms and humans to review potentially deceptive content and flag it to viewers.

Our findings show that adoption of algorithmic feedback by individuals can strongly depend on its timing in the belief formation process. Flagging content before viewers make their own judgment will likely have stronger effects on their beliefs than flags that appear after the video has been watched or the text read. The results thus contribute to understanding how to design algorithmic advice, such that its effectiveness and use by humans is substantially improved. The literature shows that information design can influence perception (Kamenica 2019, Brooks et al. 2024). Our paper suggests that an important decision in the design and implementation of algorithmic feedback in a variety of applications is its timing. If decision makers choose to introduce algorithmic advice early in the decision-making process, this feedback will likely have stronger effects on decisions than when it is used late in the decision-making process. Such decisions could affect how the human-machine interaction evolves over time and individuals' overall trust in algorithms.

Acknowledgments

This research was conducted under IRB 161827. The authors are grateful to Donja Darai for sharing her videos for this research paper. The authors are grateful to participants at numerous seminars and conferences for their feedback. The authors also thank Muriel Niederle, Sally Sadoff, Anya Samek, Isabel Trevino, Jeroen van de Ven, and Emanuel Vespa for their very helpful comments.

Endnotes

[1] If contestants have no costs associated with lying, preplay communication may result in both players claiming they will cooperate (choosing split), but in equilibrium, these claims are cheap talk (Crawford and Sobel 1982, Farrell and Rabin 1996). Experimental data do not always support the simplifying assumption of no lying costs, and the decisions of some individuals are consistent with them experiencing a psychological cost associated with lying (e.g., Gneezy 2005, Fischbacher and Föllmi-Heusi 2013, Gneezy et al. 2013). If lying is costly enough, players' promises could be informative (Charness and Dufwenberg 2006). In addition, players may have other nonselfish preferences. For example, players may prefer to match their behavior to that of their counterparts: cooperate if they do and defect if they do not (e.g., Rabin 1993, Dufwenberg and Kirchsteiger 2004). In such cases, players may promise to cooperate and follow up on this promise.

[2] In other work, individuals interact with virtual players who are present while others are tasked with detecting deception, but do not provide advice (e.g., Azaria et al. 2015, Varga et al. 2015).

[3] von Schenk et al. (2024) ask individuals to pay for advice after they have been presented with a statement that may be true or false. They find that only 33% of the individuals purchase the algorithmic advice, but if they purchase it, they follow it closely.

[4] The characteristics of the sample we use are similar to those in Turmunkh et al. (2019). In the 284 episodes they study, 54% of the contestants are women, the average age is 37 years old, the average prize is GBP 13,510, and 48% of the contestants choose steal.

[5] In experiment 1, an additional group of participants only watched muted videos of the contestant about whom the participant made a prediction (nonverbal information treatment). The results are presented in Online Appendix D.4. In this nonverbal information treatment, participants' accuracy was not better than chance, with an AUC of 0.49, and significantly worse than in the verbal treatment (χ^2 test, $p < 0.001$).

[6] In an additional experiment, participants' preferences to delegate predictions to the algorithm for flagged videos were elicited. Over half of the participants (53%) chose to delegate to the algorithm. Details are provided in Online Appendix D.5.

[7] Recently, Danz Vesterlund and Wilson (2022) highlighted that the Binarized Scoring Rule (BSR) could lead participants to report conservative beliefs. To explore how much of a concern this possibility could be in our setting, we compare the distribution of participants' beliefs and the model's guesses (see Online Appendix D.2). We observe a positive mass of beliefs at both zero and one for participants, suggesting that the incentives did not lead participants to guess away from the extremes.

[8] In wave 1 of experiment 1, Prolific participants received $2.25 as their participation fee; in wave 2, they received $3.00. Because of the increasing wages in Prolific, in experiment 2, participants received $3.50.

[9] By design, the average quartile is 2.5 (because quartiles range from one to four). Among participants, the average quartile belief is significantly lower at 2.29 (χ^2 test, $p < 0.001$).

[10] We consider *how* emotions, speech, and other covariates are related to the likelihood of steal choice in the ML model, using partial dependence plots (Hastie et al. 2008), shown in Online Appendix D.3. Partial dependence plots present the effect of a covariate on the likelihood of steal choice, accounting for the average effects of all other covariates. Consistent with the raw data and the findings in van den Assem et al. (2012), the likelihood of steal increases with the prize at stake and decreases with the age of the contestant. In addition, based on facial expressions and emotions, participants who gaze left more often are more likely to choose steal. Also, more angry and more disgusted contestants are predicted to choose steal with a higher chance. Happier and sadder contestants are predicted to choose steal with a lower chance. For words said, which can be more positive or more negative, saying more positive words is not associated with a lower chance of choosing steal but, rather, with a slightly higher one. Those contestants with a higher voice intensity (volume) are more likely to steal.

References

Abeler J, Nosenzo D, Raymond C (2019) Preferences for truth-telling. *Econometrica* 87(4):1115–1153.

Agarwal N, Moehring A, Rajpurkar P, Salz T (2023) Combining human expertise with artificial intelligence: Experimental evidence from radiology. NBER Working Paper No. 31422, National Bureau of Economic Research, Cambridge, MA.

Allen RT, Choudhury R (2022) Algorithm-augmented work and domain experience: The countervailing forces of ability and aversion. *Organ. Sci.* 33(1):149–169.

Azaria A, Richardson A, Kraus S (2015) An agent for deception detection in discussion based environments. *CSCW'15: Proc. 18th ACM Conf. Comput. Supported Cooperative Work Soc. Comput.* (Association for Computing Machinery, New York), 218–227.

Baron J (2000) *Thinking and Deciding,* 3rd ed. (Cambridge University Press, Cambridge, UK).

Becker GM, Degroot MH, Marschak J (1964) Measuring utility by a single-response sequential model. *Behav. Sci.* 9(3):226–232.

Belot M, Bhaskar V, van De Ven J (2009) Promises and cooperation: Evidence from a TV game show. *J. Econom. Behav. Organ.* 73(3): 396–405.

Belot M, Bhaskar V, van De Ven J (2012) Can observers predict trustworthiness? *Rev. Econom. Statist.* 94(1):246–259.

Belot M, van de Ven J (2017) How private is private information? The ability to spot deception in an economic game. *Experiment. Econom.* 20(1):19–43.

Benjamin DJ (2019) Errors in probabilistic reasoning and judgment biases. Bernheim D, DellaVigna S, Laibson D, eds. *Handbook of Behavioral Economics* (Elsevier Press, Amsterdam).

Bijlstra G, Dotsch R (2011) FaceReader 4 emotion classification performance on images from the Radboud Faces database. Unpublished manuscript, Department of Social and Cultural Psychology, Radboud University Nijmegen, Netherlands.

Boersma P, Weenink D (2020) Praat: Doing phonetics by computer [computer program]. V. 6.2.23. Accessed October 13, 2020, http://www.praat.org/.

Bond CF, DePaulo BM (2006) Accuracy of deception judgments. *Personality Soc. Psych. Rev.* 10(3):214–234.

Brooks B, Frankel A, Kamenica E (2024) Comparisons of signals. *Amer. Econom. Rev*. 114(9):2981–3006.

Bundorf K, Polyakova M, Tai-Seale M (2019) How do humans interact with algorithms? Experimental evidence from health insurance. NBER Working Paper No. 25976, National Bureau of Economic Research, Cambridge, MA.

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

Centorrino S, Djemai E, Hopfensitz A, Milinski M, Seabright P (2015a) Honest signaling in trust interactions: Smiles rated as genuine induce trust and signal higher earning opportunities. *Evolution Human Behav.* 36(1):8–16.

Centorrino S, Djemai E, Hopfensitz A, Milinski M, Seabright P (2015b) A model of smiling as a costly signal of cooperation opportunities. *Adapt. Human Behav. Physiology* 1:325–340.

Chan MS, Jones CR, Hall Jamieson K, Albarracín D (2017) Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psych. Sci.* 28(11):1531–1546.

Charness G, Dave C (2017) Confirmation bias with motivated beliefs. *Games Econom. Behav.* 104:1–23.

Charness G, Dufwenberg M (2006) Promises and partnership. *Econometrica* 74(6):1579–1601.

Charness G, Gneezy U, Rasocha V (2021) Experimental methods: Eliciting beliefs. *J. Econom. Behav. Organ.* 189:234–256.

Chen D, Hopfenstiz A, van Leeuwen B, van de Ven J (2019) The strategic display of emotions. CentER Discussion Paper 2019-014, Center for Economic Research (CentER), Tilburg University, Tilburg, Netherlands.

Chen X, Ita Levitan S, Levine M, Mandic M, Hirschberg J (2020) Acoustic-prosodic and lexical cues to deception and trust: Deciphering how people detect lies. *Trans. Assoc. Comput. Linguistics* 8:199–214.

Chugunova M, Sele D (2022) We and it: An interdisciplinary review of the experimental evidence on how humans interact with machines. *J. Behav. Experiment. Econom.* 99:101897.

Conlon JJ, Mani M, Rao G, Ridley M, Schilbach F (2022) Not learning from others. NBER Working Paper No. 30378, National Bureau of Economic Research, Cambridge, MA.

Constancio AS, Tsunoda DF, Silva HDFN, Silveira JMD, Carvalho DR (2023) Deception detection with machine learning: A systematic review and statistical analysis. *PLoS One* 18(2):e0281323.

Crawford VP, Sobel J (1982) Strategic information transmission. *Econometrica* 50(6):1431–1449.

Danz D, Vesterlund L, Wilson A (2022) Belief elicitation and behavioral incentive compatibility. *Amer. Econom. Rev.* 112(9):2851–2883.

Darai D, Gratz S (2013) Attraction and cooperative behavior. Department of Economics Working Paper No. 82, University of Zurich, Zurich, Switzerland.

Dell'Acqua F (2022) Falling asleep at the wheel: Human/AI collaboration in a field experiment on HR recruiters. Working paper, Laboratory for Innovation Science, Harvard Business School, Boston.

Dietvorst BJ, Simmons JP, Massey C (2018) Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Sci.* 64(3):1155–1170.

Dijkstra J (1999) User agreement with incorrect expert system advice. *Behav. Inform. Tech.* 18(6):399–411.

Dufwenberg M, Kirchsteiger G (2004) A theory of sequential reciprocity. *Games Econom. Behav.* 47(2):268–298.

Dwenger N, Lohse T (2019) Do individuals successfully cover up their lies? Evidence from a compliance experiment. *J. Econom. Psych.* 71:74–87.

Dykstra H, Exley CL, Niederle M (2022) When do individuals give up agency? The role of decision avoidance. Working paper, Harvard Business School, Boston.

Eil D, Rao JM (2011) The good news-bad news effect: Asymmetric processing of objective information about yourself. *Amer. Econom. J. Microeconom.* 3(2):114–138.

Ekman P (1970) Universal facial expressions of emotion. *Calif. Mental Health Res. Digest* 8:151–158.

Ekman P (2009) *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage* (W.W. Norton & Company, New York).

Farrell J, Rabin M (1996) Cheap talk. *J. Econom. Perspect.* 10(3): 103–118.

Fischbacher U, Föllmi-Heusi F (2013) Lies in disguise–An experimental study on cheating. *J. Eur. Econom. Assoc.* 11(3):525–547.

Friedman J (2002) Stochastic gradient boosting. *Comput. Statist. Data Anal.* 38(4):367–378.

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

Gneezy U, Rockenbach B, Serra-Garcia M (2013) Measuring lying aversion. *J. Econom. Behav. Organ.* 93:293–300.

Gneezy U, Saccardo S, Serra-Garcia M, van Veldhuizen R (2020) Bribing the self. *Games Econom. Behav.* 120:917–946.

Hastie T, Tibshirani R, Friedman J (2008) *The Elements of Statistical Learning* (Springer, New York).

Holt CA, Smith AM (2009) An update on Bayesian updating. *J. Econom. Behav. Organ.* 69(2):125–134.

Hu A, Ma S (2021) Persuading investors: A video-based study. NBER Working Paper No. 29048, National Bureau of Economic Research, Cambridge, MA.

James G, Witten D, Hastie T, Tibshirani R (2013) *An Introduction to Statistical Learning* (Springer, New York).

Kamenica E (2019) Bayesian persuasion and information design. *Annual Rev. Econom.* 11:249–272.

Kapoor S, Cantrell E, Peng K, Pham TH, Bail CA, Gundersen OE, Hofman JM, et al. (2024) REFORMS: Consensus-based recommendations for machine-learning-based science. *Sci. Adv.* 10(18): eadk3452.

Karni E (2009) A mechanism for eliciting probabilities. *Econometrica* 77(2):603–606.

Kim H, Glaeser EL, Hillis A, Kominers SD, Luca M (2024) Decision authority and the returns to algorithms. *Strategic Management J.* 45(4):619–648.

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

Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J, Mullainathan S (2018) Human decisions and machine predictions. *Quart. J. Econom.* 133(1):237–293.

Konrad K, Lohse T, Qari S (2014) Deception choice and self-selection: The importance of being earnest. *J. Econom. Behav. Organ.* 107:25–39.

Leib M, Köbis N, Rilke RM, Hagens M, Irlenbusch B (2024) Corrupted by algorithms? How AI-generated and human-written advice shape (dis) honesty. *Econom. J.* 134(658):766–784.

List JA (2006) Friend or foe? A natural experiment of the prisoner's dilemma. *Rev. Econom. Statist.* 88(3):463–471.

Loijens L, Krips O, Grieco F, van Kuilenburg H, van den Uyl M, Ivan P (2016) *FaceReader: Tool for Automatic Analysis of Facial Expressions: Reference Manual Version 7* (Noldus Information Technology, Netherlands).

Longoni C, Bonezzi A, Morewedge CK (2018) Consumer reluctance toward medical artificial intelligence: The underlying role of uniqueness neglect. Gershoff A, Kozinets R, White T, eds. *Advances in Consumer Research*, vol. 46 (Association for Consumer Research, Duluth, MN), 63–67.

Möbius MM, Niederle M, Niehaus P, Rosenblat SR (2022) Managing self-confidence: Theory and experimental evidence. *Management Sci.* 68(11):7793–8514.

Mynatt CR, Doherty ME, Tweney RD (1977) Confirmation bias in a simulated research environment: An experimental study of scientific inference. *Quart. J. Experiment. Psych.* 29(1):85–95.

Nickerson R (1998) Confirmation bias: A ubiquitous phenomenon in many guises. *Rev. General Psych.* 2(2):175–220.

Oberholzer-Gee F, Waldfogel J, White M (2010) Friend or foe? Cooperation and learning in high-stakes games. *Rev. Econom. Statist.* 92(1):179–187.

Ockenfels A, Selten R (2000) An experiment on the hypothesis of involuntary truth-telling in bargaining. *Games Econom. Behav.* 33(1):90–116.

Peer E, Rothschild D, Gordon A, Evernden Z, Damer E (2022) Data quality of platforms and panels for online behavioral research. *Behav. Res. Methods* 54:1643–1662.

Pennycook G, Bear A, Collins ET, Rand DG (2020) The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Sci.* 66(11):4944–4957.

Perez-Rosas V, Abouelenien M, Mihalcea R, Burzo M (2015) Deception detection using real-life trial data. *Proc. ACM Internat. Conf. Multimodal Interaction (ICMI 2015)* (Association for Computing Machinery, New York), 59–66.

Promberger M, Baron J (2006) Do patients trust computers? *Behav. Decision Making* 19(5):455–468.

Prome SA, Ragavan NA, Islam MR, Asirvatham D, Jegathesan AJ (2024) Deception detection using ML and DL techniques: A systematic review. *Natl. Language Processing J.* 6:100057.

Rabin M (1993) Incorporating fairness into game theory and economics. *Amer. Econom. Rev.* 83(5):1281–1302.

Saccardo S, Serra-Garcia M (2023) Cognitive flexibility or moral commitment? Evidence of demand for moral commitment. *Amer. Econom. Rev.* 113(2):396–429.

Schotter A, Trevino I (2014) Belief elicitation in the laboratory. *Annual Rev. Econom.* 6:103–128.

Serra-Garcia M, Gneezy U (2021) Mistakes, overconfidence and the effect of sharing on detecting lies. *Amer. Econom. Rev.* 111(10):3160–3183.

Stevenson M, Doleac JL (2023) Algorithmic risk assessment in the hands of humans. *Amer. Econom. J. Econom. Policy* 16(4):382–414.

Turmunkh U, van den Assem MJ, van Dolder D (2019) Malleable lies: Communication and cooperation in a high stakes TV game show. *Management Sci.* 65(10):4795–4812.

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

van Leeuwen B, Noussair CN, Offerman T, Suetens S, van Veelen M, van de Ven J (2018) Predictably angry–Facial cues provide a credible signal of destructive behavior. *Management Sci.* 64(7): 3352–3364.

Varga M, Visu-Petra G, Miclea M, Visu-Petra L (2015) The "good cop, bad cop" effect in the RT-based concealed information test: Exploring the effect of emotional expressions displayed by a virtual investigator. *PLoS One* 10(2):e0116087.

von Schenk A, Klockmann V, Bonnefon JF, Rahwan I, Köbis N (2024) Lie detection algorithms attract few users but vastly increase accusation rates. *iScience* 27(7):110201.

Wang W, Gao G, Agarwal R (2023) Friend or foe? Teaming between artificial intelligence and workers with variation in experience. *Management Sci.* 70(9):5753–5775.

Wismüller A, Stockmaster L (2020) A prospective randomized clinical trial for measuring radiology study reporting time on artificial intelligence-based detection of intracranial hemorrhage in emergent care head CT. *Proc. Medical Imaging 2020: Biomedical Appl. Molecular Structural Funct. Imaging*, vol. 11317 (International Society for Optics and Photonics, Bellingham, WA), 144–150.

Wu S, Cai W, Jin S (2015) Motivation enhances the ability to detect truth from deception in audio-only messages. *J. Investigative Psych. Offender Profiling* 12(2):119–126.