CESifo WORKING PAPERS

10518
2023

Original Version: June 2023
This Version: August 2023

Improving Human Deception Detection Using Algorithmic Feedback

Marta Serra-Garcia, Uri Gneezy

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 10518*

Improving Human Deception Detection Using Algorithmic Feedback

Abstract

Can algorithms help people detect deception in high-stakes strategic interactions? Participants watching the pre-play communication of contestants in the TV show Golden Balls display a limited ability to predict contestants' behavior, while algorithms do significantly better. We provide participants algorithmic advice by flagging videos for which an algorithm predicts a high likelihood of cooperation or defection. We find that the effectiveness of flags depends on their timing: 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.

JEL-Codes: D830, D910, C720, C910.

Keywords: detecting lies, machine learning, cooperation, experiment.

Marta Serra-Garcia
Rady School of Management
UC San Diego / CA / USA
mserragarcia@ucsd.edu

Uri Gneezy
Rady School of Management
UC San Diego / CA / USA
ugneezy@ucsd.edu

August 2023
This research was conducted under IRB 161827. We are grateful to Donja Darai for sharing her videos for this research paper. We are grateful to participants at numerous seminars and conferences for their feedback. We also thank Muriel Niederle, Sally Sadoff, Anya Samek, Isabel Trevino, Jeroen van de Ven, and Emanuel Vespa for their very helpful comments.

1 Introduction

Machine learning algorithms are often developed to assist people in making predictions. Do people use the advice algorithms provide? The empirical evidence examining the effects of algorithmic advice on decision-making provides contradictory answers to this question. Some studies find that individuals are likely to adopt algorithmic advice (e.g., Bundorf et al., 2019), while others find that they largely ignore it (e.g., Glaeser et al., 2022). Comparing algorithmic to human advice, some papers find "algorithmic appreciation" (e.g., Logg et al., 2019), while others find hesitancy to follow algorithmic advice (e.g., Longoni et al., 2019). 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 hence 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 belief. By contrast, when people first form their own belief, they put relatively more weight on their own assessment and less on the algorithmic advice.

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 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 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).

The video clips we use are 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 decide simultaneously "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.

While contestants in the TV show elect steal 46% of the time, they almost always make non-binding pre-play 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] Beliefs could be influenced by several features of the conversation, including the content of the communication (e.g., whether the counterpart promised to cooperate) and nonverbal factors (e.g., facial expressions). If participants know how these features are associated with behavior, they could help in predicting the behavior of contestants.

In a first experiment, we show participants display a limited ability to predict contestant behavior, when watching 20 video clips of pre-play communication between contestants. Using a standard metric of classification accuracy, the area under the receiver operating curve (AUC), we find that the AUC for participant predictions is 0.54, which is significantly higher than that of a random classifier (0.50), but by a small margin. This finding holds even if participants are provided with an opportunity to learn about the actual choice of the

[1]If contestants have no costs associated with lying, pre-play 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 is 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 non-selfish preferences. For example, players may prefer to match their behavior to that of their counterpart: 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.

contestant behavior after each prediction.

A simple ML algorithm provides a significant improvement in accuracy relative to our participants, with an AUC of 0.71. The ML algorithm's predcitions are based on visual, vocal, and verbal features of the contestants in the videos, including features such as promise-making, which has been shown to be important in this context (Turmunkh et al., 2019). Its predictions are strongly correlated with actual behavior, with a correlation of approximately 0.9, while participant predictions only exhibit a weak correlation of between 0.05 and 0.11.

Given its accuracy relative to participants in our setting, algorithmic advice has the potential to improve participant predictions. Our main research question is how to provide algorithmic advice such that it will be used by participants. We test whether the effectiveness of advice depends on *when* they are presented to participants.

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 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 in the extreme cases in which the algorithm provides high probability for it. For such extreme predictions, we find that the algorithm is right in 74% of cases, while participants were correct between 51% and 53% of the cases. Third, such flags are easy to apply in practice and similar real-world flagging procedures already exist.

In a 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). Theoretically, 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 feedback.

But, the data show that introducing flags significantly affects participant predictions. Flags shown *before* the participant watches the video lead to more than a 5-fold increase in the difference in participant 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.

Why is there a differential impact of algorithmic advice provided before or after participants watch the video? Our hypothesis was that timing of advice would affect beliefs due to confirmation bias, the tendency of people 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 biased updating when receiving new signals, finding evidence for it 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).

Confirmation bias would suggest that the first piece of information that participants receive becomes their prior, and they rely less on the information received later, leading to a primacy effect. Consistent with this explanation, participants report to overwhelmingly rely on their own beliefs and are more confident in their own ability when they first form their beliefs (in Flag-After). Though timing does not affect the perceived accuracy of the algorithm, participants' increase in confidence leads them to believe that 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 order of flags may also affect participants' behavior if they wish to minimize the time spent in making predictions (e.g., Dykstra et al., 2022). When they see the flag before the video, participants could choose to simply follow it, without forming their own beliefs. Leveraging time data, 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, and double than the average length of contenstant conversations.

Several papers in psychology have studied individuals' preferences for algorithmic advice relative to human advice (for a review, see Chugunova and Sele, 2022). When advice is presented early in the judgment process, several studies have found that individuals tend to follow the algorithmic advice (e.g., Dijkstra et al., 1998; Dijkstra, 1999; Promberger and Baron, 2006; Dietvorst et al., 2018). When individuals first form a judgment in their mind, Longoni et al. (2019) document hesitancy to follow algorithmic advice. Our findings complement the existing literature by focusing on when algorithmic advice is likely to influence individual decision-making.

Understanding the importance of the timing of the algorithmic advice could help organizations improve the adoption of advice. Our findings indicate that to increase its impact, algorithmic advice should be provided to people as soon as possible in the decision-making process, even if standard theoretical frameworks would not predict that the order will influence beliefs. The effectiveness of feedback will be lower when people are already familiar with the situations and have formed an initial belief. Such insight is relevant for a range of applications in which algorithmic advice is used, including predictive modelling within organizations and decision-making by experts with the aid of algorithms.

Consider again the politicians who posts videos with campaign promises or salespersons making promises about their products. 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 guidelines. In the context of online shopping, some companies (e.g., ReviewMeta) offer ML assistance for text-based reviews. Their algorithms scan publicly available data regarding the reviews and identify unnatural patterns. As in our paper, their algorithm is then used for flagging in the form of the most and least Trusted Reviews.

Our results suggest that such the timing of flags is crucially important. Flagging content before viewers make their own judgment will have stronger effects on their beliefs than flags appear after the video has been watched or the text read. Our results thus contribute to understanding how to design the timing of algorithmic advice, such that its effectiveness and use by humans is substantially improved.

2 The Setting: The *Golden Balls* TV Show

Golden Balls is a TV show that was broadcasted in the UK from 2007 to 2009. In the first three rounds of play in the show, contestants make claims about their private information on the potential prize, after which they discuss and vote against each other. The jackpot is determined in the third round.

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. In this final round, the two contestants simultaneously and privately choose split or steal. If they both choose split, they share the prize equally. If they both choose steal, neither receives anything. If one contestant chooses split while the other chooses steal, the former receives nothing while the latter receives the entire prize. The payoff matrix is presented in Table 1.

Table 1: Payoff Matrix in the Prisoner's Dilemma of *Golden Balls*

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

Prior research has documented several interesting behavioral regularities in this gameshow. 54% of contestants choose split in this high-stakes environment (e.g., Belot et al., 2012; van de Assem et al., 2012). Women choose split more often than men, particularly young male contestants, and attractiveness increases cooperation in mixed-gender pairs (van de Assem et al., 2012; Belot et al., 2012; Darai and Gratz, 2013; see also Dreber et al., 2013).

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 pre-play 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 (see also Belot et al., 2009). Among these statements, when contestants told malleable lies (statements that are malleable to ex-post interpretation as truths), they were more likely to choose steal.

The input for our analysis is the videos of the final conversation prior to the split or steal decision (last round of play in each episode). These videos were edited to facilitate the facial analysis, removing 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.

2.1 Facial Analysis, Voice, and Speech Features

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 and/or guilt and overall 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 and 2015b). 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 (2020) use nonverbal and verbal features to estimate the positiveness in videos of startup pitches and relate these emotions to funding decisions.

2.1.1 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 0 and 1.

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 0 and 1. Figure 1 shows an example of the software analyzing a participant in the TV show.

Our main algorithm uses the average of each contestant's emotions and arousal and the average rate of facial movements. Additional analyses using the standard deviation, minimum and maximum of each feature do not increase predictive accuracy (and are described in Online Appendix C).

Figure 1: Example of FaceReader Analysis

2.1.2 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. Past research shows these easy-to-interpret features correlate with lying (Serra-Garcia and Gneezy, 2021). Second, in the case of *Golden Balls*, previous work suggests the prize at stake matters. Third, we also include whether a contestant makes explicit or implicit promises, based on the classification by Turmunkh et al. (2019).

For voice, we include two simple features that describe the contestants' voice: intensity

and pitch (measured using Praat, by Boersma and Weenink (2020), a standard phonetics software). Voices are analyzed through their 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 on the meaning of these features is included in Online Appendix B.

2.2 Descriptive Statistics: Verbal and Nonverbal Behaviors

The sample of *Golden Balls* videos consists of 430 contestants in 215 episodes, from four of the show's six seasons. The average age of contestants was 36 years, and 54.0% were women. The average prize was GBP 13,444, and in 46% of the cases contestants chose steal.[2]

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. On average, 56.4% of the frames for each of these contestants could be read and the emotions analyzed by FaceReader. We find this feature important in showing the applicability of our methods to real world settings that are not created just for the use of such software. A documented 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 (Loijens et al., 2016). 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 dataset, comparing those who chose split to those who chose steal. We provide a detailed comparison in Online Appendix D. Descriptively, male contestants were 10% more likely to choose steal, whereas

[2]The characteristics of the sample we use are similar to those in Turkmukh 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 to steal.

contestants whose age was above median were 30% less likely to choose steal. In their conversations with their opponent, 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.

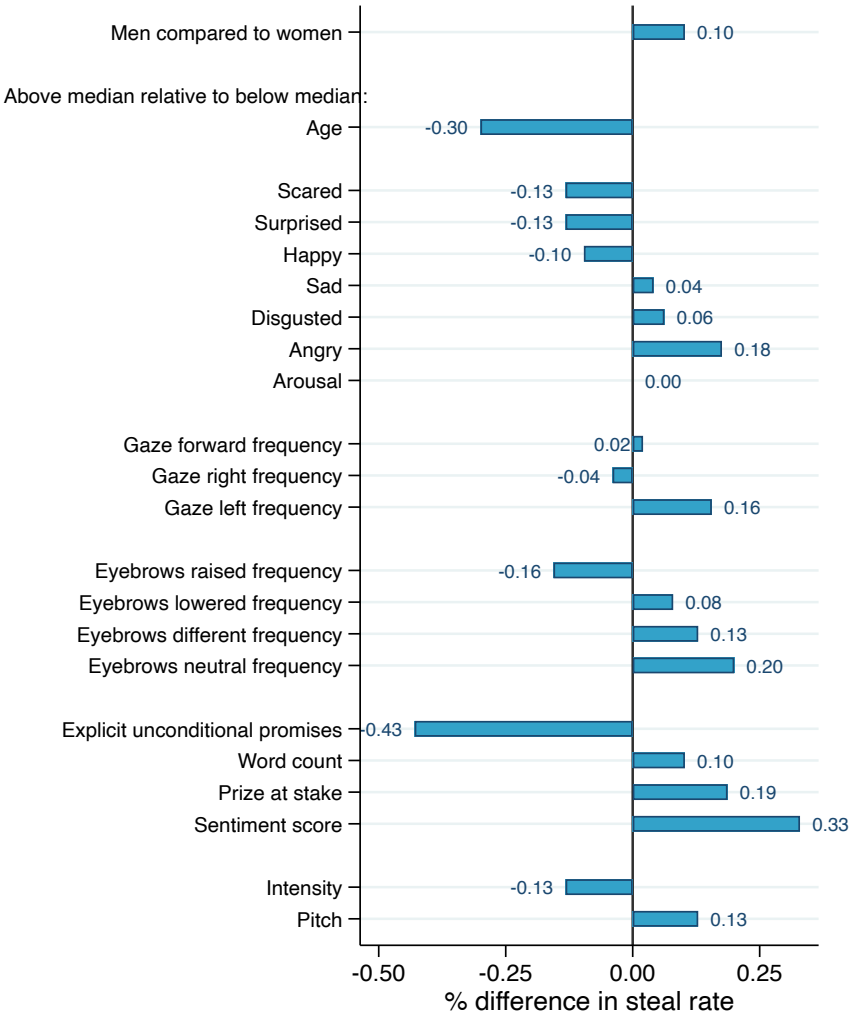
2.3 Measuring predictive accuracy

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 0 otherwise.

Second, we estimate the AUC, which is the area under the receiver operating curve (ROC). Whereas the fraction of correct predictions applies a threshold of 0.50, the ROC presents the rate of false positives (Type I errors) on the x-axis against the rate of true positives (1-Type II errors) on the y-axis, for different threshold rates (from 0 to 1). A true positive is defined as correctly predicting that the contestant will steal, whereas a false positive is defined as incorrectly predicting that the contestant will steal. An ideal ROC curve will have low false-positive rates and high true-positive rates and would be as far as possible from the 45-degree line, which would be how a "no information" classifier would perform. Random guessing would yield an AUC of 0.5, and the closer the AUC is to 1, the higher the accuracy of the model.[3]

[3]We also explore precision-recall measures. Precision is the rate of true positives over the sum of true and false positives and recall is the rate of true positives over the sum of true positives and false negatives.

Figure 2: 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 Individual and Algorithmic Predictive Ability

3.1 Individual Ability: Experimental Design

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. Table 2 provides

These measures deliver similar results.

an overview of the main experiments we conducted.

Experiment 1 consisted of two treatments (for this and all other instructions, see Online Appendix A). In the No Learning treatment, participants watched the 20 videos and provided their prediction without learning the actual decision contestants made. In the Learning treatment, after providing their prediction, participants learned the contestant's choice of split or steal. The experiment was conducted online in two waves (pre-registration #39504 and #73632) on Prolific Academic. In the first wave (N=258), participants were assigned to the No Learning treatment and the videos presented to participants in this wave were drawn from the test sample (i.e., videos not used to train the ML algorithm). The second wave included both the No Learning ($N = 52$) and Learning treatment ($N = 159$). We expanded the set of videos presented to participants to include both videos from the test and the training sample. The fraction of correct predictions and the AUC in the No Learning treatment did not vary significantly across waves (p-value = 0.84 for correct guesses and p-value=0.45 for the AUC).[4]

To examine the robustness of results, and whether accuracy would vary when participants' behavior is closely monitored, we ran an additional wave of the experiment in a laboratory, at west coast university in the US. Participants were assigned to the No Learning treatment ($N = 146$). Their accuracy did not differ significantly from that in the online experiment (p-value = 0.97 for correct guesses and p-value = 0.62 for the AUC). Hence, we pool these participants with the online sample in all analyses.

Experiment 2 examines the effect of "flagging" videos that an ML algorithm predicts have a high or low chance of steal choice (pre-registration #107116, on Prolific Academic). This experiment 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.[5]

In the Flag-Before and Flag-After treatments, participants were told that the researchers

[4]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. 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-value< 0.001).

[5]In an additional experiment, participants' preferences to delegate predictions to the algorithm, for flagged videos, were elicited. Over half of participants (54%) chose to delegate to the algorithm. Details are provided in Online Appendix D.

Table 2: Overview of Experiments

Experiment	Description and Treatments	N
1	Human Predictive Ability	
	- No Learning	456
	- Learning	159
2	Timing of Algorithmic Feedback	
	- Control	204
	- Flag-Before	202
	- Flag-After	191

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 in 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, 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 pre-play conversations.

Since 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 (p-value = 0.68 for predictions, and p-value = 0.37 for accuracy). We hence present both versions of the Control treatment pooled together.

To identify the effect of flags at the video level, we used 20 videos. Among them, 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 varied 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 saw 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 (while other videos are flagged).

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 pre-registered, if they failed to answer any one of the questions correctly, they were disqualified from participation. Overall, 1,212 participants answered the control questions correctly.

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).[6] ML models provide such a prediction, which allows for a comparison between the participants and ML on their predictions. While 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' belief regarding the algorithm's accuracy, separately for when it indicated a high likelihood of steal or a high likelihood of split. 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 US. They needed to have a previous approval rate of over 95% for studies completed on Prolific. All participants (online and in the lab) had

[6]Recently, Danz et al. (2022) highlighted that 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). We observe a positive mass of beliefs at both 0 and 1 for participants, suggesting that the incentives did not lead participants to guess away from the extremes.

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).[7]

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

3.2 Machine Learning Algorithm

Since 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). Existing prediction models often present a tradeoff between interpretability and flexibility (e.g., Hastie et al., 2008). We focus on GBMs because they are flexible, allow for nonlinearity, and they have been previously found to have high predictive accuracy. 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

[7] In wave 1 of Experiment 1, 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.

sample. For that purpose, we randomly split the sample into a training dataset (302 videos) and a testing dataset (128 videos). We analyze whether ML models can reliably predict split or steal decisions out-of-sample, only on the testing dataset. We present a detailed description of the algorithm in Online Appendix C and use the REFORMS checklist (Kapoor et al., 2023) to provide a detailed report of the ML method used (Appendix F). Descriptive statistics on the training and the testing dataset are also presented in Online Appendix D.

3.3 Hypotheses

Our hypotheses are based on the pre-registration plan. The first hypothesis relates to the literature discussed above which shows individuals display a limited ability to detect lies. Our prediction was that, due to the complexity of the task, learning contestants' decisions after making predictions will not increase accuracy. By contrast, we hypothesized that ML algorithms can detect features that correlate with choices across a large set of contestants and yield a prediction function that is better than chance. 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, even when given the opportunity to learn, while the ML algorithm will predict better than chance.

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 order, or primacy effects, has been

shown with human advice (e.g., Gneezy et al., 2020; Saccardo and Serra-Garcia, 2023). We hypothesized that the effects of order would 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 affect 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, while individuals display an accuracy that is only slightly above chance.

Hypothesis 3: Participants' predictions will be significantly affected by the flags based on the algorithm's prediction. Flags will be significantly 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 would believe the algorithm is only somewhat better than them in predicting, and that the algorithm would make some mistakes. We hypothesized that the timing of flags would not significantly affect beliefs about the algorithm's accuracy, because participants would not be (fully) aware that they were relying 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.

4 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.

4.1 Experiment 1: Predictive Performance of Participants and ML Models

Participants achieve an average AUC of 0.54 (95% CI: 0.52 - 0.55) without learning and 0.52 with learning (95% CI: 0.50 - 0.54). The AUC is significantly different from chance without learning (p-value< 0.001), though by only 4 percentage points, and marginally significantly different from chance (p-value $= 0.08$) with learning.

Participants correctly guess the decisions of contestants 54.1% of the time without learning, and 51.8% of the time with learning (using the 50% threshold). These rates of correct guesses are significantly better than chance (p-value< 0.001 and p-value $= 0.03$, respectively), though by less than 5 percentage points. The results with and without learning are not statistically different, both measured as the AUC (p-value $= 0.13$) and in terms of the fraction of correct guesses (p-value $= 0.61$).

The ML model achieves an AUC of 0.71 and correctly classifies 65.6% of the contestants' choices. We find a similar result using lasso (see Online Appendix C). The ML model is more accurate than participants (p-value< 0.001 both with and without learning opportunities).[8]

Result 1: Participants predict significantly better than chance without learning, though by a small margin (AUC = 0.54). Providing participants with feedback does not improve predictions. Machine-learning models predict significantly better than chance and better than participants.

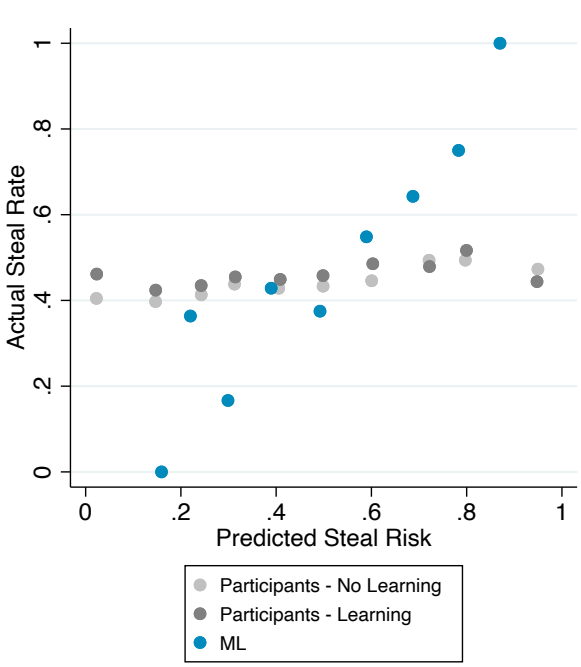
Considering the distribution of participants' performance, the ML model is more accurate than 89% of participants who do not learn about the contestant's decision after submitting their prediction and more accurate than 94% of participants who do receive such information.

The relationship between the predictions regarding steal risk and the actual steal rate is shown in Figure 3. The relationship to the ML model is stronger than for participants' predictions. This observation is confirmed by the marginal effects of regressions in which the

[8]Precision-recall analysis yields similar insights, which are most adequate when the number of observations in each outcome is significantly unequal (e.g., many more steal than split choices). For participants, the area under the precision-recall curve is 0.45, both with and without learning. For ML, the area under the precision-recall curve is 0.62.

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 shown in Online Appendix D). In the No Learning treatment, a one-percentage-point increase in participants' predicted steal risk is associated with a 0.11-percentage-point increase in the actual steal rate. This relationship is positive and significant. The association between predictions and actual steal rates is 0.05 in the Learning treatment, and it does not significantly improve as the participant watches more videos (p-value > 0.10). By contrast, for the ML model, a one-percentage-point increase in the predicted steal risk is associated with a 0.96-percentage-point increase in the actual steal rate. The coefficient is not significantly different from 1 (p-value $= 0.84$).

Figure 3: 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 contestant. The relationship for participants' predictions is shown in gray (light gray for the no-learning treatment and dark gray for the learning one). The relationship for the ML model is shown in blue. A linear fit is added for each treatment/model, separately for participants' predictions and those of the ML model.

4.2 Experiment 1: Participants' Beliefs about Accuracy

In Experiment 1, on average, participants overestimate their absolute accuracy in predicting steal choice, though not by a large magnitude. Participants believe they correctly predicted the behavior of 56.4% of contestants, which is significantly higher than their actual success rate (p-value= 0.01), and their beliefs are uncorrelated with actual performance (Spearman correlation coefficient, $\rho = 0.017$, p-value = 0.72). Although participants believe they perform significantly better than they do, the magnitude of bias is small relative to overconfidence in detecting lies in Serra-Garcia and Gneezy (2021), where individuals correctly detected lies in videos 52% to 53% of the time, but believed they were correct 64% to 67% of the time. These results suggest the task of predicting who will steal may have been perceived as difficult and that participants were aware that they were not able to accurately predict behavior.

Participants show overplacement: less than 2.7% of participants place themselves in the bottom quartile of the distribution of performance. By contrast, 53.3% believe their performance is in the second quartile of the distribution, above median but not in the top quartile, and 14.3% place themselves in the highest quartile. Whereas, by design, the average quartile is 2.5 (because quartiles range from 1 to 4), the average quartile belief is significantly lower at 2.21 (χ^2 test, p-value< 0.001).

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

4.3 Correlates of participant and algorithmic predictions

The advantage of the ML model is that its predictions correlate with the correct cues for steal choice, whereas 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.[9]

4.4 Experiment 2: Flagging Predictions of the Algorithm

Since ML algorithms are more accurate than participants, they 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, since 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 4 we compare the average prediction of participants in each treatment, conditional on whether the algorithm flagged the contestant in the video as having a very low or a very high chance of choosing steal, or if there was no flag.

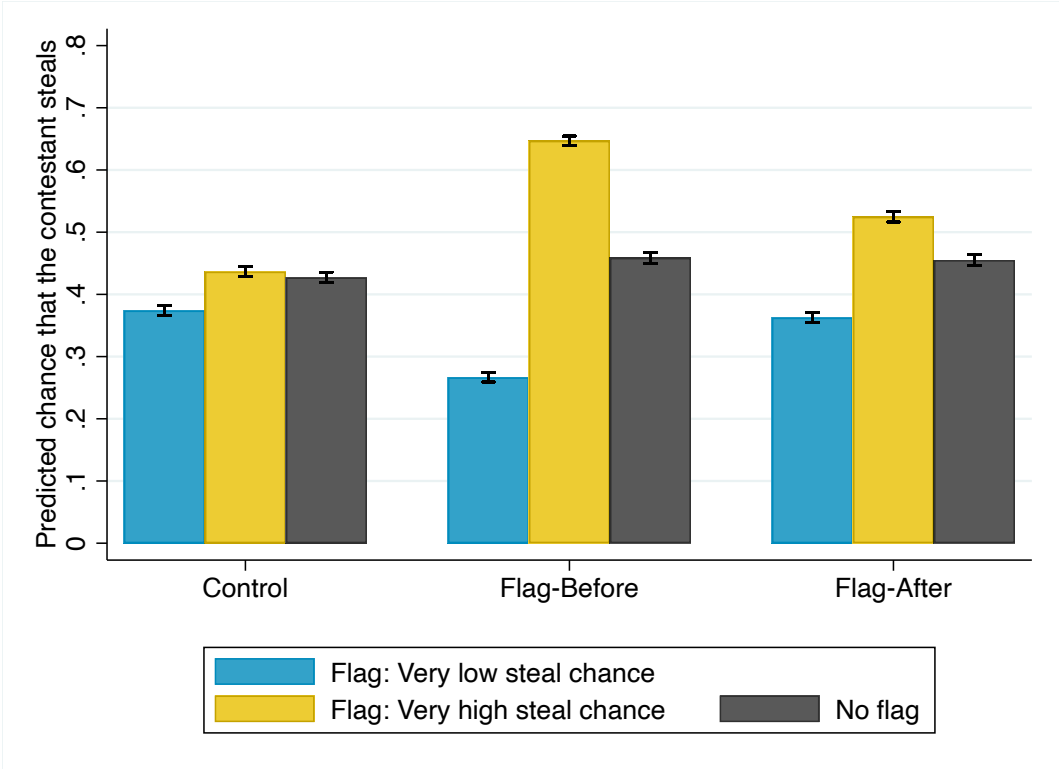
Without flags, participants believe the chance of steal choice is 37.4% for videos that the ML algorithm flagged as "very low chance of stealing," 43.7% for those flagged as "very high," and 42.7% for videos that were never flagged. The difference between videos flagged as "very low" or "very high chance of stealing" is significant (p-value = 0.004).

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

[9]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. Partial-dependence plots present the effect of a covariate xs 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, whereas more happy and more sad 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.

of stealing." Participants in Flag-After also 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."

Figure 4: 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.

Table 3 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 3 percentage points. Columns (2)-(3) show that significantly larger changes in beliefs occur for flagged videos, especially in Flag-Before.

In Flag-Before, "very low" and "very high" flags change the predicted chance of steal choice by 17 percentage points, relative to the average prediction for all videos. Relative to the same flagged videos in Control, the difference in beliefs after observing a "very low" flag in Flag-Before is 14 percentage points ($-$ 0.17 for "very low" flag + 0.03 overall). "Very

high" flags lead to an even larger difference of 20 percentage points relative to control (0.17 for "very high" flag + 0.03 overall). The difference in predictions between videos flagged as "very low" and those flagged as "very high" chance of steal choice is 34 percentage points, over 5 times larger than in the Control treatment.

Table 3: Effects of Flagging on Beliefs and Accuracy

	(1)	(2)	(3)	(4)	(5)	(6)
	Predicted chance contestant steals			Correct prediction		
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 X Flag Effects						
Flag-Before X Very low flag		-0.170***	-0.170***		0.061***	0.061**
		(0.014)	(0.014)		(0.024)	(0.024)
Flag-Before X Very high flag		0.169***	0.169***		0.144***	0.144***
		(0.014)	(0.014)		(0.028)	(0.028)
Flag-After X Very low flag		-0.071***	-0.071***		0.042	0.042
		(0.014)	(0.014)		(0.028)	(0.028)
Flag-After X 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-squared	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 for which group of flags the participant was assigned to. Demographic controls include gender, age and familiarity with the TV show. Robust standard errors clustered at the participant level are presented throughout. *, **, *** indicate significance at the 10%, 5%, and 1% levels.

In Flag-After, a "very low" flag decreases the predicted chance of steal choice by 7 percentage points while a "very high" flag increases the predicted chance by 5 percentage points. Relative to the same flagged videos in Control, the effect of a "very low" flag in Flag-After is a small, 4 percentage points ($-$ 0.07 for "very low" flag + 0.03 overall), but statistically significant (p-value = 0.04). In Flag-After a "very high" flag has a stronger effect of 8 percentage points relative to Control (p-value$<$ 0.001). For both types of flags, the effects of flags are significantly smaller in Flag-After compared to Flag-Before (p-value$<$ 0.001 in both cases). The resulting gap between videos flagged as "very low" and those flagged as

"very high" chance of steal choice is 12 percentage points in Flag-After, which is larger than in Control, but significantly smaller than in Flag-Before (F-test, p-value< 0.001).

Because flags affect beliefs and they are correct 75% of the time, the accuracy of beliefs in Flag-Before increases significantly. Table 4 shows that in the Flag-Before treatment, the fraction of correct guesses is 66.6% for videos flagged as low chance of steal choice and 60.9% for videos flagged as high chance, compared with 62.7% and 40.7%, respectively, in the Control treatment. By contrast, in Flag-After, the fraction of correct guesses does not change for videos flagged as low chance, and it increases by 6 percentage points, to 46.9%, for videos flagged as high chance. The effects of flags for flagged videos result in a significant increase in the AUC from 0.54 to 0.66 for Flag-Before (χ^2 test, p-value< 0.001), but a smaller and directional increase in the AUC to 0.58 for Flag-After (χ^2 test, p-value $= 0.14$).

Table 4: Accuracy and Beliefs about Accuracy in Experiment 2

	(1)	(2) Treatment	(3)
	Control	Flag-Before	Flag-After
Fraction 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%
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
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%

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

Considering non-flagged videos, the accuracy of participants in Flag-Before was 3 percentage points higher, also for non-flagged videos. There are two potential reasons for this difference. First, participants may learn from videos that are flagged, and this effect leads to higher accuracy in videos that are not flagged. Second, by chance, participants in Flag-Before could be more accurate. The data suggest that there is no spillover, but rather a

small difference in accuracy between the groups. In additional analyses, reported in Online Appendix D, we explore whether there is an increased accuracy in predictions for non-flagged videos after flags are observed. We observe that, even before the first flagged video is shown to participants, the accuracy of participants in Flag-Before is 3 percentage points higher. There is also no evidence of increased accuracy with the number of flagged videos observed.

Columns (4)-(6) of Table 3 examine the effects of flags on the fraction of correct predictions. After controlling for baseline differences in accuracy, participants in Flag-Before are 6 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 to "very low" flags. There is no significant increase in accuracy in Flag-After, neither in response to "very high" nor "very low" flags, though the effects are directionally positive.

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 (p-value= 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% in Flag-Before, and 57.1% in Flag-After (t-test, p-value = 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 77.5% of the time when the flag is "very low," and 64.3% when the flag is "very high."

Such patterns of beliefs are consistent with the stronger effects of flags in Flag-Before. In

this treatment, participants believe that the algorithm is either directionally more accurate (for "very high" flags, p-value = 0.30) or significantly more accurate (for "very low" flags, p-value= 0.01) than they are. In Flag-After, by contrast, participants believe they are more accurate than the algorithm when it is used to indicate a "very high" chance of steal choice (p-value = 0.03), or only directionally worse by 1.7 percentage points, when it is used to indicate a "very low" chance (p-value= 0.30).

Result 4: Participants are more confident in their ability to predict in Flag-After than in Flag-Before. In Flag-Before, they believe that ML algorithms predict better than them by a relatively small amount. In Flag-After, they believe that ML algorithms are either worse than them or only slightly better. Overall, participants' beliefs about the accuracy of ML algorithms do not depend on the timing of flags, but timing affects their perceptions about their own ability.

4.5 Understanding the Effects of Timing of Algorithmic Feedback

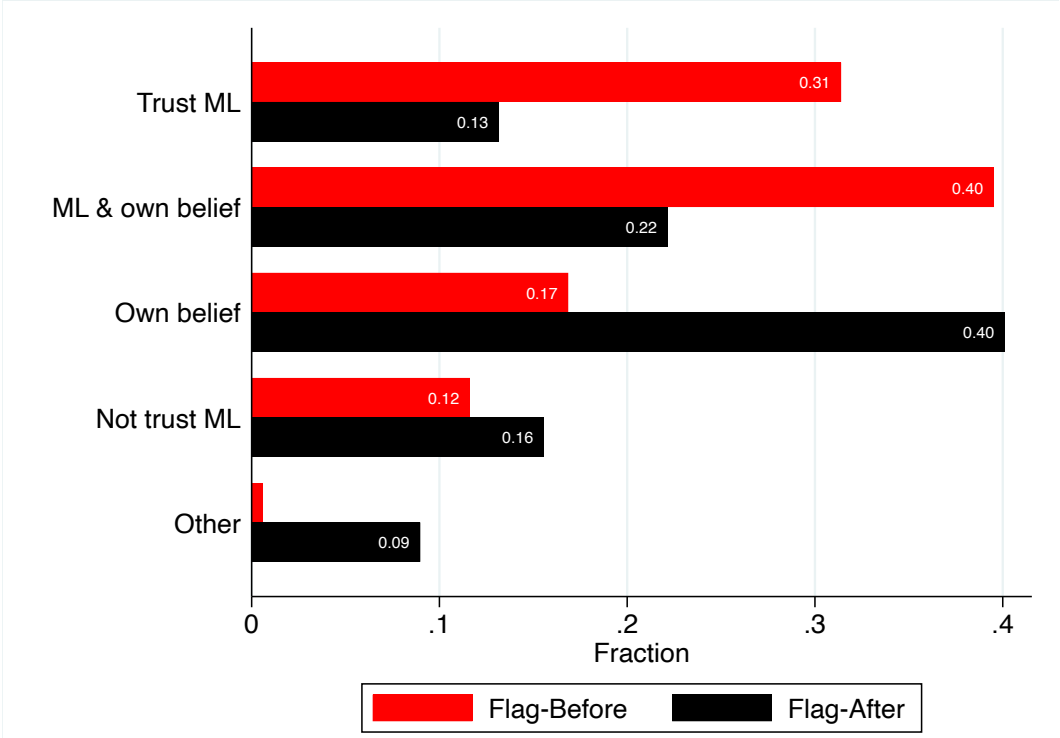
We conduct two additional analyses to better understand why the effect of algorithmic feedback depends on its timing. First, we examine participants' reports of how they made predictions when the videos were flagged. Second, we examine the time spent watching the videos across different conditions.

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, 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 (as detailed in Online Appendix D). 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 $\kappa = 0.82$).

Participants report to react differently to flags shown before and after they watch the

video (p-value$<$ 0.001). Figure 5 shows the distribution of categories for the Flag-Before and Flag-After treatment, based on the cases in which coders agreed. In Flag-Before, 31% of participants report to trust ML, while 40% combine their own belief with ML. Hence, 71% report to follow the algorithm. By contrast, in Flag-After, 40% of participants report to follow their own belief (despite the flag). They are less likely to trust ML (13%) or to combine it with their own belief (22%). Automated text analyses, reported in Online Appendix D, are consistent with these findings.

Figure 5: 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 to follow the algorithm and to rely or trust it. "ML & own belief" includes participants who report combining their own belief with the algorithm (and sometimes trusting the flags). "Own belief" includes participant explanations that are only based on their own beliefs and not the algorithm. "Not trust ML" includes participants who report not trusting the algorithm.

These findings are consistent with confirmation bias and the primacy effect, since participants' reports indicate that they rely on the information they receive first. Participants report to rely on the flag more often when it is shown first (Flag-Before). They report to rely 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 (p-value = 0.24). Relatedly, time spent on a video is uncorrelated with participants' confidence in their accuracy (Spearman's $\rho = 0.03$, p-value= 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 due to timing of flags.

5 Conclusion

Interest in the ability to detect deception has grown substantially with the emergence of social media and spread the vast amounts of video content available online. An important downside of this unfiltered communication is that it can be deceptive. In this paper we test a potential mechanism to reduce mistakes in evaluating videos using machine learning algorithms that can flag suspicious videos.

We study 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 their beliefs 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 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) to apply them to a new context that is becoming increasingly important for policy makers: how to leverage automation and algorithmic models to improve belief

accuracy. Our findings show that adoption of algorithmic feedback can strongly depend on its timing in the belief formation process.

Real-world applications of algorithmic advice in other contexts are suggestive of the importance of timing. 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 to them for sentencing, parole, and related decisions. Since by the time judges receive the algorithmic advice they are already familiar with the case, this timing might explain why the risk assessment tools have relatively small impact. For example, Stevenson and Doleac (2022) 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).

In health care, algorithms are used, for example, in assisting radiologists in a clinical reading workflow environment. Since algorithms are employed *prior* to the radiologists' evaluation of the case, they may have a relatively large impact. Wismüller and Stockmaster (2020) measured the impact of algorithmic flags for head CT scans on Turn-Around Time (TAT). Reducing TAT is important because delayed interventions may be detrimental for patient outcomes. They found that their early flagging procedure reduced TATs substantially, showing that algorithms have a substantial impact on radiologists' decision-making.

The literature shows that information design can influence perception (Kamenica, 2019; Brooks, Frankel and Kamenica, 2023). 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 on algorithms.

References

[1] Abeler, J., Nosenzo, D., and C. Raymond (2019). Preferences for Truth-telling. *Econometrica*, 87 (4), 1115–1153.

[2] Baron, J. (2000). *Thinking and Deciding* (Third ed.). Cambridge University Press.

[3] Becker, G.M., Degroot, M.H., and Marschak, J. (1964). Measuring utility by a single-response sequential model. *Behavioral Science* 9 (3), 226–232.

[4] Belot, M., Bhaskar, V., and van De Ven, J. (2009). Promises and Cooperation: Evidence from a TV Game Show. *Journal of Economic Behavior & Organization* 73(3), 396–405.

[5] Belot, M., Bhaskar, V., and van De Ven, J. (2012). Can observers predict trustworthiness? *Review of Economics and Statistics* 94 (1), 246–259.

[6] Belot, M., and van de Ven, J. (2017). How private is private information? The ability to spot deception in an economic game. *Experimental Economics* 20 (1), 19–43.

[7] Benjamin, D. J. (2019). Errors in Probabilistic Reasoning and Judgment Biases. *Handbook of Behavioral Economics*, edited by Doug Bernheim, Stefano DellaVigna, and David Laibson. Elsevier Press.

[8] Bijlstra, G., and 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, The Netherlands.

[9] Boersma, P., and Weenink, D. (2020). Praat: doing phonetics by computer [Computer program]. Version 6.2.23, retrieved 13 October 2020 from http://www.praat.org/.

[10] Bond, C.F., and DePaulo, B.M. (2006). Accuracy of Deception Judgments. *Personality and Social Psychology Review* 10 (3), 214–234.

[11] Brooks, B., Frankel, A., and Kamenica, E. (2023). Comparisons of Signals. Working paper.

[12] Bundorf, K., Polyakova, M., and Tai-Seale, M. (2019). How do humans interact with algorithms? Experimental evidence from health insurance. Technical Report. National Bureau of Economic Research.

[13] Centorrino, S., Djemai, E., Hopfensitz, A., Milinski, M., and Seabright, P. (2015). Honest Signaling in Trust Interactions: Smiles Rated As Genuine Induce Trust and Signal Higher Earning Opportunities. *Evolution and Human Behavior* 36, 8–16.

[14] Centorrino, S., Djemai, E., Hopfensitz, A., Milinski, M., and Seabright, P. (2015). A Model of Smiling as a Costly Signal of Cooperation Opportunities. *Adaptive Human Behavior and Physiology* 1, 325–340.

[15] Charness, G., and Dave, C. (2017). Confirmation Bias with Motivated Beliefs. *Games and Economic Behavior* 104, 1–23.

[16] Charness, G., and Dufwenberg, M. (2006). Promises & Partnership. *Econometrica*, 74, 1579–1601.

[17] Charness, G., Gneezy, U., and Rasocha, V. (2021). Experimental methods: Eliciting beliefs. *Journal of Economic Behavior & Organization* 189, 234–256.

[18] Chen, D., A. Hopfenstiz, B. van Leeuwen, and van de Ven, J. (2019). The Strategic Display of Emotions. CentER Discussion Paper, 2019-014.

[19] Chugunova, M., and Sele, D. (2022). We and It: An interdisciplinary review of the experimental evidence on how humans interact with machines. *Journal of Behavioral and Experimental Economics* 99, 101897.

[20] Conlon, J. J., Mani, M., Rao, G., Ridley, M., and Schilbach, F. (2022). Not Learning from Others. Working Paper.

[21] Crawford, V.P., and Sobel, J. (1982). Strategic Information Transmission. *Econometrica* 50 (6), 1431–1449.

[22] Danz, D., Vesterlund, L., and Wilson, A. (2021). Belief Elicitation and Behavioral Incentive Compatibility. *American Economic Review* 112 (9), 2851–2883.

[23] Darai, D., and Gratz, S. (2013). Attraction and Cooperative Behavior. University of Zurich Department of Economics Working Paper No. 82.

[24] Dietvorst, B. J., Simmons, J. P., and Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, 64 (3), 1155–1170.

[25] Dijkstra, J. (1999). User agreement with incorrect expert system advice. *Behaviour & Information Technology*, 18 (6), 399–411.

[26] Dijkstra, J., Liebrand, W.B.G, and Timminga, E. (1998). Persuasiveness of expert systems. *Behaviour & Information Technology* 17 (3), 155–163.

[27] Dreber, A., Gerdes, C., and Gransmark, P. (2013). Beauty queens and battling knights: Risk taking and attractiveness in chess. *Journal of Economic Behavior & Organization* 90, 1–18.

[28] Dufwenberg, M., and Kirchsteiger, G. (2004). A Theory of Sequential Reciprocity. *Games and Economic Behavior,* 47, 268–98.

[29] Dwenger, N., and Lohse, T. (2019). Do individuals successfully cover up their lies? Evidence from a compliance experiment. *Journal of Economic Psychology* 71, 74–87.

[30] Dykstra, H., Exley, C.L., and Niederle, M. (2022). When Do Individuals Give Up Agency? The Role of Decision Avoidance. Working Paper.

[31] Eil, D., and Rao, J. M. (2011). The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself. *American Economic Journal: Microeconomics*, 3 (2), 114–38.

[32] Ekman, P. (1970). Universal facial expressions of emotion. *California Mental Health Research Digest*, 8, 151-158.

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

[34] Farrell, J., and Rabin, M. (1996). Cheap talk. *Journal of Economic Perspectives*, 10 (3), 103–118.

[35] Fischbacher, U., and Föllmi-Heusi, F. (2013). Lies in disguise–an experimental study on cheating. *Journal of the European Economic Association*, 11(3), 525–547.

[36] Friedman, J. (2002). Stochastic Gradient Boosting.*Computational Statistics and Data Analysis* 38 (4), 367–378.

[37] Glaeser, E. L., Hillis, A., Kim, H., Kominers, S. D., and Luca, M. (2021). Decision authority and the returns to algorithms. Harvard Business School Working Paper.

[38] Gneezy, U. (2005). Deception: The role of consequences. *American Economic Review*, 95(1), 384–394.

[39] Gneezy, U., Rockenbach, B., and Serra-Garcia, M. (2013). Measuring lying aversion. *Journal of Economic Behavior & Organization* 93, 293–300.

[40] Gneezy, U., Saccardo S., Serra-Garcia, M., and van Veldhuizen R. (2020). Bribing the self. *Games and Economic Behavior*, 120, 917–946.

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

[42] Holt, C.A., and Smith, A.M. (2009). An update on Bayesian updating. *Journal of Economic Behavior & Organization* 69 (2), 125–134.

[43] Hu, A. and Ma, S. (2020). Human Interactions and Financial Investment: A Video-Based Approach. Working paper.

[44] Kamenica, E. (2019). Bayesian Persuasion and Information Design. *Annual Reviews of Economics* 11, 249–272.

[45] Kapoor, S., Cantrell, E., Peng, K., Pham, T. H., Bail, C. A., Gundersen, O. E., Hofman, J.M., Hullman, J., Lones, M.A., Malik, M.M., Nanayakkara, P., Poldrack, R.A., Raji, I.D., Roberts, M., Salganik, M.J., Serra-Garcia, M., Steward, B.M., Vandewiele, G., and

Narayanan, A. (2023). REFORMS: Reporting Standards for Machine Learning Based Science. arXiv preprint arXiv:2308.07832.

[46] Karni, E. (2009). A mechanism for eliciting probabilities. *Econometrica* 77, 603–606

[47] Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2018). Human Decisions and Machine Predictions. *Quarterly Journal of Economics*133(1) 237–293.

[48] Konrad, K., Lohse, T., and Qari, S. (2014). Deception choice and self-selection: The importance of being earnest. *Journal of Economic Behavior & Organization*, 107, 25–39.

[49] Logg, J.M., Minson. J.A., and Moore, D.A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151, 90–102.

[50] Loijens, L., Krips, O., Grieco, F., van Kuilenburg, H., van den Uyl, M., and Ivan, P. (2016). *FaceReader: Tool for automatic analysis of facial expressions*. Reference Manual Version 7, Noldus Information Technology, the Netherlands.

[51] Longoni, C., Bonezzi, A. and K Morewedge, C. (2018). Consumer Reluctance Toward Medical Artificial Intelligence: the Underlying Role of Uniqueness Neglect. in *NA - Advances in Consumer Research* 46, eds. Andrew Gershoff, Robert Kozinets, and Tiffany White, Duluth, MN. Association for Consumer Research, 63–67.

[52] Möbius, M. M., Niederle, M., Niehaus, P., and Rosenblat, S. R. (2022). Managing Self-Confidence: Theory and Experimental Evidence. *Management Science*, 68 (11), 7793-8514.

[53] Mynatt, C. R., Doherty, M. E., and Tweney, R. D. (1977). Confirmation bias in a simulated research environment: An experimental study of scientific inference. *Quarterly Journal of Experimental Psychology*, 29(1), 85–95.

[54] Nickerson, R. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology*, 2, 175–220.

[55] Ockenfels, A., and Selten, R. (2000). An Experiment on the Hypothesis of Involuntary Truth-Telling in Bargaining. *Games and Economic Behavior* 33 (1), 90–116.

[56] Peer, E., Rothschild, D., Gordon, A., Evernden, Z., and Damer, E. (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, 1.

[57] Pennycook, G., Bear, A., Collins, E. T., and Rand, D. G. (2020). The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science*, 66(11), 4944-4957.

[58] Promberger, M., and Baron, J. (2006). Do patients trust computers? *Journal of Behavioral Decision Making*, 19(5), 455–468.

[59] Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review*, 1281–1302.

[60] Schotter, A., and Trevino, I. (2014). Belief Elicitation in the Laboratory, *Annual Review of Economics* 6, 103–128.

[61] Saccardo, S. and Serra-Garcia, M. (2023). Cognitive Flexibility or Moral Commitment? Evidence of Demand for Moral Commitment. *American Economic Review* 113 (2), 396–429.

[62] Serra-Garcia, M., and Gneezy, U. (2021). Mistakes, Overconfidence and the Effect of Sharing on Detecting Lies. *American Economic Review* 111 (10), 3160–3183.

[63] Stevenson, M., and Doleac, J.L., 2021. Algorithmic Risk Assessment in the Hands of Humans. SSRN Working Paper, https://ssrn.com/abstract=3489440

[64] Turmunkh, U., van den Assem, M. J., and van Dolder, D. (2019) Malleable Lies: Communication and Cooperation in a High Stakes TV Game Show. *Management Science* 65 (10):4795–4812.

[65] van de Assem, M.J., van Dolder, D., and Thaler, R. H. (2012) Split or Steal? Cooperative Behavior when the Stakes Are Large. *Management Science*, 58 (1), 2–20.

[66] van Leeuwen, B., C. N. Noussair, T. Offerman, S. Suetens, M. van Veelen, and J. van de Ven (2018). Predictably Angry–Facial Cues Provide a Credible Signal of Destructive Behavior. *Management Science* 64 (7), 3352–3364.

[67] Wismüller, A., and 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. In *Medical Imaging 2020: Biomedical Applications in Molecular, Structural, and Functional Imaging*, 11317, 144–150.

Online Appendix

Improving Human Deception Detection Using Algorithmic Feedback

by Marta Serra-Garcia and Uri Gneezy

August 2023

Contents

A Data Sources **3**
A.1 Golden Balls . . . 3
A.2 Instructions . . . 3

B Facial, Audio and Speech Analysis **11**
B.1 Facial Analysis: FaceReader . . . 11
B.2 Speech Analysis . . . 12
B.3 Voice Analysis . . . 13

C Machine-learning Methods **15**
C.1 Developing Prediction Functions . . . 15
C.2 Estimation Details and Settings for GBM . . . 15
C.3 Estimation Approach for Rigorous Logistic Lasso . . . 17
C.4 Results for Rigorous Logistic Lasso . . . 18
C.5 Comparison to logit regression . . . 19

D Additional Results **20**
D.1 Descriptive Statistics . . . 20
D.2 Experiment 1: Additional Results . . . 23
D.3 Experiment 2: Additional Results . . . 25
D.3.1 Text Analysis of Participants' Reported Use of Flags . . . 25
D.4 Contestant Features, Beliefs, Predictions & Stealing . . . 27
D.4.1 Influence and Partial Dependence Plots for GBMs . . . 27
D.4.2 Correlation between Predictions and Features . . . 28
D.4.3 Lasso . . . 31
D.5 Nonverbal Information Treatment: Results on Accuracy . . . 32
D.6 Additional Experiment: Flagging and Delegation Decisions . . . 33
D.6.1 Experimental Design . . . 33
D.6.2 Results . . . 34

E Pre-registrations **40**

F REFORMS Checklist **43**

A Data Sources

Appendix A provides details on the sources of the Golden Balls episodes (Section A.1) and the experimental instructions (Section A.2).

A.1 Golden Balls

The episodes from Golden Balls were shared with us by Donja Darai for research purposes. She had previously used this TV show to study attractiveness and cooperation (Darai and Gratz, 2013). The DVDs we obtained included all episodes from Season 1 (40 episodes) and Season 2 (60 episodes). We obtained 78 episodes out of 80 in Season 3 and 42 out of 65 aired in Season 4. We focus on all episodes we received, but exclude three episodes in which the prize to be split was less than 4 GBP (1.10 and 4.24), and in which one contestant claimed an intention to steal but split after the show (1.09), as done in Turmunkh et al. (2019). For 4 contestants, emotions could not be detected by the facial analysis software and are thus excluded (Participant IDs 98, 147, 305, 412). This implies that we analyze the behavior of 76 contestants in Season 1 (since 2 episodes are excluded), for 118 contestants in Season 2 (since the emotions of 2 contestants could not be detected), for 155 contestants in Season 3 (since the emotions of 1 contestant could not be detected), and for 81 contestants in Season 4 (since the emotions of 1 contestant could not be detected). The total number of contestants is 430.

A.2 Instructions

Below are the instructions presented to participants on Prolific Academic and in the laboratory at UCSD, via Qualtrics surveys. The instructions are shown for the no learning treatment in Experiment 1, with differences with the learning treatment indicated in brackets. The additions, including flagging, for Experiment 2 are also shown in brackets. Participants were asked to guess the chance that the contestant would split or steal (which one was randomized across participants). The instructions are shown for participants guessing 'steal.'
The experiment always began with a CAPTCHA verification question and with an audio transcription question to verify that participants could listen to audio.

Instructions

- In this study, you will see 20 videos.
- In each video, you will see 1 contestant in a TV show.
- Each contestant made a SPLIT or STEAL decision.

- You will watch the conversation between the two contestants before they made their SPLIT or STEAL decision.

page break

How does the split or steal decision work?

After 3 rounds of play in the TV show, in the final round:

- Each of the 2 remaining contestants is presented with 2 golden balls, one with the word "split" and the other with the word "steal" written inside.
- The 2 contestants simultaneously have to choose either the split or the steal ball.

Consequences for the split or steal decision:

- If both decide to split, they split the jackpot equally.
- If one decides to split while the other decides to steal, the one who steals receives the entire jackpot and the one who splits goes home with nothing.
- If both decide to steal, both go home with nothing.

In what follows you will watch a brief example video, in which the TV show host explains the rules, and you will be asked to answer several questions. Please listen to the rules carefully.

page break

The SPLIT or STEAL decision

Below is an example episode in which the host of the TV show explains the rules. The jackpot amount varies in each episode.

[VIDEO OF PRESENTER EXPLAINING RULES]

Please answer the following questions carefully. There is only 1 correct answer in each question.

IF YOU FAIL TO ANSWER THE 3 QUESTIONS BELOW CORRECTLY, YOUR SUBMISSION WILL NOT BE APPROVED.

1. What happens if both contestants choose to SPLIT?

- They both go home with nothing

- They both share the jackpot equally
- They both share a chance to participate in a future episode of the show

2. What happens if both contestants choose to STEAL?

- They both go home with nothing
- They both share the jackpot equally
- They both share a chance to participate in a future episode of the show

3. What happens if one contestant STEALS and the other SPLITS?

- The one who STEALS goes home with nothing, the one who SPLITS gets the whole jackpot
- The one who STEALS gets disqualified, the one who SPLITS gets nothing
- The one who STEALS gets the whole jackpot, the one who SPLITs gets nothing

page break

Guessing task

You will watch 20 videos of the conversation before the split or steal decision.

After watching each video please carefully consider whether you think that the contestant you saw on the video chose Split or Steal.

We will ask you to guess how likely it is that the contestant chose to steal. You will provide your guess as a chance, on a scale from 0 to 100.

Below you see an example. Please move the slider from 0 to 100 to see how it works.

How likely do you think it is that the contestant on the left/right of the screen chose to steal?

[SLIDER (0 to 100)]

page break

Your bonus will depend on your guess. Specifically, your bonus will be calculated as follows:

The computer will randomly select a number R from 1 to 100.

- If the number the computer randomly selected is less than or equal to the number you selected, you will receive the $5 bonus if the contestant indeed chose to steal, and zero otherwise.
- However, if the number the computer randomly selected is greater than the number you selected, you will participate in a lottery. In this lottery, you will receive the $5 bonus with a chance exactly equal to the number drawn by the computer.

For example, suppose that you guess that a contestant chose to steal with 60% chance.

- If the number the computer chooses randomly is below 60, then you receive the bonus if the contestant indeed chose to steal, and zero otherwise.
- If the number the computer chooses randomly is above 60, then you receive the bonus with chance R%, and zero otherwise.

To maximize your chance of earning a bonus payment, you should honestly choose your guess that the contestant chose to steal.

Understanding Questions

1. What is your guessing task? After watching each video...
   - ...I will guess how likely it is that the contestant chose to steal
   - ...I will guess how likely it is that the contestant chose to split
   - ...I will guess how likely it is that the number R is above 100
2. My bonus will determined by...
   - ...a random number R, from 1 to 100. Hence the best I can do is guess randomly after each video.
   - ...a random number R, from 1 to 100, and my guess about the chance that the contestant chose to steal. Hence, the best I can do is honestly guess how likely I think it is that the contestant chose to steal.

- ...a random number R, from 1 to 100, and my guess about the chance that the contestant chose to steal. Hence, the best I can do is provide a random number between 0 and 100, to answer how likely I think it is that the contestant chose to steal.

page break

One of your guesses for one video will be randomly selected for payment, and your bonus will be determined by the decision you made. Since any choice you make may be selected, please make your decisions carefully.

page break

[Experiment 2 :
Additional Information For Your Guessing Task

We have used contestants' facial expressions and speech in the videos to train a standard machine-learning algorithm to predict when a contestant will choose split or steal.
We used this algorithm to "flag" four videos for which the algorithm either predicted that the contestant chose to split with very high chance or very low. [Flag-Before: This flagging will be shown as a message that informs you that the algorithm made such a prediction before you watch the video and provide your guess.] [Flag-After: This flagging will be shown as a message that informs you that the algorithm made such a prediction after you watch the video and think about your guess. After seeing the flag, you can submit your guess.]

Understanding Question
Which of the following is true?

- I will be provided with predictions from an algorithm trained to predict with which likelihood the contestant will [split/steal] for all videos.
- I will be informed if the video is "flagged" if the algorithm predicted that the contestant chose to split with a very high or a very low chance, for some videos.
- For every video, I will be informed if the video is "flagged" if the algorithm predicted that the contestant chose to split with a very high or a very low chance.]

page break

You are now going to watch the 20 videos. [Experiment 1 – Learning Treatment: Each time, after providing your guess, you will learn whether the contestant chose to split or to steal.]

page break

[For each of 20 randomly selected contestant videos:]

[Flag-Before: If video was flagged:
For this upcoming video: "The algorithm predicted that the contestant on the [right/left] of the screen will [split/steal] with a [very high/very low] chance."]

page break

[GOLDEN BALLS CONTESTANT VIDEO]

[Flag-After:

page break

If video was flagged:
"The algorithm predicted that the contestant on the [right/left] of the screen will [split/steal] with a [very high/very low] chance."
]

How likely do you think it is that the contestant on the left/right of the screen chose to steal?

[SLIDER (0 to 100)]

page break

[Experiment 1 – Learning Treatment: The contestant in the previous video chose to [split/steal].]

page break

After rating all 20 videos:
You have now seen all 20 videos.

We will now ask you some more questions about the videos. You can earn an **additional bonus** if you answer correctly.

In these questions, we will ask you about how well you think you did when guessing whether a contestant choose to steal or split, in the 20 videos.

When we ask you about whether your guess is correct or not, it means the following. If you guessed that the contestant chose to steal with more than a 50% chance, and the contestant

chose to steak, your answer is correct. Similarly, if you guessed that the contestant chose toÂ steal with less than a 50% chance, and the contestant chose to split, your answer is correct.

page break

[Experiment 1 – No Learning Treatment only and Experiment 2:
How many of the 20 guesses (for 20 videos) you just made do you believe are correct? If the number you choose is correct, you will earn an **additional BONUS of $1.**

[SLIDER (0 to 20)]]

page break

Compared with previous participants in this experiment, how well do you think you could guess whether a participant chose to steal or split? We ask you to choose a quartile. If you choose the correct one, you will earn an **additional bonus of $1.**

- **Quartile 4:** 75th-100th percentile (better than at least 75% of participants).
- **Quartile 3:** 50th-75th percentile
- **Quartile 2:** 25th-50th percentile
- **Quartile 1:** 0th-25th percentile (worse than at least 75% of participants)

page break

[Experiment 2 Flag-Before and Flag-After treatments:

In this question we ask you to guess how accurate the algorithm is. Consider 10 videos flagged by the algorithm, which predicted that in these videos contestants will split/steal with a very high chance.

How accurate do you guess the algorithm will be? That is, how many of the 10 contestants in the video will actually choose to split/steal? If the number you choose is correct, you will earn an **additional BONUS of $1.**

[SLIDER (0 to 10)]

page break

In this question we ask you to guess how accurate the algorithm is. Consider 10 videos flagged by the algorithm, which predicted that in these videos contestants will split/steal with a very low chance.

How accurate do you guess the algorithm will be? That is, how many of the 10 contestants in the video will actually choose to steal/split? If the number you choose is correct, you will earn an **additional BONUS of $1.**

[SLIDER (0 to 10)]]

page break

Very short questionnaire

1. What is your gender?
   - Male
   - Female
2. What is your age?
3. Were you familiar with the TV show before this study?
   - Yes, I had seen (parts of) its episodes in the past
   - No, I had never seen this show before

4. [Experiment 1: Please describe in 1-2 sentences how you made your decisions.]

4. [Experiment 2 - Control: Please describe in 1-2 sentences how you made your guesses.]

5. [Experiment 2 - Flag-Before and Flag-After: Please describe in 1-2 sentences how you made your guesses when you saw a flag for the video.]

B Facial, Audio and Speech Analysis

B.1 Facial Analysis: FaceReader

To perform the analysis of emotions and facial expressions we use in FaceReader. We first started by converting all DVDs and cutting the final conversation prior to the split or steal decision (last round of play in each episode) for analysis. We refer to this cut as the video or conversation video.

Next, each video was edited to remove all shots that did not only display the contestant of interest. The removed shots included shots of the audience, shots of multiple faces including the host, as shown in Figure B1. The reason these shots were removed is that they are not easily separated by FaceReader from the contestant's shots, and our aim was to analyze contestant facial expressions as cleanly as possible.

(a) Audience

(b) Multiple faces

(c) Host

Figure B.1: Shots removed for individual analysis of contestants

FaceReader analyzes 10 frames per second of video. It allows several options for the analysis of facial expressions. We followed the recommendations given by the software provider. We used every other frame as the sample rate, which speeds up analysis without an effect on output. We also used a continuous calibration for facial expressions, which attempts to correct for individual-specific biases in facial expressions, and smoothed classifications of emotions, which considers time between frames when calculating emotions. We gave age and gender information to FaceReader to maximize precision of the analysis. Figure 1 in the body of the paper shows an example of the software analyzing a contestant in the TV show.

Facereader is best able to analyze facial expressions on straight ahead faces with proper lighting. The software is known to not be very good at reading non-Caucasian faces, children's faces, or faces over the age of 65. A large majority of the Golden Balls contestants are Caucasian and between the ages of 18-65. The same Facereader model, the General model, was used for all contestants.

The videos from Golden Balls fall outside the optimal (lab) settings in which FaceReader can be run. Hence, for several contestants there are missing frames that could not be analyzed by the software, and in some cases no frames at all. On average, 56.5% of the frames for each contestant could be read and their emotions analyzed by FaceReader.

FaceReader measures the six basic or universal expressions as classified by Ekman (1970).

Each emotion is assigned a value between 0 and 1 in each frame of the video analyzed. In addition to measuring the six basic emotions, Facereader also provides a measure of valence. Valence indicates whether the emotional status of the contestant is positive or negative. 'Happy' is the only positive emotion. 'Sad', 'Angry', 'Scared' and 'Disgusted' are considered to be negative emotions. 'Surprised' can be either positive or negative. The valence is calculated as the intensity of 'Happy' minus the intensity of the negative emotion with the highest intensity. For instance, if the intensity of 'Happy' is 0.8 and the intensities of 'Sad', 'Angry', 'Scared' and 'Disgusted' are 0.1; 0.0; 0.05 and 0.05, respectively, then the valence is 0.7.

FaceReader also measures head, mouth, eye, gaze, and eyebrow movements. Specifically, FaceReader calculates the x-, y-, and z-head orientation (in degrees) of the individual in each frame. It measures whether the mouth is open or closed, or whether its position is unknown. We define indicator variables for whether the mouth is open, closed, or its position is unknown. FaceReader similarly measures whether both eyes are opened, closed, in different position or the position of at least one is unknown. FaceReader measures whether the eyes gaze left, right, forward or gaze is unknown, and the contestant's eyebrow movements, where we define indicator variables that take value one if both eyebrows are raised, lowered, neutral, in different positions or unknown. The software also provides a measure of quality (of the image) for each frame, which is between 0 and 1.

The sample for analysis focuses on 430 contestants, for whom FaceReader analyses of emotional states could be conducted for at least one frame. In the main analyses, all measures for a given contestant are averaged throughout the clip.

B.2 Speech Analysis

We use transcripts from the words that each contestant said during their conversation prior to the split or steal decision.

We focus on two basic features of speech, word count and sentiment score, provided by a widely used package in R (sentimentr). Word count is the number of words that the contestant says. Sentiment scores are calculated in sentimentr package in R (Rinker, 2018). This package differs from standard methods based on dictionary lookups (Bing, NRC and Afinn methods) in that it takes into account valence shifters (including negators such as 'not') and de-amplifiers (words such as 'hardly'). Additionally, we use the classification of promises by Turmunkh et al. (2019), into explicit or implicit and conditional or unconditional promises.

B.3 Voice Analysis

We examine the role of voice features, and their predictive power for split and steal decisions. We briefly explain the measures we considered in what follows.

Excellent introductions to sound analysis are provided in a variety of sources. We borrow heavily from Zhang (2019, Chapter 6) here to provide an overview of the voice analysis we conduct. Humans have vocal cords that vibrate when air from our lungs passes through them. This vibration produces vocal sounds. A sound wave is a transfer of energy as it travels away from a vibrating source. Waves are characterized by their amplitude, their length and frequency. Frequency is measured in Hertz (HZ) where 1 Hertz is 1 vibration per second.

A wavelength is the length of one cycle of sound and the inverse of the frequency. Longer wavelengths have a lower pitch. Amplitude specifies the sound's loudness. A low amplitude will produce a soft sound and a higher amplitude will produce a louder sound. Pitch and loudness are different from each other. The pitch of a sound depends on the frequency, while the loudness of a sound depends on the amplitude of sound waves. Pitch and intensity (loudness) are illustrated in Figure B.2.

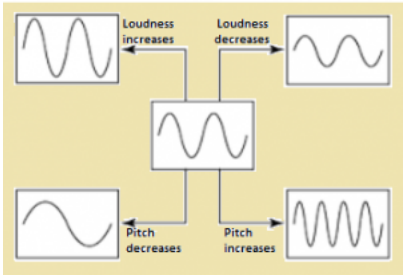


Figure B.2: Sound waves: pitch and intensity (Zhang, 2019)

We analyze the voice of contestants in Golden Balls using Praat (Boersma and Weenink, 2020; see also, the soundgen package in R by Anikin, 2020). We take the entire conversation between two contestants and sequentially mute the opponent, to analyze each contestant separately. The basis of sound analysis is the short-time Fourier transform (STFT) and the software analyzes one short segment of sound at a time (one STFT frame). STFT is based on Discrete Fourier Transformations (DFT), which decompose a time series into a sum of finite series of sine or cosine functions, which have a specified frequency and a relative amplitude. In this way, a DFT allows us to switch from the time domain to the frequency domain (a detailed introduction is provided by Sueur, 2020). An STFT computes a DFT on each slide or jump of the sound (signal).

We focus on two key and intuitive descriptive measures of the sound of the contestant's voice: pitch and loudness using intensity. The intensity of a wave is the power per unit area carried by the wave. The intensity of a sound is proportional to its amplitude squared and

it is measured in decibels. A more intense sound has larger amplitude oscillations, and it is an approximate measure of the loudness of a contestant's voice. Pitch is salient to listeners but difficult to measure accurately (Anikin, 2020). Pitch is the fundamental frequency of each sound wave. The software calculates pitch based on the autocorrelation method. We measure the mean value of pitch (in Hz) over all the frames in which the contestant speaks.

Hence, our measures of voice characteristics provide basic descriptive measures of how a contestant's voice sounded in their conversation.

C Machine-learning Methods

C.1 Developing Prediction Functions

We are interested in predicting Y, an indicator for the contestant's decision to steal. The input variables, X, consist of individual characteristics or features of the contestant, during her conversation (detailed below).

We estimate GBMs to predict the likelihood of stealing. Given that stealing is a binary variable, we use the Bernoulli distribution and set as an objective to minimize binomial deviance (or cross-entropy) loss function:

$$L(y, f(x)) = \log[1 + \exp^{-2yf(x)}]$$

where $y \in \{-1, 1\}$ (Hastie et al., 2008 Chapter 10, page 346). This approach estimates boosted regression trees, which divide the predictor space into distinct non-overlapping regions. The boosting approach learns "slowly." It starts with simply fitting the training data, with all observations in the training set receiving equal weight. However, with each successive iteration, the observation weights are modified, and those observations that were misclassified receive a higher weight. We use the standard package "gbm" in R (Ridgeway, 2020) and tune the model hyperparameters,[1] controlling for the learning rate, the interaction depth in the tree, and the fraction of observations used in each iteration, using fivefold cross-validation.

We also analyze which features the GBM uses to predict stealing. To predict, GBMs split the data according to different covariates. Once the algorithm calculates the best predictive tree ensemble, we can calculate the increase in accuracy achieved by splitting on each feature. This metric is referred to as the "relative influence" and provides a measure of the reduction in error risk that is achieved by including it in the splits of the tree (Hastie et al., 2008). A higher value of influence for a covariate means it is more important in generating a prediction by the predictive model. We explore this measure to contrast the cues correlated with GMB predictions and those correlated with participant beliefs in our experiments.

C.2 Estimation Details and Settings for GBM

Gradient boosting combines both classification and regression techniques (Kuhn and Johnson, 2013). Broadly speaking, given a loss function (binary deviance in our case) and a weak learner (regression trees), the algorithm looks for an additive model that minimizes the loss function.

The basis of gradient boosted trees are classification trees which partition the space into non-overlapping regions R_j, where $j = \{1, 2, ..., J\}$, representing the terminal nodes of the

[1] A useful practical guide is provided here.

tree (Hastie et al., 2008). For each region, a constant γ_j is defined such that the predictive model assigns an observation in that region j the value γ_j: $x \in R_j \Rightarrow f(x) = \gamma_j$. A tree is written as:

$$T(x, \Theta) = \sum_{j=1}^{J} \gamma_j I(x \in R_j),$$

with parameters $\Theta = \{R_j, \gamma_j\}_1^J$, where J is a tuning parameter controlling the number of regions or splits in each tree. A boosted tree model is a sum of M trees,

$$f_M(x) = \sum_{m=1}^{M} T(x; \Theta_m)$$

induced in a forward stagewise manner, which approximates a solution to the optimization problem sequentially by following the steps outlined in Hastie et al. (2008, Algorithm 10.3). The steps of the algorithm are the following:

1. Initialized $f_0(x) = \arg\min_\gamma \sum_{i=1}^{N} L(y_i, \gamma)$
2. For $m = 1$ to M:
   a) For $i = \{1, 2, ..., N\}$ compute

$$r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}}.$$

   b) Fit a regression tree to the targets r_{im} given terminal regions R_{jm}, with $j = \{1, 2, ..., J_m\}$.
   - For $j = \{1, 2, ..., J_m\}$ compute

$$\gamma_{jm} = \arg\min_\gamma \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma).$$

   c) Update $f_m(x) = f_{m-1}(x) + \nu \cdot \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$.
3. The output is $\hat{f}(x) = f_M(x)$.

The first step states that the algorithm starts with a single terminal node tree, finding the optimal constant model. In each iteration m, for each observation, the negative gradient is computed (step 2a) and referred to as pseudo residuals (r). Then, in each iteration, a new regression tree is fit to add to the current existing trees. The contribution of each tree is given a factor $0 < \nu < 1$, referred to as shrinkage, when it is added to the current set of trees.

The interaction level of the regression trees is limited by the tree size J. In tuning the boosted tree parameters, we consider two levels $J = 1$, an additive model, and $J = 2$ a model in which two-variable interaction effects are also allowed. We also set the minimum number of observations in each terminal node, such that the resulting regions are based on enough observations. We explore a minimum of 5, 10, 20 or 25 observations in each terminal node. We explore different values of the shrinkage parameter (or learning rate): 0.1, 0.2, 0.3 and 0.4. Since the number of observations in the training data is limited and we want to avoid overfitting, we also introduce subsampling. This implies that in each iteration only a fraction η of the training observations is used to grow the next tree. We explore values of 0.5, 0.6 and 0.7 for the subsampling parameter.

To reduce the risk of overfitting, we include in the model the following covariates regarding the contestant. First, two contestant characteristics: age and gender. Second, several measures of her facial expressions: the contestant's emotions, as measured by the average value of six emotions (excluding neutral), and gaze movements (how often the contestant gazes left, right, forward and how often gaze is unknown). Since quality may also matter, we also included three measures of recordings (quality measured by FaceReader between 0 and 1, the number of frames in the recording, and the share of analyzed frames). Third, several measures of speech: whether the jackpot was mentioned and, if so, the jackpot amount, whether the contestant makes a conditional or unconditional, implicit or explicit promise to split, her word count and sentiment score. Fourth, two simple measures of the contestant's voice during her conversation with the other contestant: pitch and intensity of her voice.

We train 100 trees ($M = 100$), use 5-fold cross-validation on the training sample, to determine the optimal interaction level of the model ($J = 2$), the minimum number of observations in each node (10), the learning rate ($\nu = 0.2$), and the subsampling parameter ($\eta = 0.7$).

C.3 Estimation Approach for Rigorous Logistic Lasso

We also estimate regularized logistic regression models with rigorous penalization (see Belloni et al, 2016; Ahrens et al., 2020) to predict the likelihood of stealing. Penalized regression methods fit models with all p predictors using a technique that regularizes the coefficient estimates, shrinking them towards zero (Tibshirani, 1996). Denote by y_i the decision to steal, which takes value 1 if the contestant steals and 0 if she splits. The vector of predictors is x_i, and the vector of parameters is β. Given N contestants, the logistic lasso has as an objective to maximize the penalized log-likelihood:

$$\frac{1}{N}\sum_{i=1}^{N} y_i(\beta_0 + x_i'\beta) - \log(1 + e^{(\beta_0 + x_i'\beta)}) - \frac{\lambda}{N}||\beta||_1$$

where λ is a tuning parameter and $||\beta||_1$ is the $\ell(1)$ vector norm.

With rigorous penalization, the value of λ is theory driven (Belloni et al, 2016) with $\lambda = \frac{c}{2}\sqrt{N}\Phi^{-1}(1-\gamma)$, where c is a slack parameter, Φ is the standard normal CDF and γ is the significance level. Following Belloni et al. (2016), we set $c = 1$ and $\gamma = \frac{0.1}{p\log(N)}$. Alternatively, the value of λ may be selected via cross-validation. We obtain qualitatively similar results using 5-fold cross-validation as an alternative approach.

C.4 Results for Rigorous Logistic Lasso

We estimate two types of models: a "lasso simple" model and a "lasso long" model. The difference between them is that the lasso simple model only includes the average value of the emotions of the contestant during her conversation prior to the split or steal decision, while the lasso long model also includes the standard deviation, the minimum and the maximum of the contestant's emotions during the conversation. Minor variations in the covariates included do not significantly affect predictive accuracy.

Table C.1: Rigorous Logistic Lasso - Coefficient Estimates

	(1)	(2)
	Likelihood of stealing	
Selected covariates	lasso simple	lasso long
Age	-0.015	-0.015
Happy	-0.271	-0.151
Max of angry		0.612
Max of disgusted		0.702
Gaze left	0.847	0.858
Implicit unconditional promise	-0.039	-0.041
Explicit unconditional promise	-0.356	-0.375
Constant	0.554	0.371

The AUC of the lasso simple model is 0.70 (95% CI is 0.60,0.79). The fraction of correct predictions, using 50% as the threshold, is 66.41%. The correlation between the model's predictions and actual stealing is significant, though the coefficient is 1.86, which is significantly higher than 1 (p-value= 0.003). Hence, a change in the predicted steal rate of 1 percentage point is associated with more than 1 percentage point increase in the likelihood of stealing. This suggests that the model's predictions may not be sensitive enough to capture the magnitudes of changes in steal risk. Similarly, the AUC of the lasso long model is 0.70 (95% CI is 0.60,0.80). The fraction of correct predictions, using 50% as the threshold, is 68.75%. The correlation between the model's predictions and actual stealing is again significant, though as in the lasso simple model the coefficient is 1.896, which is significantly higher than 1 (p-value= 0.002). The selected coefficients with each model are shown in Table C.1.

C.5 Comparison to logit regression

We also examine the accuracy of logit regression models in predicting out of sample. We run a "simple" logit regression, without a penalty term, on the training sample. Then we predict on the test sample, and examine the accuracy of predictions. The simple logit model yields an AUC of 0.69 (95% CI is 0.59, 0.78). The fraction of correct predictions, using 50% as the threshold, is 65.63%. We also estimate a "long" logit regression, including the standard deviation, the minimum and the maximum of the contestant's emotions during the conversation. The long logit model yields an AUC of 0.67 (95% CI is 0.57, 0.77). The fraction of correct predictions, using 50% as the threshold, is 65.63%.

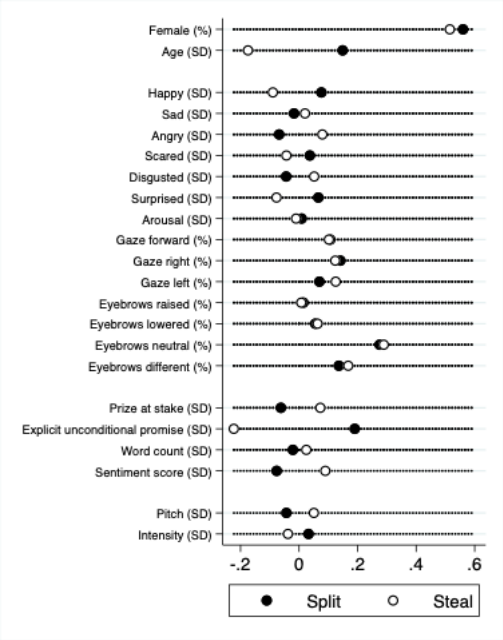
D Additional Results

In this section, we present additional analyses to complement those presented in the body of the paper.

D.1 Descriptive Statistics

Figure D.1 below shows summary statistics for contestants in the entire dataset, comparing those who split and those who steal. All continuous variables are standardized.

Figure D.1: Contestant characteristics and behavior by decision (split or steal)



Notes: This figure shows the average value of each covariate in standard deviations (for continuous variables, standardized for all videos) and in rates (for variables that range 0 to 1) by the contestant's decision to split or steal. All videos of contestants are included (N=430).

Table D.1: Descriptive Statistics of Contestants, by Steal Decision

	Split ($N = 232$)	Steal ($N = 198$)	t-test p-value
Contestant characteristics			
Female	0.56	0.515	0.35
Age	37.948	34.171	0.001
Nonverbal features			
Emotions			
Happy	0.201	0.176	0.081
Sad	0.099	0.102	0.722
Angry	0.045	0.051	0.141
Surprised	0.159	0.141	0.130
Scared	0.079	0.074	0.389
Disgusted	0.026	0.029	0.336
Arousal	0.524	0.521	0.844
Facial movements			
Mouth open	0.273	0.284	0.614
Mouth closed	0.2	0.236	0.112
Mouth unknown	0.528	0.48	0.102
Both eyes open	0.3	0.315	0.553
Both eyes closed	0.078	0.089	0.402
Eyes different	0.103	0.124	0.106
Eyes unknown	0.518	0.472	0.115
Gaze forward	0.106	0.102	0.718
Gaze left	0.07	0.125	0.001
Gaze right	0.141	0.124	0.368
Gaze unknown	0.683	0.649	0.145
Both eyebrows raised	0.014	0.008	0.109
Both eyebrows lowered	0.056	0.063	0.573
Both eyebrows neutral	0.275	0.289	0.606
Eyebrows different	0.136	0.168	0.051
Eyebrows unknown	0.519	0.472	0.112
Video features			
Quality	0.789	0.812	0.018
Nonmissing frames	0.54	0.591	0.074
Total frames	198.991	208.692	0.333
Speech features			
Jackpot (in GBP)	12246.913	14848.652	0.166
Explicit unconditional promises	0.586	0.298	0.000
Explicit conditional promises	0.065	0.091	0.388
Implicit unconditional promises	0.935	0.722	0.028
Implicit conditional promises	0.401	0.409	0.894
Sentiment score	0.074	0.101	0.083
Word count	55.81	57.379	0.635
Voice features			
Pitch	210.235	214.343	0.335
Intensity	39.981	38.865	0.469
Observations			430

Notes: This table presents the mean of each contestant characteristic or feature, separated by split and steal decision. The p-value of a t-test comparing those who steal and those who split, based on a linear regression of the decision on the characteristic or feature, is presented in the last column.

Table D.2: Descriptive Statistics of Contestants, by Training vs. Test Set

	Training Set ($N = 302$)	Test Set ($N = 128$)	t-test p-value
Steal	0.474	0.430	0.404
Contestant characteristics	0.526	0.570	0.404
Female	36.567	35.363	0.332
Age	0.940	0.953	0.583
Nonverbal features			
Emotions			
Happy	0.190	0.189	0.945
Sad	0.099	0.106	0.420
Angry	0.049	0.045	0.349
Surprised	0.149	0.156	0.613
Scared	0.076	0.078	0.819
Disgusted	0.028	0.025	0.179
Arousal	0.521	0.526	0.766
Facial movements			
Mouth open	0.275	0.285	0.686
Mouth closed	0.225	0.197	0.249
Mouth unknown	0.501	0.518	0.585
Both eyes open	0.314	0.289	0.361
Both eyes closed	0.078	0.096	0.258
Eyes different	0.115	0.107	0.538
Eyes unknown	0.492	0.508	0.610
Gaze forward	0.110	0.091	0.099
Gaze left	0.092	0.103	0.534
Gaze right	0.135	0.129	0.775
Gaze unknown	0.663	0.678	0.558
Both eyebrows raised	0.012	0.010	0.737
Both eyebrows lowered	0.060	0.057	0.813
Both eyebrows neutral	0.287	0.269	0.532
Eyebrows different	0.149	0.156	0.690
Eyebrows unknown	0.492	0.508	0.617
Video features			
Quality	0.799	0.802	0.771
Nonmissing frames	0.571	0.546	0.429
Total frames	203.063	204.391	0.908
Speech features			
Jackpot (in GBP)	13278.648	13837.229	0.789
Explicit unconditional promises	0.430	0.508	0.284
Explicit conditional promises	0.076	0.078	0.955
Implicit unconditional promises	0.831	0.852	0.848
Implicit conditional promises	0.424	0.359	0.308
Sentiment score	0.090	0.079	0.515
Word count	56.619	56.328	0.935
Voice features			
Pitch	209.880	217.426	0.103
Intensity	39.169	40.170	0.556

Notes: This table presents the mean of each contestant characteristic or feature, separated by whether the contestant was assigned to the training or test set. The p-value of a t-test comparing those who are in the test and training set, based on a linear regression of the decision on the characteristic or feature, is presented in the last column.

D.2 Experiment 1: Additional Results

Figure D.2 shows the distribution of participant beliefs in Experiment 1, in the No Learning (Panel A) and Learning (Panel B) treatments.

Figure D.2: Distribution of participant beliefs in Experiment 1

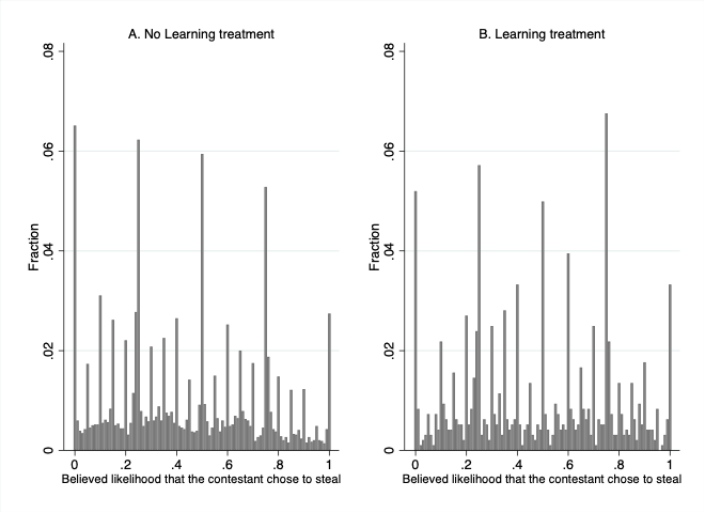


Table D.3: Average Accuracy of Participants and ML Models

	AUC	Fraction correct
No Learning	0.536	0.541
Learning	0.518	0.518
ML Model (GBM)	0.713	0.656
Differences (p-value)		
Effect of Learning	0.126	0.613
No Learning vs. GBM	0.000	0.000
Learning vs. GBM	0.000	0.000

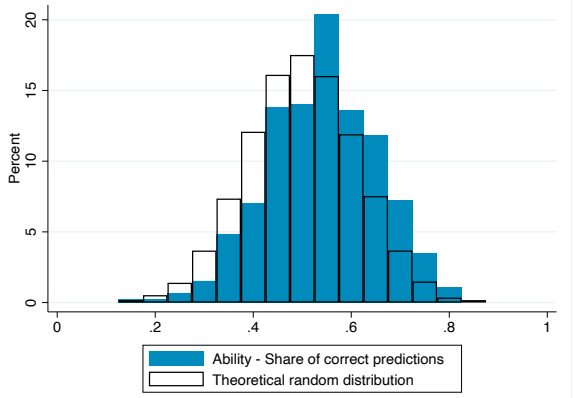
Notes: AUC and the fraction of correct guesses (using 50% threshold) are shown for each treatment and the ML model. Under "Differences (p-value)" p-values of tests of the difference between these measures are shown. When comparing the AUC, we test the equality of the AUC using the test proposed by DeLong et al. (1988) and report the p-value of their proposed χ^2-test. When comparing the fraction of correct guesses and predicted steal rate, we use linear regressions that include contestant fixed effects, with robust standard errors clustered at the participant level (for participants).

Table D.4: Predicted and Actual Steal Risk

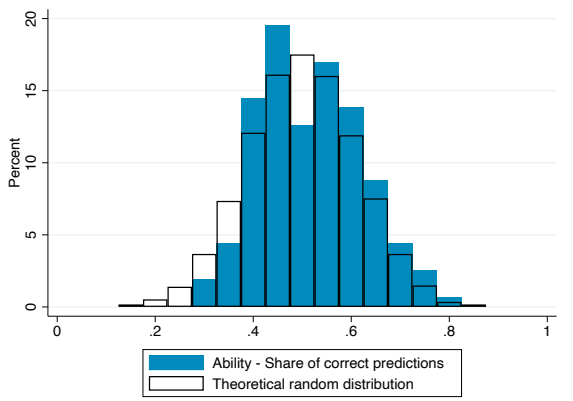
	(1)	(2)	(3)
	Likelihood that contestant steals		
	Participant prediction		
	No Learning	Learning	ML model
Predicted likelihood that contestant steals	0.109***	0.053*	0.964***
	(0.018)	(0.028)	(0.180)
Observations	9,120	3,180	128
Number of clusters	456	159	-

Notes: This table shows marginal effects of probit regressions on the likelihood that a contestant steals. Columns (1)–(2) show the relationship between participants' predicted likelihood of stealing and actual stealing for the participants in the no-learning and learning treatments, respectively. Column (3) shows the relationship between the predicted likelihood of stealing by the ML model and actual stealing. Robust standard errors, clustered at the participant level in columns (1)–(2), shown in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% levels.

Figure D.3: Distribution of accuracy, relative to chance



(a) No Learning Treatment

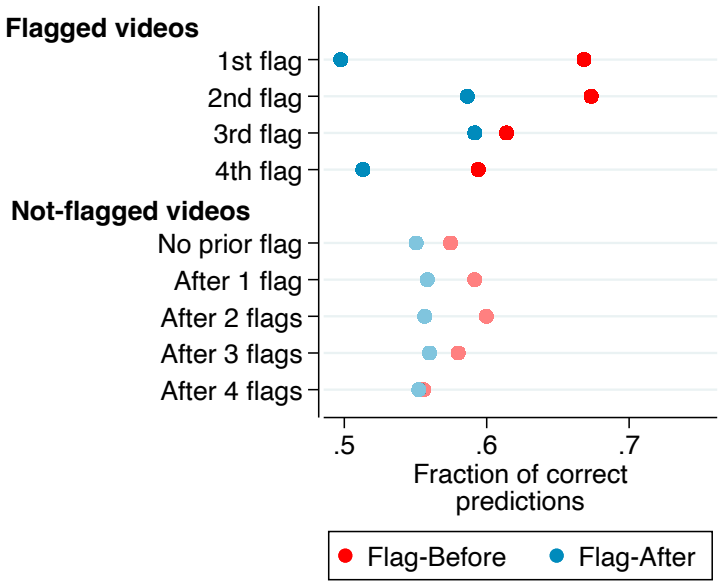


(b) Learning Treatment

D.3 Experiment 2: Additional Results

Figure D.4 shows the average correct predictions of participants in Flag-Before and Flag-After, focusing on videos that were flagged, by order of the flag, and videos that were not flagged, by their order relative to the number of flags shown.

Figure D.4: Accuracy following flags in Flag-Before and Flag-After



Notes: This figure shows the average accuracy of participants' predictions for flagged and not-flagged videos, by their order. For flagged videos, the figure shows the fraction of correct predictions after the 1st, 2nd, 3rd and 4th flags, for Flag-Before and Flag-After. For not-flagged videos, the figure shows the fraction of correct predictions prior to seeing the first flag, after 1 flag, after 2 flags, after 3 flags, and after 4 flags.

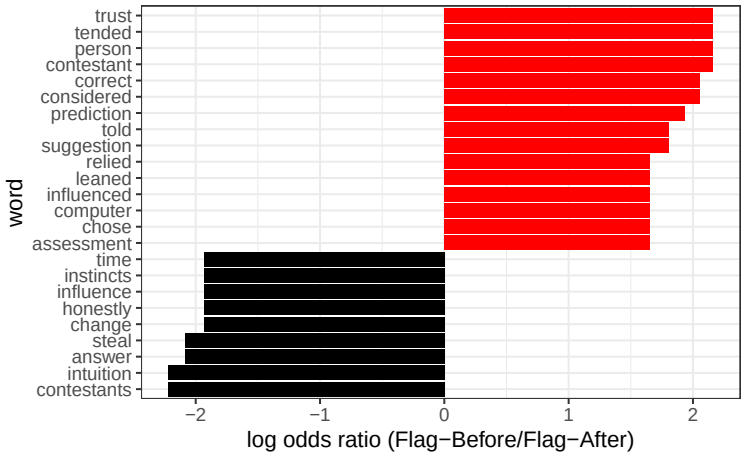
D.3.1 Text Analysis of Participants' Reported Use of Flags

At the end of Experiment 2, participants were asked to report how they made decisions and how they responded to the presence of flags for the videos.[2] In what follows, we report automated text analyses based on what participants wrote. To prepare the analysis, typos were removed and the correct spelling was used.We also used manual coding, as reported in the manuscript, and use these automated analyses to complement the results in the manuscript.

[2]There was a typo in the question, which was "Please describe how you made your guesses when you say [saw] a flag for the video."

Our first analysis calculates the log-odds ratios of words in each treatment, Flags-Before compared to Flags-After. We focus on words that are mentioned at least 5 times by the participants, and plot the words with the 20 largest absolute values of the log-odds ratio, where the odds of a word been mentioned in Flag-Before is compared to the odds of it being used in Flag-After. Figure D.5 shows the words and their log-odds ratios. Words such as "trust", "prediction", "influenced", and "computer" were more likely in Flag-Before, while words such as "instincts" and "intuition" were more likely in Flag-After.

Figure D.5: Word Use in Explaining Use of Flags



Notes: This figure shows the log-odds rations of words in Flags-Before compared to Flags-After. The sample focuses on words that were mentioned at least 5 times by participants.

We also estimate whether there are words that are "discriminating" of each flag timing using logistic lasso regressions. In particular, using rigorous lasso, we find that there are 2 selected (post logit) words that are indicative of Flag-Before are "algorithm" and "tended," while 2 words are indicative of Flag-After, "didn't" and "guess." Alternative estimations using cross validation yield qualitatively similar results. Participants often indicated trust in the algorithm, and that they tended to follow the predictions in flagged videos in Flag-Before. By contrast, in Flag-After, participants often indicated distrust of the algorithm or choosing to ignore it, paying more attention to what they saw in the videos.

D.4 Contestant Features, Beliefs, Predictions & Stealing

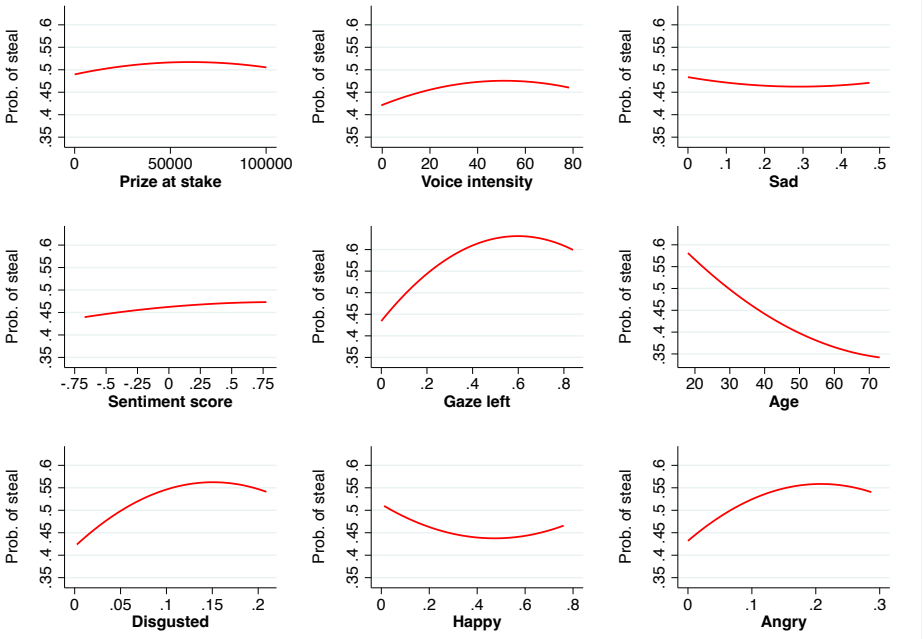
D.4.1 Influence and Partial Dependence Plots for GBMs

Table D.5: Relative Influence of Contestant Characteristics on Predicted Stealing Risk

Covariate	Type	Relative influence
Prize at stake (Jackpot amount)	Game	8.44
Voice intensity	Voice	7.92
Sad	Facial	7.66
Sentiment Score	Speech	7.14
Gaze left	Facial	7.01
Age	Contestant	6.01
Explicit, unconditional promise	Speech	5.65
Disgusted	Facial	5.63
Happy	Facial	4.67
Angry	Facial	4.17

Notes: This table shows the covariates with the highest relative influence on predictions generated by the GBM, excluding quality of the video, as rated by FaceReader (importance 6.94) and the number of non-missing frames (importance 4.98).

Figure D.6: Partial Dependence Plots

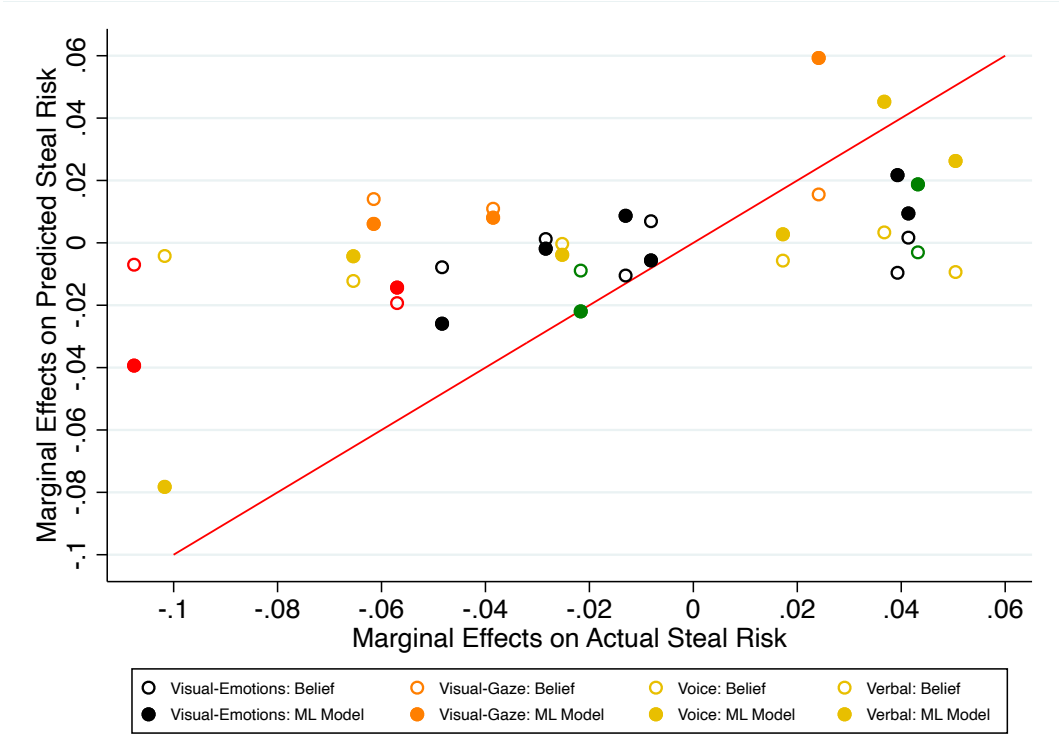


Notes: This figure shows the partial dependent plots for continuous covariates with the highest relative influences on the likelihood of stealing. Explicit, unconditional promise is not included as it only only takes values 0, 1 or 2. For this covariate, the predicted likelihood of stealing is 0.49 if the contestant does not make such a promise, while it is 0.45 if the contestant makes at least one statement containing such promise.

D.4.2 Correlation between Predictions and Features

To examine whether the ML model correlates with the correct features for stealing whereas participants' beliefs do not, we examine the relationship between the contestant's emotions, speech, and other covariates and beliefs as well as model predictions. In each case, we estimate simple linear models on these features of the contestant, which are standardized if they are continuous variables. A concern is that the ML model allows for relationships to be nonlinear, whereas this model is linear. However, we obtain similar conclusions when we focus on predictions by a linear model (e.g., the lasso), as shown in the next section. A summary of the regression analysis is presented in Figure D.7, which plots the marginal effects of each feature on the actual steal risk against those on the ML and participant predictions.

Figure D.7: Marginal Effects of Contestant Features on Predicted and Actual Steal Risk



Notes: This figure shows the marginal effect of each feature on the actual likelihood of stealing on the x-axis and the correlation between ML's prediction (using GBM) or participant's predictions (all treatments in Experiment 1) on the y-axis. The dots that are full show the relationship between actual and ML predictions for stealing, while the empty dots show the relationship between actual and participant predictions, for each feature. Detailed regression results are shown in Table D.6, and disaggregated regression results for different participant treatments are shown in Table D.7.

Table D.6: The Relationship Between Contestant Features, Beliefs, GBM Predictions & Stealing

	(1)	(2)	(3)	(4)	(5)	(6)
	Steal (=1)		ML prediction about stealing		Belief about stealing All treatments	
	Coeff.	SE	Coeff.	SE	Coeff.	SE
Female	-0.057	(0.058)	-0.014	(0.017)	-0.020*	(0.011)
Age est.	-0.108***	(0.023)	-0.039***	(0.007)	-0.006	(0.006)
Happy	-0.048*	(0.027)	-0.026***	(0.008)	-0.008	(0.006)
Sad	-0.028	(0.026)	-0.002	(0.009)	0.002	(0.004)
Angry	0.039	(0.026)	0.022***	(0.008)	-0.010**	(0.004)
Surprised	-0.013	(0.030)	0.009	(0.009)	-0.010*	(0.005)
Scared	-0.008	(0.027)	-0.006	(0.007)	0.007	(0.005)
Disgusted	0.041	(0.027)	0.009	(0.008)	0.002	(0.004)
Gaze forward	-0.039	(0.034)	0.008	(0.011)	0.010*	(0.006)
Gaze left	0.024	(0.043)	0.059***	(0.015)	0.016**	(0.007)
Gaze right	-0.062	(0.047)	0.006	(0.015)	0.014	(0.009)
Word count	0.050	(0.045)	0.026*	(0.014)	-0.010	(0.009)
Word sentiment score	0.037	(0.023)	0.045***	(0.007)	0.003	(0.005)
Explicit, unconditional promise	-0.102***	(0.023)	-0.078***	(0.007)	-0.004	(0.005)
Explicit, conditional promise	0.017	(0.022)	0.003	(0.005)	-0.006	(0.004)
Implicit, conditional promise	-0.025	(0.024)	-0.004	(0.007)	-0.000	(0.004)
Implicit, unconditional promise	-0.065***	(0.024)	-0.004	(0.007)	-0.013**	(0.005)
Jackpot mentioned	0.055	(0.087)	0.168***	(0.020)	0.032**	(0.015)
Jackpot mentioned X Jackpot amount	0.038*	(0.022)	0.019***	(0.007)	-0.017***	(0.004)
Voice pitch	0.043	(0.027)	0.019**	(0.007)	-0.002	(0.005)
Voice intensity	-0.022	(0.030)	-0.022**	(0.009)	-0.008	(0.006)
Quality of image	0.058	(0.038)	0.043***	(0.011)	-0.012*	(0.007)
Fraction of video analyzed by FaceReader	0.032	(0.048)	-0.010	(0.017)	0.001	(0.009)
Nr. of video frames	0.015	(0.038)	0.014	(0.012)	0.022***	(0.008)
Constant	0.440***	(0.087)	0.314***	(0.020)	0.406***	(0.018)
Observations	430		430		12,300	
Clusters	-		-		615	
R-squared	0.168		0.528		0.019	

Notes: This table presents the coefficients and standard errors from linear regression models on the relationship between actual stealing (columns 1-2), GBM predictions about stealing (columns 3-4), and participant beliefs in Experiment 1, all conditions pooled (columns 5-6), with contestant features in the conversation prior to the steal decision. Robust standard errors are computed for predictions (columns 1-2) and clustered at the contestant level for the regression in which human beliefs are the dependent variable (columns 5-6). ***, **, and * indicate 1%, 5%, and 10% significance levels, respectively.

Table D.7: Detailed Analysis of the Relationship Between Contestant Features, Beliefs, GBM Predictions & Stealing, by Treatment and Sample

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	GBM		Belief about stealing					
	ML prediction about stealing		No Learning Prolific		No Learning UCSD Laboratory		Learning	
	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
Female	-0.014	(0.017)	-0.014	(0.013)	-0.012	(0.016)	-0.018	(0.014)
Age est.	-0.039***	(0.007)	-0.002	(0.009)	-0.003	(0.007)	-0.020***	(0.006)
Happy	-0.026***	(0.008)	-0.007	(0.008)	-0.011	(0.009)	-0.007	(0.008)
Sad	-0.002	(0.009)	0.002	(0.006)	0.002	(0.006)	0.002	(0.006)
Angry	0.022***	(0.008)	-0.007	(0.006)	-0.012**	(0.006)	-0.009*	(0.005)
Surprised	0.009	(0.009)	-0.010	(0.007)	-0.014*	(0.008)	-0.007	(0.007)
Scared	-0.006	(0.007)	0.009	(0.007)	-0.007	(0.007)	0.014**	(0.006)
Disgusted	0.009	(0.008)	0.001	(0.005)	-0.003	(0.007)	0.010*	(0.005)
Gaze forward	0.008	(0.011)	0.013*	(0.008)	-0.000	(0.008)	0.011	(0.009)
Gaze left	0.059***	(0.015)	0.024**	(0.010)	-0.005	(0.011)	0.013	(0.010)
Gaze right	0.006	(0.015)	0.022**	(0.011)	-0.009	(0.012)	0.007	(0.011)
Word count	0.026*	(0.014)	-0.005	(0.013)	-0.015	(0.012)	-0.012	(0.011)
Word sentiment score	0.045***	(0.007)	-0.000	(0.006)	0.010	(0.007)	0.004	(0.006)
Explicit, unconditional promise	-0.078***	(0.007)	-0.008	(0.006)	0.003	(0.007)	-0.002	(0.006)
Explicit, conditional promise	0.003	(0.005)	-0.008	(0.006)	-0.002	(0.006)	-0.007	(0.005)
Implicit, conditional promise	-0.004	(0.007)	0.004	(0.005)	-0.004	(0.006)	-0.005	(0.007)
Implicit, unconditional promise	-0.004	(0.007)	-0.010	(0.007)	-0.019***	(0.006)	-0.007	(0.007)
Jackpot mentioned	0.168***	(0.020)	0.011	(0.017)	0.037*	(0.022)	0.017	(0.021)
Jackpot mentioned X Jackpot amount	0.019***	(0.007)	-0.016***	(0.005)	-0.019***	(0.006)	-0.012**	(0.006)
Voice pitch	0.019**	(0.007)	-0.008	(0.007)	-0.003	(0.007)	0.003	(0.006)
Voice intensity	-0.022**	(0.009)	-0.012	(0.009)	-0.006	(0.009)	-0.001	(0.007)
Quality of image	0.043***	(0.011)	-0.021**	(0.009)	0.001	(0.010)	-0.000	(0.009)
Fraction of video analyzed by FaceReader	-0.010	(0.017)	-0.001	(0.012)	0.016	(0.013)	-0.005	(0.012)
Nr. of video frames	0.014	(0.012)	0.022*	(0.012)	0.021**	(0.010)	0.020**	(0.009)
Constant	0.314***	(0.020)	0.399***	(0.021)	0.435***	(0.032)	0.430***	(0.030)
Observations	430		6,200		2,920		3,180	
Clusters	-		310		146		159	
R-squared	0.528		0.023		0.026		0.031	

Notes: This table presents the coefficients and standard errors from linear regression models on the relationship between GBM predictions about stealing (columns 1-2), and participant beliefs in the no learning condition (columns 3-4) and in the learning condition (columns 5-6), with contestant features in the conversation prior to the steal decision. Robust standard errors are computed for predictions (columns 1-2) and clustered at the participant level for the regressions in which human beliefs are the dependent variables (columns 3-6). ***, **, and * indicate 1%, 5%, and 10% significance levels, respectively.

D.4.3 Lasso

We conduct the same correlational analyses in this section, using LASSO predictions instead of GBMs. The results are qualitatively similar.

Table D.8: The Relationship Between Contestant Features, Beliefs, Lasso Predictions & Stealing

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
			Belief about stealing					
	ML prediction about stealing		No Learning Prolific		No Learning UCSD laboratory		Learning Prolific	
	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
Female	0.000	(0.000)	-0.019**	(0.009)	-0.015	(0.014)	-0.022	(0.012)
Age est.	-0.043***	(0.000)	-0.003	(0.004)	-0.004	(0.006)	-0.020***	(0.006)
Happy	-0.010***	(0.000)	-0.004	(0.004)	-0.010	(0.007)	-0.006	(0.006)
Sad	0.000	(0.000)	0.001	(0.004)	0.001	(0.006)	0.000	(0.006)
Angry	0.000	(0.000)	-0.006	(0.005)	-0.012**	(0.006)	-0.008	(0.006)
Surprised	-0.000	(0.000)	-0.010*	(0.005)	-0.014*	(0.008)	-0.007	(0.007)
Scared	-0.000	(0.000)	0.008**	(0.004)	-0.007	(0.007)	0.014**	(0.006)
Disgusted	-0.000	(0.000)	0.001	(0.005)	-0.003	(0.005)	0.010	(0.006)
Gaze forward	0.000	(0.000)	0.015**	(0.006)	-0.000	(0.007)	0.012	(0.008)
Gaze left	0.035***	(0.000)	0.024***	(0.007)	-0.005	(0.011)	0.014	(0.010)
Gaze right	0.000	(0.000)	0.023***	(0.007)	-0.009	(0.011)	0.009	(0.010)
Word count	-0.000	(0.000)	-0.003	(0.007)	-0.014	(0.009)	-0.012	(0.011)
Word sentiment score	0.000	(0.000)	0.000	(0.004)	0.010**	(0.005)	0.005	(0.005)
Explicit, unconditional promise	-0.060***	(0.000)	-0.009**	(0.004)	0.003	(0.005)	-0.002	(0.006)
Explicit, conditional promise	-0.000	(0.000)	-0.007**	(0.004)	-0.002	(0.005)	-0.007	(0.005)
Implicit, conditional promise	0.000*	(0.000)	0.005	(0.004)	-0.003	(0.005)	-0.004	(0.006)
Implicit, unconditional promise	-0.009***	(0.000)	-0.010***	(0.004)	-0.017***	(0.006)	-0.005	(0.006)
Jackpot mentioned	-0.001***	(0.000)	-0.004	(0.009)	-0.002	(0.011)	-0.022**	(0.006)
Jackpot mentioned X Jackpot amount	0.001*	(0.000)	-0.009***	(0.005)	-0.013**	(0.005)	-0.007	(0.005)
Voice pitch	0.000	(0.000)	-0.007*	(0.004)	-0.003	(0.005)	0.003	(0.006)
Voice intensity	0.000	(0.000)	-0.012**	(0.005)	-0.008	(0.006)	-0.000	(0.006)
Quality of image	-0.000	(0.000)	-0.021***	(0.006)	0.001	(0.008)	0.001	(0.009)
Fraction of video analyzed by FaceReader	-0.000	(0.000)	-0.002	(0.008)	0.015	(0.012)	-0.010	(0.010)
Nr. of video frames	-0.000	(0.000)	0.020***	(0.006)	0.019***	(0.010)	0.019**	(0.009)
Constant	0.461***	(0.000)	0.444***	(0.010)	0.482***	(0.015)	0.517***	(0.012)
Observations	430		6,200		2,920		3,180	
Clusters	-		310		146		159	
R-squared	0.998		0.018		0.020		0.015	

Notes: This table presents the coefficients and standard errors from linear regression models on the relationship between lasso predictions about stealing (columns 1-2), participant beliefs in the no learning condition (columns 3-4) and in the learning condition (columns 5-6), with contestant features in the conversation prior to the steal decision. Robust standard errors are computed for predictions and actual stealing (columns 1-2) and clustered at the participant level for the regressions in which human beliefs are the dependent variables (columns 3-8). ***, **, and * indicate 1%, 5%, and 10% significance levels, respectively.

D.5 Nonverbal Information Treatment: Results on Accuracy

In Experiment 1, 240 participants made predictions about contestants' decisions to steal or split, by watching only muted videos, in the Nonverbal information treatment. These participants took part in first wave of Experiment 1, and hence we compare their behavior to that of those individuals in the main (Verbal information treatment) in the same wave ($N = 258$). In the Nonverbal information treatment, the accuracy of participants, as measured by the AUC was 0.4945 (95% CI is 0.47793, 0.51099). This accuracy is significantly lower than that in the main (Verbal information) treatment, reported in the main text (χ^2-test, p-value$<$0.001), showing that participants paid attention and use features of contestant's conversations in their predictions. The fraction of correct guesses, using 50% as the threshold, was 49.75%. This fraction is also significantly lower than the fraction of correct guesses in the main (Verbal information) treatment (t-test, p-value$<$0.001).

Compared to the Verbal treatment, participants were less confident in their absolute ability to guess correctly, but their relative confidence was not significantly different. They believed to have correctly guesses contestant behavior in 52.6% of the cases in the Nonverbal information treatment, while they believed to have guessed correctly in 58.5% of the cases with Verbal information (t-test, p-value$<$0.001). The belief about the quartile of the distribution in which a participant's ability lied was 2.25 in the Nonverbal information treatment, while it was 2.17 in the Verbal information treatment. The difference is not significant (χ^2-test, p-value=0.553).

D.6 Additional Experiment: Flagging and Delegation Decisions

D.6.1 Experimental Design

In a separate experiment, we examined individuals' preferences to delegate their predictions to an algorithm. In this experiment, flags were introduced at the same time as the video, and their timing was not varied across treatments. The experiment consisted of three treatments: control ($N = 150$), ML-flags ($N = 256$), and ML-delegation ($N = 245$). The control condition did not provide participants with any information about the ML algorithm's prediction.

In the ML-flags treatment, as in Experiment 2 in the main paper, participants were told that the researchers had used the contestants' facial expressions and speech to train a standard ML algorithm to predict contestant decisions. 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 with very high chance or very low."

The ML-delegation treatment gave participants the option to delegate their prediction for the four videos that the ML algorithm would flag at the beginning of the experiment. Participants knew that if they chose to delegate, their prediction would be that of the ML algorithm. They still saw the video of the contestant and the same flag, but the option to make a prediction was removed. If they chose not delegate, the videos and flags regarding the ML predictions were the same as in the ML-flags treatment.

As in Experiment 2, to identify the effect of flags at the video level, we used a group of 20 videos, which included four videos for which the ML algorithm had predicted over a 70% likelihood of steal, and four videos for which it had predicted over a 70% likelihood of split. All flags provided participants correct feedback. Each participant could be presented with one of two groups of "flags." In the first group, each participant saw a message indicating a high predicted likelihood of stealing for two videos and a message indicating a high predicted likelihood of splitting for two videos. In the second group, the other two (out of four) videos with a high predicted likelihood of splitting (or stealing) were flagged for the participant. This way, all participants in the ML-flags treatment saw four flagged videos in total. But, across the groups, we varied which videos were flagged.

After completing the 20 predictions, we elicited three beliefs from participants. First, 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 to which they thought they belonged. Third, in the treatments with flags, we elicited participants' belief regarding the algorithm's accuracy. In all cases, they 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.

Participants received received \$3.00, as a fixed participation fee, in addition to incentives depending on their predictions for one randomly-selected video and their answers about

their ability and that of the algorithm. The proportion of female participants was 47.9%, the average age of participants was 33.29, and 90.0% reported never having seen the TV show before the study.

D.6.2 Results

We start by describing the decision to delegate in the ML-delegation treatment, and then examine how flags affect beliefs and accuracy.

When given the opportunity to delegate, we find 53.5% of participants choose to delegate their predictions for videos flagged by the algorithm. In probit models, we examine whether individual characteristics, accuracy, and beliefs regarding the algorithm's accuracy as well as own performance are related to the delegation decision. Table D.9 shows that women are not significantly more likely to delegate than men, but older people exhibit a higher propensity to delegate. A participant who is 10 years older than the mean participant, who is 34 years old, exhibits a 0.7-percentage-point higher likelihood of delegating her predictions (p-value= 0.013).

Table D.9: Determinants of Delegation to the Algorithm

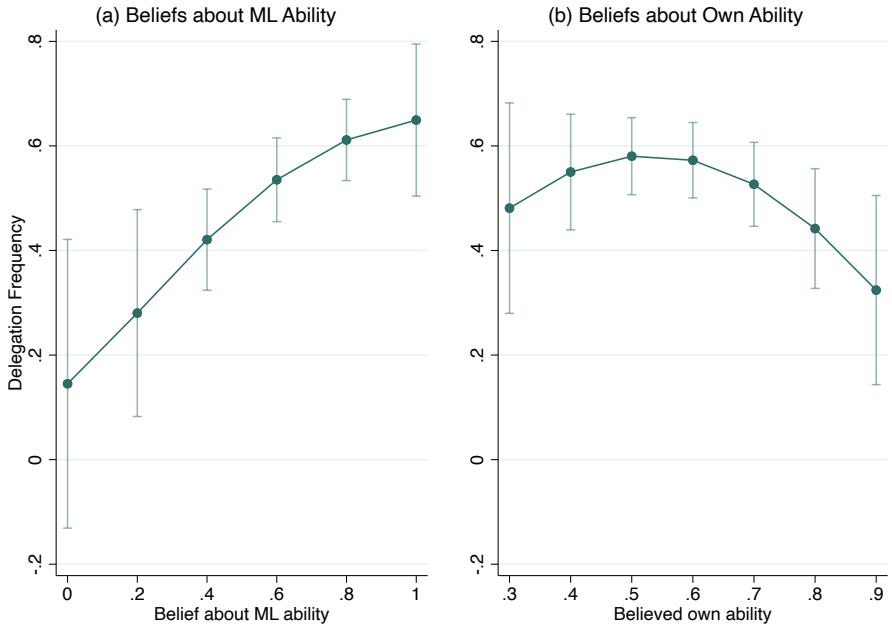
	(1)	(2)	(3)
	Delegate = 1		
Female	0.017	0.016	0.025
	(0.064)	(0.064)	(0.063)
Age	0.007***	0.007***	0.007***
	(0.003)	(0.003)	(0.003)
Familiarity with TV show	-0.130	-0.133	-0.137
	(0.096)	(0.096)	(0.094)
Believed own ability		-0.154	-0.275
		(0.212)	(0.212)
Believed ability ML			0.463***
			(0.139)
Observations	245	245	245

Notes: This table presents the marginal effects, calculated at the means of the regressors, in probit regression models of participants' decision to delegate. Believed own ability is the fraction of correct predictions that the participant believed she made (using a 50% threshold). Believed ability ML is the fraction of correct predictions (out of 4) that the participant believed the ML made. Standard errors are shown in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% levels.

Participants who believe the ML is more accurate are significantly more likely to delegate. On average, participants who delegate believe that the algorithm is accurate 68.1% of the time, whereas those who choose not to delegate believe the flags by the algorithm are correct

59.6% of the time (p-value= 0.002). Participants' beliefs about their own ability are not significantly related to their delegation decision. Figure D.8 shows the relationship between delegation and beliefs about own and ML ability, based on a probit regression on the likelihood of delegating and allowing the relationship with beliefs to be quadratic. Consistent with Table D.9, higher beliefs about ML ability increase the likelihood of delegation, whereas beliefs about own ability do not exhibit a significant relationship with delegation.

Figure D.8: Beliefs about Own and ML Ability and Delegation



Notes: This figure shows the relationship between participants' beliefs about their own ability and delegation frequency, as well as the relationship between participant beliefs about ML ability to predict stealing and delegation frequency. The figure plots predictive margins based on a probit regression in which the delegation decision is the dependent variable. Beliefs about ML ability and about own ability are allowed to be nonlinear (quadratic). The regression includes the participants' gender, age, and familiarity with the TV show as covariates. Whisked bars indicate 95% confidence intervals.

Table D.10 extends the main results, exploring nonlinearity in the relationship between beliefs and delegation, and examining the relationship with actual accuracy. The table shows the complete regression coefficients underlying Figure D.8 (in column (2)).

We next examine the effects of flags on participants' predictions, in the ML-Flags and ML-Delegation treatments, compared to Control. On average, participants react to the flags of the algorithm. Figure D.9 shows the average prediction in each treatment, conditional on whether the algorithm flagged the contestant in the video as having a very low or very high chance of stealing, or if there was no flag.

Without flags, participants believe the chance of stealing is 39.1% for videos that the ML

Table D.10: Determinants of the Delegation Decision: Robustness

	(1)	(2)	(3)	(4)
		Delegate = 1		
Female	0.066	0.055	0.027	-0.007
	(0.168)	(0.169)	(0.170)	(0.173)
Age	0.018**	0.019**	0.019**	0.020***
	(0.007)	(0.007)	(0.007)	(0.008)
Familiar with show	-0.373	-0.348	-0.350	-0.309
	(0.267)	(0.272)	(0.269)	(0.275)
Believed ability ML	1.237***	2.728	1.353***	2.915
	(0.396)	(2.057)	(0.400)	(2.055)
Believed ability ML^2 (squared)		-1.223		-1.310
		(1.566)		(1.571)
Believed own ability	-0.736	5.456	-0.917	5.801*
	(0.573)	(3.337)	(0.578)	(3.319)
Believed own ability^2 (squared)		-5.151*		-5.558**
		(2.724)		(2.713)
Actual own ability			-1.810***	5.462
			(0.700)	(5.606)
Actual own ability^2 (squared)				-6.309
				(4.725)
Constant	-0.599	-2.790**	0.471	-3.892**
	(0.504)	(1.159)	(0.654)	(1.969)
Observations	245	245	245	245

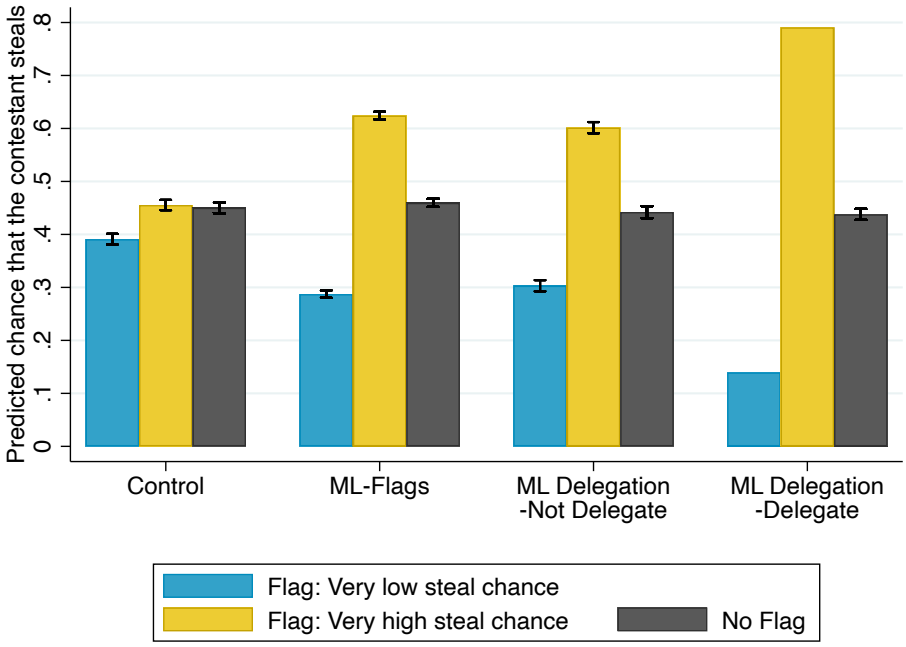
Notes: This table presents the coefficients and standard errors from probit regression models on the relationship between delegation decisions and participant characteristics and beliefs. Standard errors are presented in parentheses. ***, **, and * indicate 1%, 5%, and 10% significance levels, respectively.

algorithm flagged as "very low chance of stealing," 45.5% for those flagged as "very high," and 45.03% for videos that were never flagged. The difference between videos flagged as "very low" or "very high chance of stealing" is significant (p-value= 0.004) at 6 percentage points. Flags significantly change predictions. Participants in the ML-flags treatment reduce their predicted chance of stealing to 28.7% when the video is flagged as "low chance of stealing" and increase it to 62.4% when the video is flagged as "high chance of stealing." A similar effect is observed in ML-delegation for those participants who do not delegate. The predictions of participants who delegate are the predictions of the algorithm (14% chance of stealing for those videos flagged as low chance, and 79% for those videos flagged as high chance).

Table D.11 examines the effects of the treatments and the flags on beliefs and the accuracy of predictions, including contestant fixed effects (as well as an indicator for whether participants were assigned to the first or second group of flags). Columns (1)–(3) reveal no treatment effects for videos that were not flagged. All changes in beliefs occur in response to flagging. A "very low" flag decreases the predicted chance of stealing by 13.2 percentage points in the ML-flags treatment by 9.9 percentage points among those who do not delegate in the ML-delegation treatment, and by 25.8 percentage points when participants delegate and hence their predictions are those of the algorithm.

Because flags affect beliefs and they are correct, the accuracy of beliefs for flagged videos

Figure D.9: The Effects of Flags on Beliefs



Notes: This figure shows the average predicted chance that the contestant steals, by treatment in the Delegation Experiment, separating those who do not delegate and those who delegate in the ML-delegation treatment. Whisked bars denote 95% confidence intervals. Confidence intervals are not included for those who delegate in the ML-delegation treatment, because these represent the average prediction of the algorithm for the four flagged videos.

increases. Table D.12 shows that in the ML-flags treatment, the fraction of correct guesses is 81.4% for videos flagged as low chance of stealing and 68.6% for videos flagged as high chance, compared with 71.1% and 39.9%, respectively, in the control treatment. A similar finding is observed for participants in the ML-delegation treatment who choose not to delegate to the algorithm. The effect of flags is restricted to videos that are flagged, because flags do not increase accuracy for those videos that were not flagged. Across all treatments, the accuracy for videos that were not flagged is between 56.8% and 57.8%.

Column (4) of Table D.11 reveals that flagging videos increases accuracy overall, by 4 to 9 percentage points, although only 4 out of 20 videos were flagged. Columns (5) and (6) confirm that the effects stem from increases in accuracy for flagged videos.

Table D.11: Effects of Flagging on Beliefs and Accuracy

	(1)	(2)	(3)	(4)	(5)	(6)
	Predicted chance contestant steals			Correct prediction		
Treatment Effects						
ML Flag	0.014	0.012	0.014	0.040***	-0.002	-0.001
	(0.015)	(0.015)	(0.015)	(0.011)	(0.011)	(0.011)
ML Delegation-Not Delegate	-0.001	-0.004	-0.005	0.039***	0.008	0.007
	(0.018)	(0.018)	(0.018)	(0.015)	(0.015)	(0.015)
ML Delegation-Delegate	-0.002	-0.009	-0.005	0.092***	0.002	0.001
	(0.016)	(0.018)	(0.017)	(0.011)	(0.013)	(0.013)
Treatment X Flag Effects						
ML Flag X Very low flag		-0.132***	-0.132***		0.160***	0.160***
		(0.012)	(0.012)		(0.021)	(0.021)
ML Flag X Very high flag		0.147***	0.147***		0.253***	0.253***
		(0.013)	(0.013)		(0.025)	(0.025)
ML Delegation-Not Delegate X Very low flag		-0.099***	-0.099***		0.116***	0.116***
		(0.018)	(0.018)		(0.031)	(0.031)
ML Delegation-Not Delegate X Very high flag		0.137***	0.137***		0.189***	0.189***
		(0.019)	(0.019)		(0.037)	(0.037)
ML Delegation-Delegate X Very low flag		-0.258***	-0.258***		0.342***	0.342***
		(0.013)	(0.013)		(0.016)	(0.016)
ML Delegation-Delegate X Very high flag		0.330***	0.330***		0.560***	0.560***
		(0.014)	(0.014)		(0.017)	(0.017)
Participant Age			-0.001***			0.000
			(0.001)			(0.000)
Female participant			0.004			0.018**
			(0.011)			(0.009)
Familiar with show			0.016			-0.031**
			(0.018)			(0.015)
			-0.007			-0.001
Constant	0.549***	0.552***	0.598***	0.536***	0.576***	0.560***
	(0.015)	(0.015)	(0.024)	(0.021)	(0.021)	(0.024)
Observations	13,000	13,000	13,000	13,000	13,000	13,000
R-squared	0.111	0.157	0.161	0.073	0.107	0.108

Notes: This table presents the 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)). All specifications include video (contestant) fixed effects and an indicator for which group of flags the participant was assigned to. Robust standard errors clustered at the participant level are presented throughout. *, **, *** indicate significance at the 10%, 5%, and 1% levels.

Table D.12: Accuracy and Beliefs about Accuracy

	(1) Control	(2) ML-Flags	(3) ML-Delegation	(4)
			Delegate	Not delegate
Fraction correct guesses				
Videos flagged as low chance	71.10%	81.40%	78.10%	100.00%
Videos flagged as high chance	39.90%	68.60%	63.60%	100.00%
Not-flagged videos	56.80%	56.90%	57.80%	57.10%
AUC				
Flagged videos	0.57	0.82	0.79	1.00
Not-flagged videos	0.60	0.59	0.62	0.61
Beliefs				
Absolute ability	57.00%	57.20%	59.50%	59.00%
Relative ability (average quartile, 1-4)	2.3	2.26	2.24	2.22
ML accuracy	-	60.50%	59.60%	68.10%

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

E Pre-registrations

As Predicted:"*GoldenBalls – Experiment 1*" (#39504)

Created: 04/19/2020 02:46 PM (PT)

Author(s)
Marta Serra-Garcia (University of California, San Diego) - mserragarcia@ucsd.edu
Uri Gneezy (University of California San Diego) - ugneezy@ucsd.edu

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?
There are two main research questions: (a) how good are people at detecting lies in high-stakes environments? And (b) how confident are people about their ability to detect lies?
Participants will be shown videos of contestants in the TV show Golden Balls. These videos will show the conversations of contestants prior to making a split (cooperate) or steal (defect) decision. They will be incentivized to correctly guess whether the contestant in the video chose to split or to steal, using a binarized scoring rule.
Our main hypotheses are: (1) People will not be better than chance at guessing whether the person splits or steals. This will include believing a false statement (type I error) and not believing a true one (type II error). (2) We expect people to be overconfident about their ability to predict behavior.

3) Describe the key dependent variable(s) specifying how they will be measured.
The key dependent variables are:
- Individual belief that the contestant will steal or split
- Individual number of correct guesses of each participant (using 50 as threshold), and sensitivity / specificity of their accuracy
- Individual belief about performance (belief about correctness of guesses)
- Individual belief about performance relative to others (which quartile of the distribution of scores they believe they are in)
- Likelihood that a split decision is detected and likelihood that a steal decision is detected on aggregate

4) How many and which conditions will participants be assigned to?
There will be 2 conditions. In the muted condition subjects will view muted videos of one contestant at time, drawn from the conversation with their counterpart. In the conversation condition subjects will view videos with sound, featuring the conversation between two contestants.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
We will test:
(a) whether participants are better than chance, and whether the distribution of scores differs from a distribution based on chance
(b) whether participants are better or worse than a machine-learning algorithm trained to predict split and steal decisions based on contestants' characteristics and their facial expressions and movements.
(c) whether participants are better at predicting split or steal decisions
(d) whether they are overconfident about their absolute ability to predict behavior, and their relative ability,
(e) whether there is "wisdom of the crowd", that is whether the aggregate belief of the crowd is better than chance
(f) whether individuals' confidence is correlated with their actual ability
(g) whether men are more overconfident in their ability to detect lies than women
We will test whether the two conditions (muted/conversation) affect the ability of predict split decisions, and if they do not, we will pool these in the analyses.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
will exclude individuals who answer the control questions wrong

7) How many observations will be collected or what will determine sample size?
No need to justify decision, but be precise about exactly how the number will be determined.
We will focus on showing participants a random subsample of approximately 125-130 episodes from Seasons 1 to 4 of Golden Balls. We aim to recruit approximately 500 participants, evenly split across the two conditions.

8) Anything else you would like to pre-register?
(e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
We will collect information on gender, age, educational background, and familiarity with the TV show. We will also measure the time individuals take to make their guesses. We plan to run exploratory analyses based on these participant characteristics/behaviors.

The accuracy of human predictions will be compared to the predictive accuracy of regression and decision-tree models (machine-learning).

'*GoldenBalls - Flags_Timing*'
(AsPredicted #107116)

Created: 09/15/2022 11:49 AM (PT)

Author(s)
Marta Serra-Garcia (University of California San Diego) - mserragarcia@ucsd.edu
Uri Gneezy (University of California, San Diego) - ugneezy@ucsd.edu

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?
Our main research questions are: (1) do people change their predictions when provided with a machine-learning algorithm's "flags" of videos indicating a behavior (cooperation) is very likely or unlikely, (2) does the timing of the flag matter? Are people more likely to follow flags if they are presented before they form their own beliefs about behavior?

Participants will be shown 20 videos of contestants in the TV show Golden Balls. They will be incentivized to correctly guess whether the contestant in the video chose to split or to steal, using a binarized scoring rule.

Our main hypotheses are: 1. Based on previous findings, we expect people's guesses to be significantly affected by the machine-learning algorithm's prediction in the form of a flag (ML-flag). 2. We expect people's guesses to adjust more to the ML-flags more when they are shown before they form their beliefs about behavior, compared to after.

3) Describe the key dependent variable(s) specifying how they will be measured.
The key dependent variables are:
- Individual belief that the contestant will steal or split
- Individual number of correct guesses of each participant (using 50% as threshold), and sensitivity / specificity of their accuracy
- Individual belief about performance (belief about correctness of guesses)
- Individual belief about performance relative to others (which quartile of the distribution of scores they believe they are in)
- Individual beliefs about performance of the machine-learning algorithm.

4) How many and which conditions will participants be assigned to?
There will be 3 conditions: Before, where flags will be shown for 4 videos before the video is shown; After, where flags will be shown for 4 videos after the video is shown; and Control, without any flags. In Control, half of the subjects will see the videos and submit their guess in the same manner as Before, without flags, and half will see the videos and submit their guess in the same order as After, again without flags.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
We will test whether participants are better than chance; whether participants' predictions are affected by machine-learning predictions based on "flagging" some videos; whether flags affect predictions more strongly if they are shown before, rather than after, participants watch the video; whether participants believe machine-learning algorithms are better than chance and them at predicting steal/split decisions.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
We will exclude individuals who answer the control questions wrong.

7) How many observations will be collected or what will determine sample size?
No need to justify decision, but be precise about exactly how the number will be determined.
We will recruit 200 participants in each condition.

8) Anything else you would like to pre-register?
(e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
We will collect information on gender, age, and familiarity with the TV show. We will also measure the time individuals take to make their guesses. We plan to run exploratory analyses based on these participant characteristics/behaviors. The accuracy of human predictions will be compared to the predictive accuracy of regression and decision-tree models (machine-learning).

'*GoldenBalls - Experiment 1 Wave 2 and Experiment FlagsDelegation*'
(AsPredicted #73632)

Created: 08/30/2021 01:29 PM (PT)

Author(s)
Marta Serra-Garcia (University of California San Diego) - mserragarcia@ucsd.edu
Uri Gneezy (University of California San Diego) - ugneezy@ucsd.edu

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?
There are two main research questions: (a) how good are people at predicting behavior in high-stakes environments, and (b) do people change their beliefs when provided with a machine-learning algorithm's "flags" of videos with a very high/low likelihood of cooperation.
Participants will be shown videos of contestants in the TV show Golden Balls. We were able to collect 214 episodes of the show that we plan to use in the study. These videos will show the conversations of contestants prior to making a split (cooperate) or steal (defect) decision. They will be incentivized to correctly guess whether the contestant in the video chose to split or to steal, using a binarized scoring rule. Our main hypotheses are:
1. Based on previous findings, people will be modestly better than chance at guessing whether the person splits or steals. This will include believing a false statement (type I error) and not believing a true one (type II error).
2. Learning: Providing people with feedback after their predictions will only modestly increase their accuracy over time.
3. We expect people's guesses to be significantly affected by the machine-learning algorithm's prediction. We also predict that a significant proportion will be willing to delegate some predictions to the algorithm.

3) Describe the key dependent variable(s) specifying how they will be measured.
The key dependent variables are: Individual belief that the contestant will steal or split; Individual number of correct guesses of each participant (using 50% as threshold), and sensitivity / specificity of their accuracy; Individual belief about performance (belief about correctness of guesses); Individual belief about performance relative to others (which quartile of the distribution of scores they believe they are in); Individual belief about performance of the machine-learning algorithm; Choice to delegate to the machine-learning algorithm.

4) How many and which conditions will participants be assigned to?
There will be 2 experiments. First, in experiment A we will test whether providing feedback to individuals about the accuracy of their prediction. In this experiment, we will randomly sample from all available videos, and assign participants to the "Feedback" or "No Feedback" condition. Videos will be either "muted" (nonverbal only) or "voiced" (verbal information).
Second, in experiment B, we will test whether machine-learning predictions affect human predictions. In this experiment we will use a balanced sample of 20 videos which are not muted, with three conditions: (1) "NoML" which is a control without machine-learning predictions; (2) "MLflag" which will "flag" videos in which the algorithm predicts a very high or a very low chance of stealing, and (3) "MLdelegate" in which participants will choose whether to delegate their predictions for videos "flagged" by machine-learning as having very high or very low chance of stealing.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
We will test: (a) whether participants are better than chance; (b) whether participants are better or worse than a machine-learning algorithm trained to predict split and steal decisions based on contestants' characteristics and their facial expressions, speech and voice features; (c) whether they are overconfident about their absolute ability to predict behavior, and their relative ability; (d) whether participants' predictions are affected by machine-learning predictions based on "flagging" some videos; (e) whether participants are willing to delegate their predictions to the algorithm for "flagged" videos; (f) whether participants believe machine-learning algorithms are better than chance and than them at predicting steal/split decisions.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
We will exclude individuals who answer the control questions wrong.

7) How many observations will be collected or what will determine sample size?
No need to justify decision, but be precise about exactly how the number will be determined.
For experiment A, we aim to recruit 400 participants, assigning 100 to the "NoFeedback" condition and "300" to the "Feedback" condition. If behavior is similar, we plan to pool behavior in the "NoFeedback" condition with the results of a previous study using that treatment only (Aspredicted #39504). Otherwise, we will control for differences using regression analyses. For experiment B, we aim to recruite approximately: 150 participants for "NoML", 250 participants for "MLflag" and 250 participants for "MLdelegate".

8) Anything else you would like to pre-register?
(e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
We will collect information on gender, age, and familiarity with the TV show. We will also measure the time individuals take to make their guesses. We plan to run exploratory analyses based on these participant characteristics/behaviors. The accuracy of human predictions will be compared to the predictive accuracy of regression and decision-tree models (machine-learning).

F REFORMS Checklist

REFORMS checklist

Visit reforms.cs.princeton.edu for the latest version.

About. The REFORMS checklist lists items that should be reported in a scientific study that uses machine learning (ML) methods. It is intended to accompany the paper or report that introduces an ML model: for instance, as an appendix or supplemental material. The checklist consists of 32 questions spread across 8 modules. For each item, either list the section name, section number, or page number in the paper where the item is reported, or justify why a given item is not filled out. Note that not all of these items need to be reported in the main text of the paper; they could be reported in an appendix or supplementary files.

Some items in the checklist could be hard to report for specific studies. For instance, including a reproduction script to computationally reproduce all results (2e.) might not be possible for studies performed on academic computing clusters or those which use private data that cannot be released. Instead of requiring strict adherence for each item, we suggest authors and referees decide which items are relevant for a study and where details can be reported better. The items in our reporting standards could be a helpful starting point.

The text in italics below provides the answer to each of the questions in the checklist.

Checklist for reporting ML-based science

Module 1: Study goals

1a. Population or distribution about which the scientific claim is made.

Individuals who are tasked with detecting deception by contestants in the TV show "Golden Balls", in online experiments conducted on Prolific Academic and UC San Diego. Detailed information about the TV show is provided in Section 2 of the paper. The details about the individuals who participate in the experiments are provided in Section 3 of the paper.

1b. Motivation for choosing this population or distribution (1a.).

This population is chosen because it has been shown to be a highly attentive sample of participants that exhibits behaviors that are also observed in representative samples of the US population (Peer et al., 2022).

1c. Motivation for the use of ML methods in the study.

Section 3, subsection 3.2.:
"Since 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."

Module 2: Computational reproducibility

All items in this module will be provided in the Replication Package of the paper.

2a. Dataset used for training and evaluating the model along with link or DOI to uniquely identify the dataset.

2b. Code used to train and evaluate the model and produce the results reported in the paper along with link or DOI to uniquely identify the version of the code used.

2c. Description of the computing infrastructure used.

- Hardware infrastructure: CPU, GPU, RAM, disk space etc.
- Operating system.
- Software environment: Programming language and version, documentation of all packages used along with versions and dependencies (e.g., through a requirements.txt file).
- An estimate of the time taken to generate the results.

2d. README file which contains instructions for generating the results using the provided dataset and code.

2e. Reproduction script to produce all results reported in the paper[1].

Module 3: Data quality

3a. Source(s) of data, separately for the training and evaluation datasets (if applicable), along with the time when the dataset(s) are collected, the source and process of ground-truth annotations, and other data documentation.

- *TV show episodes: obtained from Donja Darai for research purposes. Online Appendix A provides more details, and Footnote 2 specifies the source of the data for the TV show.*
- *Facial and Vocal Analyses: Obtained from FaceReader and Praat. Section 2.1.1 and Online Appendix B describe the source of data for the nonverbal features in the videos.*
- *Text Analyses: Transcripts obtained from Turmunkh et al. (2019), analyzed using Sentimentr package in R. Section 2.1.2 and Online Appendix B describe the sources of data for the verbal features in the videos.*
- *README file and Online Appendix B provide further details on the editing, processing, and feature extraction for all videos.*
- *Behavioral Data: Obtained via Qualtrics surveys, from participants in Prolific Academic and UCSD. Online Appendix A provides details about the instructions shown to participants.*

3b. Distribution or set from which the dataset is sampled (i.e., the sampling frame).

- *Section 2.2. provides additional information on the sample, listing exclusions.*
- *Online Appendix A.1. provides further details for the sampling of videos.*

3c. Justification for why the dataset is useful for the modeling task at hand.

It was the largest number of videos that we could obtain for this well-known TV show. The reasons for studying this TV show are provided in Section 2.

3d. The definition of the outcome variable of the model along with descriptive statistics, if applicable.

The outcome variable is the decision to split or steal. 54% of contestants split (see Section 2).

[1] Note that this is a high bar for computational reproducibility. It might not be possible to provide such a script—for instance, if the analysis is run on an academic computing cluster, or if the dataset does not allow for programmatic download.

3e. Number of samples in the dataset.

430 videos. 302 in the training set, and 128 in the test set. Details on the samples are provided in Online Appendix D.

3f. Percentage of missing data, split by class for a categorical outcome variable.

No missing data on the outcome variable.

3g. Justification for why the distribution or set from which the dataset is drawn (3b.) is representative of the one about which the scientific claim is being made (1a.).

We obtain as many episodes of the TV show Golden Balls as possible and show in Footnote 2 that the features of contestants on the show are very similar to those exhibited in a sample including 284 episodes in Turkmunkh et al. (2019), which has 69 more videos than our sample of 215.

Module 4: Data preprocessing

4a. Identification of whether any samples are excluded with a rationale for why they are excluded.

Online Appendix A.1. describes the exclusions and the rationale for them.

4b. How impossible or corrupt samples are dealt with.

Online Appendix B.1. describes the samples that could not be extracted from the videos using FaceReader.

4c. All transformations of the dataset from its raw form (3a.) to the form used in the model, for instance, treatment of missing data and normalization.

No transformations.

Module 5: Modeling

5a. Detailed descriptions of all models trained, including:

- All features used in the model (including any feature selection).
- Types of models implemented (e.g., Random Forests, Neural Networks).
- Loss function used.

"We randomly split the sample into a training dataset (302 videos) and a testing dataset (128 videos)." (Section 3).

Online Appendix C.1 provides the model, loss function and decision-making rule for GBMs. Online Appendix C.3. provides the same details for Rigorous Logistic Lasso. Regarding the inputs and outputs, detailed information on inputs is provided in Online Appendix C.2. which lists all inputs. Only one output is used throughout, the choice to steal (or split).

5b. Justification for the choice of model types implemented.

"We focus on a supervised learning approach: generalized boosted regression trees (GBM, see Friedman, 2002). Existing prediction models often present a tradeoff between interpretability and flexibility (e.g., Hastie et al., 2008). We focus on GBMs because they are flexible, allow for nonlinearity, and they have been previously found to have high predictive accuracy. 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." (Section 3).

5c. Method for evaluating the model(s) reported in the paper, including details of train-test splits or cross-validation folds.

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.

5d. Method for selecting the model(s) reported in the paper.

The hyperparameter tuning is selected using 5-fold cross-validation, as described below.

5e. For the model(s) reported in the paper, specify details about the hyperparameter tuning:

- Range of hyper-parameters used and a justification for why this range is reasonable.
- Method to select the best hyper-parameter configuration.
- Specification of all hyper-parameters used to generate results reported in the paper.

In Online Appendix C.2: "The interaction level of the regression trees is limited by the tree size J. In tuning the boosted tree parameters, we consider two levels J = 1, an additive model, and J = 2 a model in which two-variable interaction effects are also allowed. We also set the minimum number of observations in each terminal node, such that the resulting regions are based on enough observations. We explore a minimum of 5, 10, 20 or 25 observations in each terminal node. We explore different values of the shrinkage parameter (or learning rate): 0.1, 0.2, 0.3 and 0.4. Since the number of observations in the training data is limited and we want to avoid overfitting, we also introduce subsampling. This implies that

in each iteration only a fraction of the training observations is used to grow the next tree. We explore values of 0.5, 0.6 and 0.7 for the subsampling parameter. (…)

We train 100 trees (M = 100), use 5-fold cross-validation on the training sample, to determine the optimal interaction level of the model (J = 2), the minimum number of observations in each node (10), the learning rate (0.2), and the subsampling parameter (0.7)."

5f. Justification that model comparisons are against appropriate baselines.

We compare GBM to rigorous logistic lasso and logit regression. We explain the methods for rigorous logistic lasso in Online Appendix C.3., and describe the logit regression in Online Appendix C.5.

Module 6: Data leakage

6a. Justification that pre-processing (Section 4) and modeling (Section 5) steps only use information from the training dataset (and not the test dataset).

All videos for which the facial analysis could be conducted in at least one frame are included (Online Appendix B). The train-test split is performed at the contestant level and no contestant appears both in the training and test set.

6b. Methods to address dependencies or duplicates between the training and test datasets (e.g. different samples from the same patients are kept in the same dataset partition).

Each contestant only appears in the train or the test set.

6c. Justification that each feature or input used in the model is legitimate for the task at hand and does not lead to leakage.

The justification for the features considered for each video is provided in Section 2: "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 and/or guilt and overall 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 and 2015b). Serra-Garcia and Gneezy (2021) use simple probit models to relate facial expressions to truth-telling by experimental participants. In the study, participants were recorded in 30-second videos making either true or false statements. Several nonverbal features that are associated with the sender's truthfulness.

Hu and Ma (2020) use nonverbal and verbal features to estimate the positiveness in videos of startup pitches and relate these emotions to funding decisions."

Online Appendix C.2. provides further descriptions and justification for all the variables included.

Module 7: Metrics and uncertainty

7a. All metrics used to assess and compare model performance (e.g., accuracy, AUROC etc.). Justify that the metric used to select the final model is suitable for the task.

"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 0 otherwise. Second, we estimate the AUC." (Section 2.3)

7b. Uncertainty estimates (e.g., confidence intervals, standard deviations), and details of how these are calculated.

Confidence intervals are calculated for the AUC. They are calculated based on the test proposed by DeLong et al. (1988).

7c. Justification for the choice of statistical tests (if used) and a check for the assumptions of the statistical test.

We use standard statistical tests to compare the performance of the ML algorithm to that of humans (t-tests).

Module 8: Generalizability and limitations

8a. Evidence of external validity.

The paper focuses on a particular context, the TV show Golden Balls. The episodes are representative of the show, but external validity to other TV shows is not guaranteed.

8b. Contexts in which the authors do not expect the study's findings to hold.

We do not know whether the same ML-model would perform similarly in other TV shows and in other instances in which individuals are recorded in conversations where there are incentives to lie.

References

[1] Ahrens, A, Hansen, C.B., and Schaffer, M. (2020). LASSOPACK: Stata module for lasso, square-root lasso, elastic net, ridge, adaptive lasso estimation and cross-validation.

[2] Anikin, A., (2020). soundgen: Acoustic analysis with soundgen. R package version 1.8.1.

[3] Belloni, A., Chernozhukov, V., Hansen, C. and Kozbur, D. (2016). Inference in High-Dimensional Panel Models With an Application to Gun Control, *Journal of Business & Economic Statistics*, 34 (4), 590–605.

[4] Friedman, J. H., (2002). Stochastic Gradient Boosting. *Computational Statistics and Data Analysis* 38 (4), 367–378.

[5] Friedman, J. H., (2001). Greedy Function Approximation: a Gradient Boosting Machine. *The Annals of Statistics* 29 (5), 1189–1232.

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

[7] Kuhn, M., and Johnson, K. (2013). *Applied Predictive Modeling.* Springer, New York.

[8] Rinker, T. W. (2017). sentimentr: Calculate Text Polarity Sentiment. University at Buffalo/SUNY. http://github.com/trinker/sentimentr.

[9] Sueur, J., (2020). Seewave: A very short introduction to sound analysis for those who like elephant trumpet calls or other wildlife sound. Available at: https://cran.r-project.org/web/packages/seewave/vignettes/seewave_analysis.pdf

[10] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological),* 58(1), 267–288.

[11] Zhang, Z. (2019). Practical Data Processing for Social and Behavioral Research Using R. University of Notre Dame.