**Supplementary Information**

*for*

**Catch me if you can: Using machine learning and behavioral interventions to reduce unethical behavior**

Oliver P. Hauser, Michael Greene, and Katherine DeCelles

**1. Previous Work on Behavioral Interventions**

Many types of messages have been tested across the marketing, nudge, and tax compliance literatures, but not always found to be successful. We review the literature below, summarizing lab and field evidence for these messages which we use in our field experiment.

**Social Norms.** Social norms are a widely used intervention (Benartzi et al. 2017). Based on several decades of controlled lab studies (e.g., Schultz et al. 2007), social norms have recently been used in applied settings, affecting behavior in domains ranging from energy savings (Allcott 2011, Jachimowicz et al. 2018) to water conservation (Ferraro and Price 2013) to voting (Gerber and Rogers 2009). In particular, social norms have been shown to be successful to increase compliance and disclosure behavior (e.g. Chirico et al. 2016, Del Carpio 2013, Perez-Truglia and Troiano 2018). For example, Hallsworth et al. (2017) sent letters to citizens who were late on their tax payments, finding that making salient the tax payment behavior of peer groups leads to more timely disclosure of tax returns, bringing forward tax payments and reducing administrative costs.

But social norm messages are not consistently effective (e.g., Blumenthal et al. 2001, Cranor et al. 2020, De Neve et al. 2021, John and Blume 2018). For example, in an effort to reduce evasion of public broadcasting taxes, Fellner et al. (2013) show that a social norm nudge had a positive effect on compliance when compliance is believed to be common but a negative effect when rare and, in some cases, social norms only led to more compliance when combined with a penalty message. Larkin et al. (2019) sent nudge variations of letters to increase tax compliance of council tax in the UK, finding that social norms do better than a sanctioning threat message in driving compliance.[[1]](#footnote-1)

**Impact Others.**Several moral appeals to behave in accordance with the law and for the benefit of the greater good have been tested in the literature on compliance (e.g., Blumenthal et al. 2001, Bott et al. 2019, Chirico et al. 2016). While Blumenthal et al. (2001) and Kettle et al. (2017) find limited effects of their moral appeals and public goods letters, Bott et al. (2019) find that moral suasion had a positive effect on taxpayer’s disclosure of foreign income. Bott et al.’s treatment made salient the benefits of truthfully disclosing earnings on publicly funded institutions (e.g. education and health), which is theorized to trigger a reciprocity motive, where paying your fair share is recognized in exchange for receiving benefits from the government.

**Audits & Verification.**Slemrod et al. (2001) demonstrate that making salient that the government conducts regular audits and monitors tax contributions carefully led to more disclosure, which Bérgolo et al. (2017) suggest can be explained by tax filers’ fear of being caught out even when the audit probabilities are small. In addition, Kleven et al. (2011) find that disclosure rates increased when the probability of a known audit increased. Taken together, these papers suggest that audit threats can increase disclosure.

**Penalties.**Becker (1968) theorized that the decision to engage in fraudulent or criminal activity can be viewed as a rational trade-off between the benefits of enaging in that activity (e.g., not having to pay taxes or gaining social benefits despite earning income) and costs (e.g., the probability of getting caught and the fine one would have to pay if caught). Several studies have since tested this theory by varying the perceptions of the probabilities and severity of penalty threats.

Schwartz and Orleans (1967) provide some of the earliest evidence that penalties messages can affect tax disclosure behavior. De Neve et al. (2021) offer evidence from four large-scale experiments in Belgium that deterrence interventions can have a positive and sustained effect on tax disclosure, with the mention of escalating sanctions in follow-up letters having had larger treatment effects. Cranor et al. (2020) find that more details about the exact nature of the penalties yielded larger effects than generic penalty threats. However, not all deterrence messages have shown success: For example, Kettle et al. (2017) find that two variants of compliance messages—including one that required that tax filers actively acknowledged they understood the potential penalties they faced if they behaved unethically—did not lead to more tax compliance.

**Reminders.** Finally, since many claimants are busy and pre-occupied during unemployment, dealing with short-term needs (e.g. working to find income during this period) and long-term plans (e.g. job search), it is possible that some claimants simply forget to disclose earnings from the prior week. A reminder message may therefore be sufficient to prompt disclosure. Reminder messages have been used to encourage medication adherence (Dai et al. 2017), patient show-ups for appointments (Prasad and Anand 2012) and savings commitments (Karlan et al. 2016). In the context of disclosure, Bott et al. (2019) show that a message that included general information and reminders about disclosing foreign income led to a modest but significant increase in disclosure rates and disclosed amount.

**2. Additional information on algorithm training**

**Algorithm overview for predicting unethical behavior.** The algorithm was first developed, trained, and validated using historical data before applying it to the current study. To do so, we used an optimization procedure to create a predictive analytics model, using claimants’ behavior on the online platform to predict which claimants were most likely to be flagged for misreporting income by an adjudicator and determined to be fraudulent. The algorithm took in 122,755 prior claimants who submitted 527,854 claims which did not overlap with the study period. There was no overlap between the claimants used in the training phase of the algorithm and those used for the experiment later on, ensuring that predictions are unbiased and “honest” (Athey and Imbens, 2016).

Due to confidentiality concerns and the sensitive data used for training, we cannot disclose some specific details about the algorithm and the data used in the predictive analytics training stage of the project. However, there are several procedural steps that we are able to share. The algorithm we used was constructed using a logistic regression with a modified backward variable selection approach (e.g., Abe 2005). While standard random forest can achieve decent accuracy (Liu et al. 2015, Sahin and Duman 2011), one challenge for these types of algorithms with fraud detection is that the number of actual (detected) fraud cases, relative to all observations, is extremely low (Dal Pozzolo et al. 2015). We therefore opted for logistic regressions that do not suffer from the same issue and can be used for imbalanced datasets (e.g., Cartus et al. 2020; Oommen, Baise and Vogel 2011; Ruiz-Gazen and Villa 2008). The backward variable selection approach is a commonly used selection algorithm which reduces the dimensionality of a dataset with many predictors.

Over 300 potential predictor variables were generated to be tested for inclusion in the algorithm; variables were selected based on predictive power, completeness, stability, and reasonability. While we are not at liberty to disclose the exact variables used in the algorithm (so as to not give away information that could be used to “game” the public service system), we are able to share high-level information about the categories of variables that were used to initially train the algorithm and how many of them were retained in the final algorithm (see SI Table S5). On the one hand, the model used as input factors “obvious” predictors of unethical behavior such prior incidences of misreporting, past claim submissions, and job and industry information. On the other hand, the model also scored seemingly unrelated factors on their predictability, including the day of the week and the time of day of claim submission, which other information claimants provide during the submission process (e.g., how often they claim to have looked for employment in the past week), and any deviations from previously established habits using the online platform, as well as the interactions between all of these variables.

**Algorithm training and validation.** The outcome variable for the algorithm training and validation phase were “adjudication” outcomes, which were concluded from an adjudication process conducted by (human) government specialists. The government department has a regulated process for adjudicating and determining whether a claim is fraudulent. First, computer systems for the state routinely match each claim against new hire and wage databases reported by employers. If a claim is identified as potentially misreporting earnings the claim is passed to a human adjudicator for review and determination. In addition, adjudicators conduct stratified, random audits across different subpopulations in the state (e.g., industries, geographical counties). Once a case has been opened by an adjudicator, the process of reviewing all relevant materials, including searching employment databases and contacting employers, can take several months to years. If a claimant has been found to have committed fraud, this information is stored in the database with the claimant’s other information in the weeks where fraud was determined.

A standard three-part training/test/validation random split procedure was used (40%-30%-30%). The algorithm was iteratively trained and tested on the earliest 70% of the data (randomly divided into 40% for training and 30% for testing) with the most recent 30% of the data reserved for a hold-out validation dataset – which the algorithm had not previously seen. The results about the algorithm that we report below are on the 30% hold-out validation dataset. In the hold-out validation step, we calculated the Area Under the Curve (AUC) using the Receiver Operating Characteristic (ROC) curve (e.g. Bradely 1997; Horton 2016). We utilized the AUC statistic in evaluating model performance at multiple classification thresholds. One reason ROC is a useful choice to evaluate the performance of the algorithm (relative to a standard accuracy statistic) is because it is insensitive to class imbalance (Horton 2016). Using the ROC, the AUC for our algorithm in the hold-out sample is 0.779. Analyses of this holdout procedure show that the algorithm was able to accurately separate claimants with a greater probability of committing fraud from claimants with a lower probability (Figure S1 shows the hold-out predictions graphically).

Finally, in addition to the AUC, another way to verify that our algorithm correctly distinguished between low and high-RAR claimants, we ran a logistic regression using the algorithmically generated RAR value, grouped into bins of 10 points on the RAR scale, as the predictor variable and the human-determined fraud indicator as the dependent variable. The algorithm reliably assigned those with a greater-than-average likelihood to commit fraud to the highest RAR bins (comparing relative rates of RAR values between 80-89 and 90-100 with average fraud rate, all *p*s < 0.001), while those with lower likelihood to commit fraud were in lower-risk bins and, conversely, claimants who were less likely to do so to lower bins (comparing each of the lower RAR buckets below a RAR of 70 and average fraud rate, all *p*s < 0.001). In other words, claimants in the lowest bucket of RAR values between 1 and 10 were 89% less likely to behave unethically than the average claimant. Conversely, claimants with a RAR between 80 and 89, and claimants with a RAR between 90 and 100, behaved unethically at 74% and 256% higher rates than average.

**3. Additional Statistical analyses**

Among the messages we tested, several of them were alike, which we grouped in the analysis above (see Table S1 for all messages). Specifically, the Audits & Verification treatment contains two similar messages, while the Penalties treatment is made up of five similar messages. When we regress disclosure on the interaction of RAR and each of the two messages[[2]](#footnote-2) in the Audits & Verification treatment, compared to the control condition, we find that both the Audits message and the Verification message are independently significant for high-RAR claimants (interaction between message and RAR; Audits: *b* = 0.096, *SE* = 0.036, FDR-adjusted *p* = 0.014; Verification: *b* = 0.081, *SE* = 0.036, FDR-adjusted *p* = 0.038). This suggests both variations of the message are effective in increasing disclosure rates among high-RAR claimants.

The Penalties treatment, however, presents a less clear picture. While some messages seem more effective than others, when we regress disclosure rate on RAR interacted with each of the five Penalties interventions relative to the control group, only three reach significance without FDR-adjustments (interaction between message and RAR; Fraud: *b* = 0.082, *SE* = 0.037, *p* = 0.029; Take Action: *b* = 0.078, *SE* = 0.0257, *p* = 0.002; Lose Benefits: *b* = 0.102, *SE* = 0.051, *p* = 0.047) and only one after adjusting for multiple comparisons (Take Action message: FDR-adjusted *p* = 0.007). We therefore conservatively conclude that only the Take Action message[[3]](#footnote-3) leads to significantly increased disclosure among high-RAR claimants.

**Figure S1**. Graphical Illustration of the Effectiveness of the Machine Learning Algorithm in Identifying Claimants Most Likely to Commit Fraud



*Note*. The algorithm was tested in an independent test dataset: claimants who were predicted to behave unethically (higher RAR values, on the right-hand side of the graph) were significantly more likely to have commsitted fraud, as determined by a (human) adjudicator. Conversely, those claimants predicted to behave ethically received lower RAR values (left-hand side of the graph) and were less likely to have committed fraud.

**Figure S2.** Screenshot from the Reporting Screen and a Treatment Message in the Experiment



**Table S1.** All Behavioral Messages Used in the Field Experiment

|  |  |  |
| --- | --- | --- |
| **Name** | **Pooled?** | **Message** |
| Social Norm |  | *98 [or 99] out of 100 people in* [ClaimantCounty] *County report their earnings accurately. If you worked between* [ReportingPeriod]*, please ensure you report these earnings.* |
| Impact Others |  | *If you misreport your earnings, you may impact other unemployed people in* [U.S. State]*. If you worked between* [ReportingPeriod]*, please ensure you report these earnings.* |
| Audits | Audits & Verification  (2 variations) | [Claimant name]*: The Department of Workforce Solutions performs audits every week to verify claim accuracy. If you worked between [ReportingPeriod], please ensure you report these earnings.* |
| Verification | *We verify your employment and earnings information. The Department of Workforce Solutions has a right to recover any overpaid benefits you receive as a result of inaccurately reporting your earnings.* |
| Fraud | Penalties  (5 variations) | [ClaimantFirstName]*: Unemployment Insurance fraud is a serious offense. Penalties for fraud include:*  *• Repayment of overpaid benefits plus penalty*  *• Inability to draw benefits for up to one year*  *• Loss of income, including liens, garnished wages, or other collections actions* |
| Penalty | [ClaimantFirstName]: *Unemployment Insurance fraud is a serious offense. If it is determined that you received overpaid benefits as a result of fraud, you may be required to:*  *• Repay any overpaid benefits in full*  *• Pay an additional fraud penalty of 25% of the overpayment amount*  *• Serve a four week compensable week penalty for each week that is determined to have been fraud* |
| Repay | *If you commit fraud, you will be required to repay all benefits plus you will receive substantial penalties.* |
| Take Action | *If any balance owed is not recovered in full within 30 days, the Department of Workforce Solutions has the right to take the following actions:*  *• File a Warrant of Levy and Lien*  *• Intercept any state tax refund you are owed or take other collections actions*  *• Garnish wages and take other legal actions* |
| Lose Benefits | *If you earned money last week and intentionally do not report it, you can possibly lose unemployment benefits for up to a year.* |
| Reminder |  | *Reminder: If you worked between* [ReportingPeriod]*, you are required to report these earnings even if you have not yet been paid.* |

**Table S2.** Robustness Check: Logistic Regression

|  |  |  |
| --- | --- | --- |
|  | | |
|  | *Dependent variable:* | |
|  |  | |
|  | Likelihood of disclosure | |
|  | (1) | (2) |
|  | | |
| Baseline | -3.070\*\*\* | -3.001\*\*\* |
|  | (0.072) | (0.146) |
|  |  |  |
|  |  |  |
| Treatment (relative to baseline) | 0.110 | -0.737 |
|  | (0.077) | (0.181) |
|  |  |  |
| RAR |  | -0.001 |
|  |  | (0.003) |
|  |  |  |
| Treatment \* RAR |  | 0.013\*\*\* |
|  |  | (0.003) |
|  |  |  |
|  |  |  |
|  | | |
| Observations | 22,457 | 22,457 |
| AIC | 8,628 | 8,565 |
|  | | |

*Note*. Unit of analysis is claimants’ weekly submissions. All results are robust to changing the model: This model uses a logistic regression (instead of LPM). Standard errors are clustered at the claimant level. *P*-values are adjusted for multiple comparisons using FDR. \* *p* < .05; \*\* *p* < .01; \*\*\* *p* < .001.

**Table S3.** Robustness Check: Ten RAR Bins

|  |  |  |  |
| --- | --- | --- | --- |
|  | | | |
|  | *Dependent variable:* | | |
|  |  | | |
|  | Likelihood of disclosure | | |
|  | (1) | (2) | |
|  | | | |
| Baseline | 4.437\*\*\* | | 4.706\*\*\* | |
|  | (0.306) | | (1.168) | |
| RAR bin 2 |  | -0.305 | |
|  |  | (1.398) | |
| RAR bin 3 |  | 0.412 | |
|  |  | (1.417) | |
| RAR bin 4 |  | -0.918 | |
|  |  | (1.363) | |
| RAR bin 5 |  | 0.199 | |
|  |  | (1.392) | |
| RAR bin 6 |  | 0.006 | |
|  |  | (1.369) | |
| RAR bin 7 |  | -0.659 | |
|  |  | (1.315) | |
| RAR bin 8 |  | 0.732 | |
|  |  | (1.906) | |
| RAR bin 9 |  | -1.034 | |
|  |  | (1.815) | |
| RAR bin 10 |  | -1.335 | |
|  |  | (1.706) | |
| Treatment (relative to baseline) | 0.490 | -1.167 | |
|  | (0.336) | (1.315) | |
| Treatment \* RAR bin 2 |  | 0.938 | |
|  |  | (1.733) | |
| Treatment \* RAR bin 3 |  | -0.368 | |
|  |  | (1.598) | |
| Treatment \* RAR bin 4 |  | 1.716 | |
|  |  | (1.628) | |
| Treatment \* RAR bin 5 |  | -0.748 | |
|  |  | (1.564) | |
| Treatment \* RAR bin 6 |  | 0.302 | |
|  |  | (1.528) | |
| Treatment \* RAR bin 7 |  | 0.252 | |
|  |  | (1.497) | |
| Treatment \* RAR bin 8 |  | 0.026 | |
|  |  | (2.059) | |
| Treatment \* RAR bin 9 |  | 2.905 | |
|  |  | (1.978) | |
| Treatment \* RAR bin 10 |  | 6.320\* | |
|  |  | (1.951) | |
|  | | | |
| Observations | 22,457 | 22,457 | |
| R2 | 0.0001 | 0.006 | |
| Adjusted R2 | 0.0001 | 0.005 | |
|  | | | |

*Note*. Unit of analysis is claimants’ weekly submissions. All findings are robust to changing the RAR categorization: while there continues to be no main effect of the intervention in Model 1, we find an intervention effect in the highest bin in Model 2. This further suggests that most of the effect of the intervention is driven by the claimants predicted to be most likely to behave unethically. Standard errors are clustered at the claimant level. *P*-values are adjusted for multiple comparisons using FDR. \* *p* < .05; \*\* *p* < .01; \*\*\* *p* < .001.

**Table S4**. Exploratory Analyses of Differential Impact of Messages by (Continuous) RAR

|  |  |
| --- | --- |
|  | |
|  | *Dependent variable:* |
|  |  |
|  | Likelihood of disclosure |
|  | |
| Baseline (No Message) | 4.728\*\*\* |
|  | (0.633) |
| Social Norm | -2.307\* |
|  | (1.043) |
| Impact Others | -3.760\*\* |
|  | (1.163) |
| Audits & Verification | -5.203\*\*\* |
|  | (1.527) |
| Penalties | -3.719\*\* |
|  | (1.093) |
| Reminder | -2.638 |
|  | (1.955) |
| RAR | -0.006 |
|  | (0.013) |
| Social Norm \* RAR | 0.059\* |
|  | (0.023) |
| Impact Others \* RAR | 0.091\*\*\* |
|  | (0.027) |
| Audits & Verification \* RAR | 0.089\*\* |
|  | (0.027) |
| Penalties \* RAR | 0.058\*\* |
|  | (0.019) |
| Reminder \* RAR | 0.034 |
|  | (0.034) |
|  | |
| Observations | 0.005 |
| R2 | 0.004 |
| Adjusted R2 | 4.728\*\*\* |
|  | |

*Note*. Unit of analysis is claimants’ weekly submissions. Standard errors are clustered at the claimant level. *P*-values are adjusted for multiple comparisons using FDR. \* *p* < .05; \*\* *p* < .01; \*\*\* *p* < .001.

**Table S5.** Number of Variables During Algorithm Training and Retained in Final Algorithm

|  |  |  |
| --- | --- | --- |
|  | **Number of variables** | |
| **Category** | **Potential predictors before variable selection** | **Retained in final algorithm** |
| Past claim information | 5 | 1 |
| Current week claim information | 33 | 11 |
| Prior misreporting or audits | 22 | 1 |
| Claimant descriptives | 83 | 20 |
| Job and industry | 82 | 10 |
| Real-time behaviors on the platform | 41 | 9 |
| Geographic information | 28 | 2 |
| Time and day information | 16 | 16 |
| **Total** | **310** | **70** |

**Additional References**

Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy*, 76(2), 169-217.

Dal Pozzolo A, Boracchi G, Caelen O, Alippi C, Bontempi G (2015) Credit card fraud detection and concept-drift adaptation with delayed supervised information. *2015 International Joint Conference on Neural Networks (IJCNN)*. (IEEE, Killarney, Ireland), 1–8.

Leets L, Sprenger A, Hartman R, Kohn N, Simon Thomas J, Vu C, Aguirre S, Wijesinghe S (2020) Effectiveness of nudges on small business tax compliance behavior. *JBPA* 3(2).

Liu C, Chan Y, Alam Kazmi SH, Fu H (2015) Financial Fraud Detection Model: Based on Random Forest. *IJEF* 7(7):p178.

Oommen T, Baise LG, Vogel RM. Sampling bias and class imbalance in maximum-likelihood logistic regression. *Mathematical Geosciences*. 2011 Jan;43:99-120.

Ruiz-Gazen A, Villa N (2008) Storms Prediction: Logistic Regression vs. Random Forest for Unbalanced Data. *Working Paper*.

Sahin Y, Duman E (2011) Detecting credit card fraud by ANN and logistic regression. 2011 *International Symposium on Innovations in Intelligent Systems and Applications*. (IEEE, Istanbul, Turkey), 315–319.

1. Specifically, their most effective intervention had two noteworthy features: first, it highlighted a local (not national) comparison group and second, it pointed to a particularly high tax compliance rate (of 96%), both of which are mirrored in our Social Norm condition – high compliance rate (in our setting, the actual rate is 98% or 99% depending on the county) and the mention of a local area (the name of the county the claimant is based in). [↑](#footnote-ref-1)
2. The Audits message read: “[Claimant name]: The Department of Workforce Solutions performs audits every week to verify claim accuracy. If you worked between [ReportingPeriod], please ensure you report these earnings.” The Verification message read: “We verify your employment and earnings information. The Department of Workforce Solutions has a right to recover any overpaid benefits you receive as a result of inaccurately reporting your earnings.” [↑](#footnote-ref-2)
3. The Take Action message read: “If any balance owed is not recovered in full within 30 days, the Department of Workforce Solutions has the right to take the following actions: File a Warrant of Levy and Lien, Intercept any state tax refund you are owed or take other collections actions, Garnish wages and take other legal actions.” [↑](#footnote-ref-3)