Behavioural Public Policy (2025), 1–18
doi:10.1017/bpp.2024.50

CAMBRIDGE UNIVERSITY PRESS

FINDINGS FROM THE FIELD

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

Oliver P. Hauser[1], Michael Greene[2] and Katherine DeCelles[3]

[1]Department of Economics and Institute for Data Science and Artificial Intelligence, University of Exeter, Exeter, UK; [2]Deloitte Consulting LLP, Boston, MA, USA and [3]Rotman School of Management, University of Toronto, Toronto, Canada
Corresponding author: Oliver P. Hauser; Email: o.hauser@exeter.ac.uk

(Received 11 April 2024; accepted 28 August 2024)

Abstract
We report the results of a field experiment designed to increase honest disclosure of claims at a U.S. state unemployment agency. Individuals filing claims were randomized to a message ('nudge') intervention, while an off-the-shelf machine learning algorithm calculated claimants' risk for committing fraud (underreporting earnings). We study the causal effects of algorithmic targeting on the effectiveness of nudge messages: Without algorithmic targeting, the average treatment effect of the messages was insignificant; in contrast, the use of algorithmic targeting revealed significant heterogeneous treatment effects across claimants. Claimants predicted to behave unethically by the algorithm were more likely to disclose earnings when receiving a message relative to a control condition, with claimants predicted to most likely behave unethically being almost twice as likely to disclose earnings when shown a message. In addition to providing a potential blueprint for targeting more costly interventions, our study offers a novel perspective for the use and efficiency of data science in the public sector without violating citizens' agency. However, we caution that, while algorithms can enable tailored policy, their ethical use must be ensured at all times.

Keywords: algorithm; behavioral science; field experiment; public sector; unethical behavior; machine learning

Introduction

Public benefits fraud is common in the U.S. across services such as social security (Social Security Administration 2016) and unemployment insurance (Committee of Ways and Means, 2002). However, combatting public benefits fraud is difficult. First, those who commit public benefits fraud are often from precarious or vulnerable populations, making punitive policies controversial. Financial penalties can potentially escalate financial hardship among these populations, which can then increase the

likelihood of further fraudulent activity (Baron, 2007; Felson *et al.*, 2012). Second, investigations and audits are both labor intensive and expensive, and it is unlikely that agencies can successfully recover the amounts lost to an impoverished population; costs of detection, investigation, and recovery efforts may even cost more than the fraudulent act itself (Gustafson, 2011). Third, there is limited evidence that deterrence efforts are effective (Regev-Messalem, 2013).

Given the costs of welfare fraud are at an all-time high (estimated at $89 billion annually, see Office of Inspector General (2021)), as well as the difficulties in enforcing the ethical use of welfare services, one potential promising avenue from behavioral economics is the use of 'nudges' to try to guide individuals using these services towards more ethical behavior (Thaler and Sunstein, 2009; Grimmelikhuijsen *et al.*, 2017; John, 2018; Tummers, 2019). Governmental agencies across the globe are increasingly using 'nudging' interventions (as opposed to changing economic incentives) which are typically more cost effective for trying to achieve policy goals (Benartzi *et al.*, 2017). For instance, an intervention might leverage the use of defaults to automatically enroll individuals where it leads to socially desirable outcomes (Johnson and Goldstein, 2003). However, interventions are not always effective in changing average behavior (Hauser *et al.*, 2018). To address the question for whom an intervention works, we turn to algorithm prediction as a tool to target the intervention at those in whom behavior change is desired and most needed (Athey, 2017; Lee *et al.*, 2022).

In this paper, we propose and demonstrate that combining behavioral interventions with algorithmic prediction can help improve the usefulness of such interventions, especially in the public sector (De Vries *et al.*, 2016; Tummers, 2020). The goal of this study is to provide a direct comparison of algorithmically targeted to untargeted nudges. We illustrate this approach in the context of the public sector—a US state agency—to which claimants return on a weekly basis to claim unemployment benefits on their platform. To be eligible for unemployment benefits, claimants need to disclose (truthfully) whether and how much they earned income in the previous week. We find that the interventions causally increase disclosure, relative to a control group, but *only* among claimants who are predicted by an algorithm to behave unethically in that week. Correspondingly, these claimants who are both in the high-risk group report higher earnings when they are in the treatment group relative to being in the control group. In exploratory analyses, we investigate the variation of the effectiveness of different messages, finding consistent effects for messages that include information on social norms, impact on others, and audits and verification, while penalty messages yield some but less robust effects. Aside from encouraging more truthful disclosures, higher disclosed earnings in the treatment group have positive financial consequences for the state agency (both directly through disclosures and indirectly through saved administrative costs of recovering owed money).

Our work contributes to three streams of literature. First, an emerging body of research studies uses behavioral (or 'nudging') interventions to reduce unethical behavior. Nudges are interventions that 'alter people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives' (Thaler and Sunstein, 2009, p. 6). Ample extant literature documents that such interventions can change behavior in the public and private sector (Thaler and Sunstein, 2009; Benartzi *et al.*, 2017). In some cases, nudges have helped to achieve policy outcomes efficiently and cheaply, such as increasing energy savings (Allcott, 2011;

https://doi.org/10.1017/bpp.2024.50 Published online by Cambridge University Press

Allcott and Rogers, 2014; Jachimowicz *et al.*, 2018), savings (Karlan *et al.*, 2016; Kessler *et al.*, 2019), vaccination uptake (Milkman *et al.*, 2011), diversity in hiring (Arslan *et al.*, 2024) and giving to charity (Barasz *et al.*, 2017). At the same time, nudge-based interventions are not always effective (Sunstein, 2017; Hauser *et al.*, 2018).[1] Studies have found null results for similar nudges in different contexts (e.g. Kettle *et al.*, 2017, Bird *et al.*, 2021) and sometimes nudges have even led to opposite effects of those predicted by researchers and policy experts (Beshears *et al.*, 2015; Robinson *et al.*, 2019), prompting new theories of ways that nudging can be improved (e.g. Nudge +, see Banerjee and John, 2021). Even for relatively well-studied interventions, such as penalties threats and deterrence messages (Ariel, 2012; De Neve *et al.*, 2021) or social norms (Larkin *et al.*, 2019), results can be mixed. Consequently, relatively little is known about the conditions under which interventions are most likely to succeed. In the current study, we examine one potential avenue for increasing the effectiveness of interventions to achieve policy outcomes: the use of predictive algorithms to identify subgroups most likely to benefit from receiving an intervention.

Second, we draw on the fast-growing literature on algorithmic targeting, while also advancing this literature by adding causal investigations in a real-world field setting (Hofman *et al.*, 2021; Lee *et al.*, 2022). Predictive algorithms are already commonly used to analyze user and consumer behavior (Webb *et al.*, 2001) and, in the context of unethical behavior, these algorithms have mostly been employed for the detection of consumer credit card fraud and firm-level financial statement fraud (Ngai *et al.*, 2011; West *et al.*, 2015). They also have the potential to be used to increase social welfare through optimal policy design (Balaguer *et al.*, 2022; Koster *et al.*, 2022, 2024). The idea that such prediction algorithms can be used to identify people who might benefit from an intervention has recently received increasing attention (Athey and Imbens, 2016; Xiao *et al.*, 2024), in particular in marketing. Matz *et al.* (2017) use Facebook 'Likes' to infer 'Big Five' personality types, increasing response to targeted advertisements, whilst Matz *et al.* (2023) demonstrate a similar targeting approach for savings behaviors. Ghose *et al.* (2019) show that fine-grained location data from cell phones can be used to target advertisements at nearby shoppers, leading to greater purchasing, while Peˆer *et al.* (2019) demonstrate that customizing nudges to suit a person's decision-making style can improve their effectiveness. Hauser *et al.* (2009) demonstrate that a website that automatically 'morphs' its design to suit different cognitive styles leads to more sales. In sum, research demonstrates algorithms are promising to selectively target individuals to increase sales.

Our final contribution relates to the literature on behavior change by focusing on the use of targeting as a means to achieve the most scope for behavior change. In our context, we demonstrate that our treatments only have an effect among claimants who are predicted to behave unethically. This should be unsurprising: those claimants had something to disclose in the first place. However, all too often, nudges in the field are

[1]Note, however, that lack of behavior change is not unique to 'nudges' or behavioral interventions; indeed, standard economic interventions, such as redistribution through cash transfers (Jaroszewicz *et al.*, 2022) or medical debt relief (Kluender *et al.*, 2024), may also result in unexpected (lack of) behavior change. Expert predictions are increasingly being employed to assess the research community's priors about interventions before they are tested (DellaVigna *et al.*, 2019).

https://doi.org/10.1017/bpp.2024.50 Published online by Cambridge University Press

applied to the whole population, but doing so might underestimate the actual treatment effect among the *relevant* population. This point was raised in early field experiments by Slemrod *et al.* (2001) and Blumenthal *et al.* (2001) who both observe heterogeneous treatment effects among populations who had more (versus less) of an opportunity to evade taxes. Relative to these early papers which focused on heterogeneous treatment effects by demographic and tax group characteristics *post hoc*, our study offers a more generalizable, *a prior* targeting approach with machine learning (Lee *et al.*, 2022): by targeting claimants algorithmically, we do not need to impose any structure or assumptions on which variables are most likely to be predictive of unethical behavior; we can take advantage of non-linearities and interactions of variables that might be predictive; and we introduce an approach by which claimants can be targeted in real time and, therefore, the interventions can be employed efficiently for only the relevant subgroup. This combination of algorithmic targeting and causal investigations under one umbrella falls into a relatively understudied area of computational social science (Hofman *et al.*, 2021). Our paper shows a first proof-of-concept of combining algorithmic targeting with a causal study of the effectiveness of 'nudging' messages in the public sector in the field, which we hope will be a blueprint for future studies and for policy-makers.

Methods

Study context. The field context for our study is the U.S. public sector: in collaboration with an unemployment office in a U.S. state, we conducted a large-scale field experiment to attempt to reduce dishonesty in unemployment claims. This context involves a population of unemployed workers who claim state unemployment benefits, an understudied population. This population is likely without significant wealth, creating a predicament for them in filing accurate unemployment claims, which can reduce their income when filing honestly. Yet, an individual who is caught filing a fraudulent claim also incurs financial penalties and potential exclusion from the system as well as legal trouble, placing an already vulnerable population in even greater financial and employment difficulty. Furthermore, the state has an incentive to minimize fraudulent claims prospectively as it reduces government spending on monitoring and recouping fraudulent claims.

The experiment was conducted on the government's online platform with all submissions over a three-week period in August–September of 2016, for a total of 9,833 unique claimants and 22,457 submissions. This type of unobtrusive procedural field experiment (Harrison and List, 2004; Hauser *et al.*, 2017) can help to reduce potential concerns regarding demand effects and self-selection. Claimants were randomized into control (where they received no message) and treatment, which is comprises one of several 'nudging' messages (Table 1), based on the extant literature (see Supplementary Information (SI) Section 1) and to which claimants were randomly assigned within the treatment condition.

Algorithmic targeting. We employed an off-the-shelf machine learning algorithm that calculates each week a Risk Assessment Result (RAR) value (as a continuous value from 1 to 100) for each claimant who submitted an unemployment claim to the state government. The RAR is a measure of likelihood of behaving fraudulently in any given

https://doi.org/10.1017/bpp.2024.50 Published online by Cambridge University Press

Table 1. Examples of intervention messages used in the field experiment. See Table S1 in the appendix for the full list of messages

Name	Example message	Example from prior literature
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.*	Schultz *et al.* (2007); Allcott (2011); Fellner *et al.* (2013); Hallsworth *et al.* (2017)
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.*	Blumenthal *et al.* (2001); Chirico *et al.* (2016); Bott *et al.* (2019); De Neve *et al.* (2021)
Audits & Verification (2 variations)	*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.*	Slemrod *et al.* (2001); Kleven *et al.* (2011); Fellner *et al.* (2013); Hallsworth *et al.* (2017); Kasper and Alm (2022)
Penalties (5 variations)	*If you commit fraud, you will be required to repay all benefits plus you will receive substantial penalties.*	Kettle *et al.* (2017); Cranor *et al.* (2020); De Neve *et al.* (2021); Yogama *et al.* (2024)
Reminder	*Reminder: If you worked between* [ReportingPeriod], *you are required to report these earnings even if you have not yet been paid.*	Karlan *et al.* (2016); Hallsworth *et al.* (2017)

week, based on the predictive algorithm. Claimants are assigned a value by the algorithm from 1 to 100, with higher RAR values indicating a greater likelihood of behaving fraudulently. The RAR takes into account various historic and real-time behavioral variables (e.g. type and date of submission, behaviors on the platform, last employment and industry, etc.). A new RAR value is calculated each week, and can thus vary within a claimant over time (whereas the randomized assignment to treatment or control group does not change over time).

The algorithm returns a RAR value that is calculated each week and varies from 1 to 100 for each claimant who submitted an unemployment claim to the state government. We later created 'RAR bins' for ease of presentation, but our results hold regardless of whether the RAR value is categorized into bins or on a continuous scale. The RAR value is the variable we use in the field experiment to conduct analyses of the causal treatment effect of the messages in each RAR bin, where high RAR values imply that the algorithm predicts that the claimant is more likely to behave unethically at that time.

Further details on the algorithmic training procedure can be found in SI Section 2. Note that, due to the sensitive nature of the data and the research question in collaboration with a public sector field partner, several details of the algorithm are confidential and proprietary. Specifically, the algorithm and data used in the training stage are confidential due to potential concerns by the field partner that revealing these details could lead to 'gaming the system.' However, we can share several procedural steps in how we arrived at the final algorithm (see SI Section 2, Figure S1 and Table S5). We also emphasize that the algorithm used here is neither the primary goal, nor of key interest, of this research; we relied on an off-the-shelf, carefully calibrated machine learning algorithm that uses sensitive data and variables that we are not allowed to disclose at this point. Future research may choose to apply a similar algorithmic targeting approach by choosing the appropriate statistical algorithm based on their research or policy question and the data at hand, and by calibrating it for the specific context.

Behavioral interventions. Table 1 summarizes the messages used in the experiment. These messages can be classified as 'nudges' because, while they may include information about the presence or severity of a sanction, the government's ability to take action against a fraudulent claimant, or simply information about the behavior of others (i.e. social norms), none of these messages materially change the underlying incentive structure. In all conditions (control or treatment) in our field context, claimants are required by law to truthfully disclose information as requested by the state agency and failing to do so can result in fines, penalties, and other sanctions. The nudge therefore does not change the 'economic incentives' (Thaler and Sunstein, 2009, p. 6) and preserves the decision-makers' ultimate choice but it may change the likelihood that claimants make a disclosure if they have earnings to disclose.

Benefits claim setup and process. When a claimant loses their job and is eligible for unemployment insurance, they set up an online account with the government agency responsible for distributing unemployment benefits. (A fraction of claimants (approx. 9% in our data) do not set up an online account, either because they do not have access to a computer or do not feel comfortable navigating the system online. They can instead have an in-person or phone conversation with a government official each week to certify their eligibility. Because they do not use the online platform, they are not part of

https://doi.org/10.1017/bpp.2024.50 Published online by Cambridge University Press

our experiment.) Once the account is set up, they need to start an 'initial claims' process. The randomization took place for each claimant at the time when they started the initial claim. Once the initial claim is approved by the government agency, claimants need to return to the online portal on a weekly basis to apply for unemployment benefits, through a process known as 'weekly certification.' Claimants are eligible if they do not hold a regular job and have not earned more (through, for example, occasional jobs) than the amount of the unemployment benefits in the week prior to completing the certification. If a claimant has earned money in the previous week, the amount of unemployment benefits they are eligible for this week is reduced accordingly.[2]

Treatment delivery. For claimants in the treatment group, the treatment is incorporated into the weekly eligibility certification process (see Figure S2 in the SI for a screenshot of an illustrative message in the treatment). Claimants in the treatment are exposed to a popup message on the same page during the eligibility certification process, after they clicked the submit button of the form where our dependent variable was collected. Specifically, in the treatment group, the popup was shown after participants had selected 'no earnings' and submitted the page. This meant that behavioral information collected on this page could also be fed back into the algorithm, which calculated the RAR value as the page was submitted based on both historic and real-time data; showing the popup as the page loads would have been another possibility but would not have incorporated the most recent behavioral patterns into the RAR calculation, which was an important consideration to the field partner.

Once randomized, claimants in the treatment group see the same intervention every week when they came back to complete the eligibility certification process (and they see no message if they are in the control group). This means that claimants can see the same message multiple times for a few weeks in a row. The number of times a claimant in the treatment group is exposed to the treatment is a function of the number of times they return to claim unemployment benefits. Conversely, if they were assigned to the control group, they do not see any message at any point, no matter how often they return to the online platform to claim unemployment benefits.

Dependent variable in the field experiment. The dependent variable, which we refer to as 'disclosure,' is whether claimants answered 'yes' versus 'no' in response to the question 'Did you work during the reporting period listed above?' during the weekly certification process. (The 'reporting period' refers to two dates, spanning the beginning to the end of the previous week, at the top of the page.) Claimants who answered 'yes' to this question are asked to indicate the amount earned. Of course, selecting 'no income to report' during the weekly certification process does not automatically equate to fraud. Only those who did work, but who deliberately do not disclose their earnings, are behaving unethically. If the claimant reported earnings, the amount is deducted from the weekly payment. Not disclosing earned income could be considered fraud, illegal, and punishable by state law, and the state has multiple mechanisms for determining actual earnings. For example, employers report wages to the state and new employment relationships are included in a database. (There is a lag between when

[2]We acknowledge that this summary is a simplification of the certification process. There are more nuances involved in this process that go beyond the scope of this paper but they do not affect the experiment.

https://doi.org/10.1017/bpp.2024.50 Published online by Cambridge University Press

the earnings are reported by the claimant and employer.) Overpayments must be paid back in full. If, in addition, the intention behind the misreporting is deemed fraudulent by the state it is considered fraud and additional penalties apply. However, in the absence of an audit, there is no way to know whether a particular individual behaved unethically when they reported no earnings. While full audit would be the most accurate and preferential outcome variable, they are only conducted in few cases and the available audit data is too small for the duration of this study.

Therefore, while the 'disclosure' outcome does not necessarily measure fraudulent behavior on an *individual* level, it is a useful *proxy* of fraudulent behavior when aggregated on a higher level (i.e. condition level): That is, with a randomized experimental design, we can observe whether the treatment increased reporting *relative* to the control group. Since claimants are randomly assigned to conditions and the only difference between them is the treatment, any difference in disclosure rates is the result of the intervention. This condition-level proxy measure of fraudulent behavior in a field experimental design mirrors the commonly used condition-level measures of dishonest in laboratory settings (e.g. die-roll task, see Fischbacher and Föllmi-Heusi, 2013, or matrix task, see Gino *et al.*, 2009).

Statistical analysis in the field experiment. Results are estimated using a linear probability model (LPM) predicting the likelihood of disclosure with robust standard errors clustered at the claimant level. Our results are robust to using variation in econometric specifications (e.g. logistic regressions). To help with interpretation and ease of reading of the coefficients in the regression tables, the disclosure outcome variable is represented as a percentage (0–100) rather than a fraction (0–1).

In all our analyses, we take a conservative approach by adjusting for multiple comparisons. We apply the False Discovery Rate (FDR) procedure to reduce the likelihood of detecting false-positives. Introduced by Benjamini and Hochberg (1995), the FDR is one of the most widely used procedures to adjust p-values when multiple tests are being performed. The FDR provides more statistical power than Bonferroni-type corrections while ensuring that p-values (also sometimes called q-values after application of the FDR) are adjusted on the basis of a cut-off value. Modern implementations of FDR algorithms automatically determine the cut-off based on the empirical distribution of p-values for the given context. We use the 'p.adjust' function in the 'stats' core package in R. All our regressions that involve multiple comparison groups (e.g. different messages or multiple RAR bins) have been adjusted using this FDR procedure.

This research was approved by the ethics board at Harvard University (IRB16-0813).

Results

Untargeted average treatment effect. We begin by looking at the average treatment effect (i.e. without the use of algorithmic targeting). We start with this analysis because it represents the status quo of the kind of empirical analysis one would conduct when analyzing such a field experiment in the absence of any algorithmic targeting. This analysis reveals what the treatment effect is for an intervention across the board for everyone, not taking into consideration the targeting approach we have outlined above and which we analyze below.

https://doi.org/10.1017/bpp.2024.50 Published online by Cambridge University Press

Using a LPM, we regress the disclosure rate by the treatment status, testing whether the treatment (relative to control) increases the average likelihood of disclosure across all claimants. While the coefficient on treatment is positive, there is no significant main effect of the intervention relative to the control group (see Table 2 Column 1: $b = 0.490$, $SE = 0.336$, $p = 0.145$).

This result, thus far, suggests that the average behavior in the treatment group is not significantly different from the average behavior in the control group in the absence of algorithmic targeting. We note that, if this study were conducted in this 'traditional' fashion (i.e. with randomly assigning nudge messages across all claimants but without application of a targeting algorithm), a policy-maker would likely conclude at this stage that the interventions had failed (or, at least, that there is a lack of evidence for the treatment being effective). This is, however, where we turn to the use of algorithmic targeting to illustrate the added value of using such an approach.

Algorithmic-targeted treatment effects. We next consider the role of the algorithmic targeting approach based on the real-time generated RAR variable. The analytical approach is as follows: we take advantage of the random assignment of claimants to control or treatment groups, and then interact the treatment assignment with the RAR variable. The RAR variable captures the extent to which a claimant is predicted to behave fraudulently (with higher values indicating a greater likelihood of behaving fraudulently).

Note that, in the *absence* of the control group, the use of the RAR variable would simply be targeting nudges for everyone and would not allow for a *causal* estimation of the treatment and targeting approach; however, because a control group is present (i.e. a group of claimants across all RAR values, for whom no nudge is shown), we can compare claimants with a similar RAR values between the control and treatment groups, in order to identify the causal effect of the nudges relative to the control for those specific claimants.[3]

Using LPM with both treatment assignment and the RAR value as independent variables, we find that there exists a significant interaction between the treatment and RAR (Table 2 Column 2; $b = 0.056$, $SE = 0.015$, $p < 0.001$).[4] As Figure 1 illustrates, the treatment effect is positive and significant *only* for claimants with the highest RAR values, suggesting that the treatment increased disclosure among claimants that were predicted to behave unethically.

[3]In some sense, this approach is similar to a conventional heterogeneity analysis by separating those with high and low RAR value from each other and running the analysis separately for each subgroup or as an interaction between the treatment and the subgroups. However, while heterogeneity analyses are typically conducted post-hoc (i.e. after the study has finished for exploratory purposes), the RAR values in our setting are assigned in 'real time' and no segmentation or analysis after the study would be necessary in a real policy context: in fact, the ultimate goal of having RAR values calculated on the fly is to use them for targeting in real time and target the nudges only at claimants who have high RAR values. For the purposes of our study, however, we believe it is instructive to analyse the causal effect of the combination the algorithmic targeting and nudges, which is why we adopt this analytical framework.

[4]Note that, in this specification, the treatment effect for the lowest RAR values is significantly negative ($b = -2.941$, $SE = 0.710$, $p < 0.001$): while intriguing, this negative effect only appears in this specification and does not replicate in other specifications that categorizes RAR values into bins (see Column 3 in Table 2 and Table S3 for regressions using discretization of 5 RAR bins and 10 RAR bins, respectively).

https://doi.org/10.1017/bpp.2024.50 Published online by Cambridge University Press

Table 2. Differential impact of intervention on the likelihood of disclosure by RAR. Unit of analysis is claimants' weekly submissions. Column 1 is the main effect without algorithmic targeting by RAR. Column 2 studies the effect of algorithmic targeting and finds an interaction effect between treatment and RAR (as a continuous variable). Column 3 corroborates this finding by showing that the treatment effect is concentrated in the 5th RAR bin (with RAR values between 81 and 100). Standard errors are clustered at the claimant level. P-values are adjusted for multiple comparisons using FDR.

	Dependent variable:		
	Likelihood of disclosure		
	(1)	(2)	(3)
Baseline	4.437***	4.728***	4.520***
	(0.306)	(0.633)	(0.715)
Treatment (relative to baseline)	0.490	−2.941***	−0.613
	(0.336)	(0.710)	(0.800)
RAR		−0.006	
		(0.013)	
Treatment * RAR		0.056***	
		(0.015)	
RAR bin 2			−0.098
			(0.891)
RAR bin 3			0.285
			(0.890)
RAR bin 4			−0.154
			(0.917)
RAR bin 5			−0.999
			(1.197)
Treatment * RAR bin 2			0.174
			(1.047)
Treatment * RAR bin 3			−0.754
			(0.997)
Treatment * RAR bin 4			0.092
			(1.035)
Treatment * RAR bin 5			4.158*
			(1.346)
Observations	22,457	22,457	22,457
R^2	0.0001	0.003	0.004
Adjusted R^2	0.0001	0.003	0.004

Note: *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

By way of example of how the algorithm moderates the effect of the treatment, consider the disclosure rates in the control and treatment groups in lowest and the highest RAR bins (Figure 1). In the lowest RAR bin (with RAR values 0-19), the control and

https://doi.org/10.1017/bpp.2024.50 Published online by Cambridge University Press

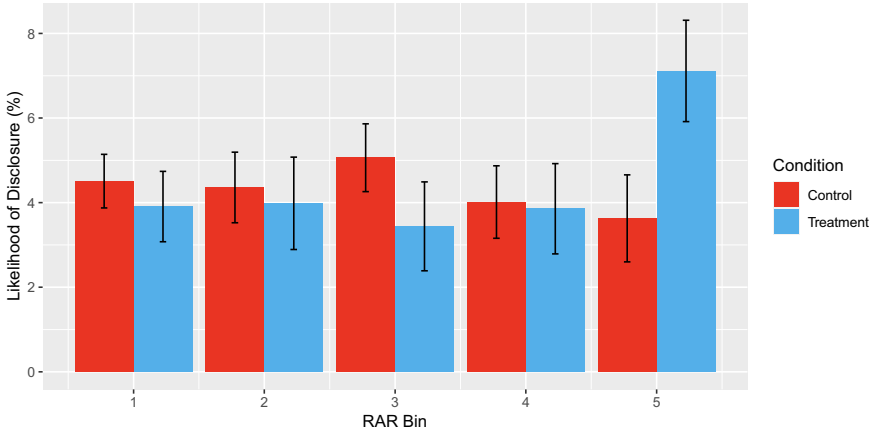

Figure 1. Intervention only increases disclosure among claimants with high RAR values. Relative to the control group (red bars), the treatment (blue) significantly increases the likelihood of disclosure among claimants in the highest RAR bin but not among claimants in lower RAR bin. Error bars are standard errors from the mean (clustered at the claimant level).

treatment groups have comparable disclosure rates of 4.52% and 3.91%, respectively. In contrast, disclosure rates of the control and treatment groups in the highest RAR bin (with RAR values 80–100) are notably different: relative to the control group (3.52%), the treatment doubles disclosure rates (7.07%) among 'high-risk' claimants. Since, by design, all claimants are part of our randomized field experiment, this difference in disclosure rates in the top bin can be attributed to the causal effect of the treatment.

Robustness analyses. We conduct several robustness checks on our data. First, in our main specification, we used a LPM to estimate the effect of the treatment on disclosure rates. However, since the dependent variable is binary, a logistic regression is another common analytical approach in the literature. We initially chose LPM over a logistic regression approach for our main analysis since logistic models can be problematic for estimates of causal treatment effects, especially when interaction terms are involved (Ai and Norton, 2003), making linear models a safer choice (Gomila, 2020). However, as we show in Table S2, our results are robust to an alternative choice of statistical model.

Next, to understand where along the RAR spectrum this treatment effect is concentrated, we repeat the same analysis using a discretization of the RAR spectrum into five equally-sized 'bins' (i.e. 0–19, 20–39, 40–59, 60–79, and 80–100). We find that the treatment effect is driven exclusively by an increase in disclosure rates among the treatment group in the highest RAR bin ($b = 4.158$, $SE = 1.346$, FDR-adjusted $p = 0.010$). (This categorization is also used in Figure 1).

In addition, we further explore a finer discretization by creating 10 equally spaced bins along the RAR spectrum. As shown in Table S4 in the appendix, these results demonstrate that the treatment effect is especially pronounced in the highest RAR bin (i.e., RAR values 90–100) where the disclosure rate in the treatment group is 2.5 times larger than in the control group (8.52% vs. 3.37%, respectively).

https://doi.org/10.1017/bpp.2024.50 Published online by Cambridge University Press

Exploratory analysis of individual messages. In this section, we examine the variation in treatment messages on the likelihood of disclosure. We acknowledge that this is an exploratory analysis and as such, our findings in this section should be interpreted with caution. To avoid spurious findings, we take two steps. First, we adjust for multiple comparisons in all our regressions using the FDR procedure described above. Second, we conduct the same analyses as above where we look at the interaction between the treatment messages and RAR.

We consider two ways that RAR could be operationalized, first as a continuous variable and, second, RAR discretized into five bins. The second analysis is more conservative than the first because, by virtue of having multiple RAR bins, we also adjust for more multiple comparisons. We only conclude that a message is reliably effective for high-RAR claimants if both analyses agree. That is, we require that both the interaction between the message and the continuous RAR as well as the interaction between the message and discretized highest-RAR bin are significant after FDR-adjusting for multiple comparisons, for us to conclude that the algorithm-moderated treatment effect is real and robust. We believe that this procedure ensures that false positives are minimized.

The results of the two complementary analyses can be summarized as follows: the Social Norm, Impact Others, Audits & Verification and Penalties message are significant *only* for claimants with the highest RAR values in both analyses (FDR-adjusted ps < 0.05 in Column 5 in Table 3, which uses discrete RAR bins; see also Table S4, which uses the continuous RAR variable). In contrast, the Reminder message is not significant in either analyses. In sum, after adjusting for multiple comparisons in both analyses, our results suggest that the Social Norm, Impact Others, Audits & Verification and Penalties messages significantly increase disclosure rates among high-RAR claimants, whereas the Reminder message does not. (For further analyses on minor variations in message, see SI Section 2.)

Discussion

While nudges have been shown to be practical and cost-effective in many domains (Benartzi *et al.*, 2017), they have not consistently replicated across different settings (Hummel and Maedche, 2019). Here, we proposed that nudges can be made more effective with algorithmic targeting. Providing experimental evidence of nudge interventions that worked in the field helps strengthen the external validity of interventions, which is both theoretically important and policy relevant (Hauser *et al.*, 2017). We created an algorithm to detect likely unethical behavior, which then enabled us to create and direct targeted interventions: The interventions that significantly increased disclosure among high-risk claimants included impact on others, social norms, and audits and verification messages, consistent with recent studies in other domains (e.g. Kleven *et al.*, 2011; Hallsworth *et al.*, 2017; Bott *et al.*, 2019).

By targeting interventions at claimants with the most opportunity to evade, our results may also shed light on some contradicting findings in the past literature, where one group of scholars may have found that these interventions work while others have not found an effect. For example, in our sample of claimants of social benefits, if we had only considered the impact of the interventions on the *average* claimant, one

https://doi.org/10.1017/bpp.2024.50 Published online by Cambridge University Press

Table 3. Exploratory analyses of differential impact of messages by discretized RAR. Linear probability model predicting disclosure by type of message interacted with RAR bin. For ease of reading, the regression table is presented such that the regression coefficients of the interactions between each RAR bin and each message are shown in the columns and rows, respectively. Specifically, the coefficients in Column 1 in the table are the baseline effects in RAR bin 1 while Columns 2–6 show the interaction coefficients between the message and the corresponding RAR bin. Standard errors are clustered at the claimant level. *P*-values are adjusted for multiple comparisons using FDR.

	Dependent variable:				
	Likelihood of disclosure				
	(1) RAR bin 1	(2) RAR bin 2	(3) RAR bin 3	(4) RAR bin 4	(5) RAR bin 5
Baseline (control)	4.520***	−0.098	0.285	−0.154	−0.999
	(0.715)	(0.892)	(0.890)	(0.918)	(1.198)
Social norm	−0.362	0.413	−1.183	1.043	5.915*
	(1.066)	(1.400)	(1.341)	(1.495)	(2.151)
Impact others	0.656	−1.787	−2.600	0.292	8.499*
	(1.158)	(1.423)	(1.415)	(1.618)	(2.823)
Audits & verification	−1.871	1.295	0.047	1.135	7.040*
	(1.480)	(2.109)	(1.727)	(1.730)	(2.265)
Penalties	−1.587	1.549	1.129	0.148	4.132*
	(1.084)	(1.510)	(1.468)	(1.327)	(1.562)
Reminder	−2.051	0.690	1.214	1.435	2.955
	(2.537)	(3.143)	(3.277)	(3.223)	(3.215)

Note: *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

might have assumed that the treatment had not worked; however, it would be wrong to assume that the treatment should work on claimants who have already made an honest decision of whether earnings needed to be disclosed (which is the vast majority of claimants). We therefore *a priori* employed an algorithm to identify which claimants would be most likely to behave fraudulently and *only* predicted a treatment effect to occur for this relevant population. Future researchers may benefit from targeting more precisely claimants who they believe are able to change their behavior, rather than applying an intervention to all available claimants.

Our results had real-world, financial implications for the public sector organization. While the treatment did not produce an average significant effect on disclosure and disclosed earnings, it did show significant increases in disclosure rates for high-risk claimants. Based on back-of-the-envelope estimates, we can approximate the additional earnings that the government was able to collect by exposing the highest RAR claimants to the treatment. During the three-week experimental period, the treatment was observed 5,440 times by claimants in the highest RAR bin, leading to approximately $23,011 in earnings disclosed by claimants, which would otherwise not have been collected by the government. Extrapolating from these numbers, the treatment could generate up to $400,000 per year in additional revenue that claimants would otherwise not disclose voluntarily and would be lost to the government. In addition, further savings for the public administration would also result from not having

https://doi.org/10.1017/bpp.2024.50 Published online by Cambridge University Press

to investigate and recover any earnings from claimants who have not truthfully disclosed such earnings. However, this rough back-of-the-envelope calculation does not consider potential general equilibrium effects, including potential habituation to the treatments or learning effects, which could affect the response to the treatment and remains an important area of future research. Furthermore, both our research design and the cost savings implications are contextual and may vary depending on the specific behavior to be nudged, the algorithm's ability to predict such behavior, and the success of the interventions used. To generalize to other contexts, we encourage future researchers to apply similar approaches that go beyond prediction alone and aim for behavior change (Athey, 2017), while also rigorously evaluating the policy outcomes and cost savings of such interventions (H.M. Treasury, 2007).

Finally, our paper also offers proof of concept and practical contributions surrounding the effectiveness of algorithms to detect risk levels of unethical behavior over time. Algorithmic targeting is fast becoming ubiquitous, from social media to savings to health (Benartzi *et al.*, 2017; Matz *et al.*, 2017, Nahum-Shani *et al.*, 2018), but the same algorithms have not been used widely in identifying unethical behaviors. However, future applications can be imagined: identifying illegal transactions, employee mistreatment, or even domestic abuse. By identifying potentially unethical subgroups using an algorithm and applying interventions to those subgroups only may provide a practical and non-invasive means to change, or even pre-empt, unethical behavior. We believe the greatest use of this application is where nudges are costly to organizations or cumbersome for users, or where the majority of benefits are concentrated in a small subset of individuals (Einav *et al.*, 2018). Recent advances with generative AI technology, such as large language models, further add to the possibility of greater personalization of persuasive messages (Matz *et al.*, 2024) or other uses of creative outputs co-created with AI (Doshi and Hauser, 2024).

However, it is critical to note that there are important ethical considerations at play in organizations'—and especially the public sector's—use of machine learning to detect risk of unethical behavior. On the one hand, nudges have been compared in their use to manipulation to achieve certain behavioral change (e.g., see Wilkinson, 2013). Artificial intelligence attracts similar worries: The in-depth monitoring, the use of dynamic artificial intelligence systems, and attempts to subtly influence behavior by organizations could all be manipulative, autonomy-reducing, and invasive. On the other hand, nudging can improve welfare and, by not constraining choice, this approach emphasizes the importance of preserving freedom of choice in attempts to influence behavior (Sunstein, 2015). Combining nudges with algorithms can potentially resolve this argument: while the algorithm in our study identified individuals with elevated risk of unethical behavior, it was not interpreted by the government as unethical behavior itself (and no action was taken), leaving the decision and behavior up to the individual. Importantly, the way in which algorithms are described to citizens can affect their acceptance, which is a particularly important consideration for their applications in the public sector (Sunstein and Reisch, 2023).

It is important to acknowledge that both the use of machine learning and nudges could potentially be used in nefarious ways, such as identifying individuals for the purposes of discrimination (Wang and Kosinski, 2018). While responsible use of these techniques has the power to decrease the negative effects of human biases, temptations,

https://doi.org/10.1017/bpp.2024.50 Published online by Cambridge University Press

and assumptions, private and public sector organizations must educate themselves on the possibilities for negative side effects or misguided use. Furthermore, regulators need to step up to oversee and manage what is currently an unregulated and potentially problematic tool that can be used on vulnerable populations. As organizations and governments embrace the use of machine learning and algorithms, a set of standards and rules ought to be developed, maintained and observed, ensuring transparency of the algorithms used and individuals' privacy respected (Horvitz and Mulligan, 2015; Athey, 2017).

Supplementary material. To view supplementary material for this article, please visit https://doi.org/10.1017/bpp.2024.50.

Acknowledgements. We are grateful to the New Mexico Department of Workforce Solutions (DWS) and Deloitte for their collaboration and support. In particular, we wish to thank Sue Anne Athens (DWS), Celina Bussey (formerly DWS), Joy Forehand (formerly DWS), Jim Guszcza (formerly Deloitte), and Scott Malm (Deloitte) for their instrumental support in making this study succeed. For the purpose of open access, the authors have applied a 'Creative Commons Attribution' (CC BY) licence to any Author Accepted Manuscript version arising.

References

Ai, C. and E. C. Norton (2003), 'Interaction terms in logit and probit models', *Economics Letters*, **80**(1): 123–129.

Allcott, H. (2011), 'Social norms and energy conservation', *Journal of Public Economics*, **95**(9–10): 1082–1095.

Allcott, H. and T. Rogers (2014), 'The short-run and long-run effects of behavioral interventions: experimental evidence from energy conservation', *American Economic Review*, **104**(10): 3003–3037.

Ariel, B. (2012), 'Deterrence and moral persuasion effects on corporate tax compliance: findings from a randomized controlled trial', *Criminology*, **50**(1): 27–69.

Arslan, C., E. H. Chang, S. Chilazi, I. Bohnet and O. P. Hauser (2024). 'Just-in-time diversity training leads to more diverse hiring in a global engineering firm'. *Working Paper*.

Athey, S. (2017), 'Beyond prediction: using big data for policy problems', *Science*, **355**(6324): 483–485.

Athey, S. and G. Imbens (2016), 'Recursive partitioning for heterogeneous causal effects', *Proceedings of the National Academy of Sciences of the United States of America*, **113**(27): 7353–7360.

Balaguer, J., R. Koster, C. Summerfield and A. Tacchetti (2022), 'The good shepherd: an oracle agent for mechanism design', arXiv preprint arXiv:2202.10135.

Banerjee, S. and P. John (2021), 'Nudge plus: incorporating reflection into behavioral public policy', *Behavioural Public Policy*, **8**: 1–16.

Barasz, K., L. K. John, E. A. Keenan and M. I. Norton (2017), 'Pseudo-set framing', *Journal of Experimental Psychology: General*, **146**(10): 1460–1477.

Baron, S. W. (2007), 'Street youth, gender, financial strain, and crime: exploring broidy and agnew's extension to general strain theory', *Deviant Behavior*, **28**(3): 273–302.

Benartzi, S., J. Beshears, K. L. Milkman, C. R. Sunstein, R. H. Thaler, M. Shankar, W. Tucker-Ray, W. J. Congdon and S. Galing (2017), 'Should governments invest more in nudging?', *Psychological Science*, **28**(8): 1041–1055.

Benjamini, Y. and Y. Hochberg (1995), 'Controlling the false discovery rate: a practical and powerful approach to multiple testing', *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**(1): 289–300.

Beshears, J., J. J. Choi, D. Laibson, B. C. Madrian and K. L. Milkman (2015), 'The effect of providing peer information on retirement savings decisions: peer information and retirement savings decisions', *The Journal of Finance*, **70**(3): 1161–1201.

Bird, K., B. Castleman, J. Denning, J. Goodman, C. Lamberton and K. O. Rosinger (2021), 'Nudging at scale: experimental evidence from FAFSA completion campaigns', *Journal of Economic Behavior & Organization*, **183**: 105–128.

Blumenthal, M., C. Christian, J. Slemrod and M. G. Smith (2001), 'Do normative appeals affect tax compliance? evidence from a controlled experiment in minnesota', *National Tax Journal*, **54**(1): 125–138.
Bott, K. M., A. W. Cappelen, S. Eø and B. Tungodden (2019), 'You've Got Mail: A Randomized Field Experiment on Tax Evasion', *Management Science*, **66**(7): 2801–2819.
Chirico, M., R. P. Inman, C. Loeffler, J. MacDonald and H. Sieg (2016), 'An experimental evaluation of notification strategies to increase property tax compliance: free-riding in the city of brotherly love', *Tax Policy and the Economy*, **30**(1): 129–161.
Committee of Ways and Means (2002), *Report*.
Cranor, T., J. Goldin, T. Homonoff and L. Moore (2020), 'Communicating tax penalties to delinquent taxpayers: evidence from a field experiment', *National Tax Journal*, **73**(2): 331–360.
DellaVigna, S., D. Pope and E. Vivalt (2019), 'Predict science to improve science', *Science*, **366**(6464): 428–429.
De Neve, J. E., C. Imbert, J. Spinnewijn, T. Tsankova and M. Luts (2021), 'How to improve tax compliance? evidence from population-wide experiments in Belgium', *Journal of Political Economy*, **129**(5): 1425–1463.
De Vries, H., V. Bekkers and L. Tummers (2016), 'Innovation in the public sector: a systematic review and future research agenda', *Public Administration*, **94**(1): 146–166.
Doshi, A. R. and O. P. Hauser (2024), 'Generative AI enhances individual creativity but reduces the collective diversity of novel content', *Science Advances*, **10**(28): eadn5290.
Einav, L., A. Finkelstein, S. Mullainathan and Z. Obermeyer (2018), 'Predictive modeling of U.S. health care spending in late life', *Science*, **360**: 1462–1465.
Fellner, G., R. Sausgruber and C. Traxler (2013), 'Testing enforcement strategies in the field: threat, moral appeal and social information', *Journal of the European Economic Association*, **11**(3): 634–660.
Felson, R. B., D. W. Osgood, J. Horney and C. Wiernik (2012), 'Having a bad month: general versus specific effects of stress on crime', *Journal of Quantitative Criminology*, **28**(2): 347–363.
Fischbacher, U. and F. Föllmi-Heusi (2013), 'Lies in disguise - an experimental study on cheating', *Journal of the European Economic Association*, **11**(3): 525–547.
Ghose, A., B. Li and S. Liu (2019), 'Mobile targeting using customer trajectory patterns', *Management Science*, **65**(11): 5027–5049.
Gino, F., S. Ayal and D. Ariely (2009), 'Contagion and differentiation in unethical behavior', *Psychological Science*, **20**(3): 393–398.
Gomila, R. (2020), 'Logistic or linear? Estimating causal effects of experimental treatments on binary outcomes using regression analysis', *Journal of Experimental Psychology: General*, **150**(4): 700.
Grimmelikhuijsen, S., S. Jilke, A. L. Olsen and L. Tummers (2017), 'Behavioral public administration: combining insights from public administration and psychology', *Public Administration Review*, **77**(1): 45–56.
Gustafson, K. S. (2011), *Cheating Welfare: Public Assistance and the Criminalization of Poverty*, New York, NY, USA: NYU Press.
Hallsworth, M., J. A. List, R. D. Metcalfe and I. Vlaev (2017), 'The behavioralist as tax collector: using natural field experiments to enhance tax compliance', *Journal of Public Economics*, **148**: 14–31.
Harrison, G. W. and J. A. List (2004), 'Field Experiments', *Journal of Economic Literature*, **42**(4): 1009–1055.
Hauser, J. R., G. L. Urban, G. Liberali and M. Braun (2009), 'Website morphing', *Marketing Science*, **28**(2): 202–223.
Hauser, O. P., F. Gino and M. I. Norton (2018), 'Budging beliefs, nudging behaviour', *Mind and Society*, **17**(1–2): 15–26.
Hauser, O. P., E. Linos and T. Rogers (2017), 'Innovation with field experiments: studying organizational behaviors in actual organizations', *Research in Organizational Behavior*, **37**: 185–198.
Hofman, J. M. et al. (2021), 'Integrating explanation and prediction in computational social science', *Nature*:595(7866): 181–188.
Horvitz, E. and D. Mulligan (2015), 'Data, privacy, and the greater good', *Science*, **349**(6245): 253–255.
Hummel, D. and A. Maedche (2019), 'How effective is nudging? A quantitative review on the effect sizes and limits of empirical nudging studies', *Journal of Behavioral and Experimental Economics*, **80**: 47–58.
Jachimowicz, J. M., O. P. Hauser, J. D. O'Brien, E. Sherman and A. D. Galinsky (2018), 'The critical role of second-order normative beliefs in predicting energy conservation', *Journal of Quantitative Criminology*, **2**(10): 757–764.

https://doi.org/10.1017/bpp.2024.50 Published online by Cambridge University Press

Jaroszewicz, A., O. Hauser, J. Jachimowicz and J. Jamison (2022). 'How effective is (more) money? Randomizing unconditional cash transfer amounts in the US'. Working Paper. SSRN: 10.2139/ssrn.4154000

John, P. (2018), *How Far to Nudge?: Assessing Behavioural Public Policy*, Edward Elgar Publishing: Cheltenham, UK.

Johnson, E. J. and D. Goldstein (2003), 'Do defaults save lives?', *Science*, **302**(5649): 1338–1339.

Karlan, D., M. McConnell, S. Mullainathan and J. Zinman (2016), 'Getting to the top of mind: how reminders increase saving', *Management Science*, **62**(12): 3393–3411.

Kasper, M. and J. Alm (2022), 'Audits, audit effectiveness, and post-audit tax compliance', *Journal of Economic Behavior & Organization*, **195**: 87–102.

Kessler, J. B., K. L. Milkman and C. Y. Zhang (2019), 'Getting the rich and powerful to give', *Management Science,* **65**: 4049–4062.

Kettle, S., M. Hernandez, M. Sanders, O. Hauser and S. Ruda (2017), 'Failure to CAPTCHA attention: null results from an honesty priming experiment in Guatemala', *Behavioral Sciences*, **7**(4): 28.

Kleven, H. J., M. B. Knudsen, C. T. Kreiner, S. Pedersen and E. Saez (2011), 'Unwilling or unable to cheat? Evidence from a tax audit experiment in Denmark', *Econometrica*, **79**(3): 651–692.

Kluender, R., N. Mahoney, F. Wong and W. Yin (2024), *The Effects of Medical Debt Relief: Evidence from Two Randomized Experiments (No. w32315)*. Cambridge, MA, USA: National Bureau of Economic Research.

Koster, R., J. Balaguer, A. Tacchetti, A. Weinstein, T. Zhu, O. Hauser, D. Williams, L. Campbell-Gillingham, P. Thacker, M. Botvinick and C. Summerfield (2022), 'Human-centred mechanism design with democratic AI', *Nature Human Behaviour*, **6**(10): 1398–1407.

Koster, R., M. Pîslar, A. Tacchetti, J. Balaguer, L. Liu, R. Elie, O. P. Hauser, K. Tuyls, M. Botvinick and C. Summerfield (2024). 'Using deep reinforcement learning to promote sustainable human behaviour on a common pool resource problem'. *arXiv preprint arXiv*:2404.15059.

Larkin, C., M. Sanders, I. Andresen and F. Algate (2019), 'Testing local descriptive norms and salience of enforcement action: a field experiment to increase tax collection', *Journal of Behavioral Public Administration*, **2**.

Lee, A., I. Inceoglu, O. Hauser and M. Greene (2022), 'Determining causal relationships in leadership research using Machine Learning: the powerful synergy of experiments and data science', *The Leadership Quarterly*, **33**(5): 101426.

Matz, S. C., J. J. Gladstone and R. A. Farrokhnia (2023), 'Leveraging psychological fit to encourage saving behavior', *American Psychologist*, **78**: 901–917.

Matz, S. C., M. Kosinski, G. Nave and D. J. Stillwell (2017), 'Psychological targeting as an effective approach to digital mass persuasion', *ProcNatlAcadSciUSA*, **114**(48): 12714–12719.

Matz, S. C., J. D. Teeny, S. S. Vaid, H. Peters, G. M. Harari and M. Cerf (2024), 'The potential of generative AI for personalized persuasion at scale', *Scientific Reports*, **14**(1): 4692.

Milkman, K. L., J. Beshears, J. J. Choi, D. Laibson and B. C. Madrian (2011), 'Using implementation intentions prompts to enhance influenza vaccination rates', *Proceedings of the National Academy of Sciences*, **108**(26): 10415–10420.

Nahum-Shani, I., S. N. Smith, B. J. Spring, L. M. Collins, K. Witkiewitz, A. Tewari and S. A. Murphy (2018), 'Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support', *Annals of Behavioral Medicine*, **52**(6): 446–462.

Ngai, E. W. T., Y. Hu, Y. H. Wong, Y. Chen and X. Sun (2011), 'The application of data mining techniques in financial fraud detection: a classification framework and an academic review of literature', *Decision Support Systems*, **50**(3): 559–569.

Office of Inspector General (2021), *DOL-OIG Oversight of the Unemployment Insurance Program*.

Pe'er, E., S. Egelman, M. Harbach, N. Malkin, A. Mathur and A. Frik (2019), 'Nudge me right: personalizing online nudges to people's decision-making styles', *Computers in Human Behavior,* **109**: 106347.

Regev-Messalem, S. (2013), 'Claiming citizenship: the political dimension of welfare fraud', *Law and Social Inquiry*, **38**(04): 993–1018.

Robinson, C. D., J. Gallus, M. G. Lee and T. Rogers (2019), The Demotivating Effect (And Unintended Message) of Awards, *Organizational Behavior and Human Decision Processes*, **163**: 51–64.

Schultz, P. W., J. M. Nolan, R. B. Cialdini, N. J. Goldstein and V. Griskevicius (2007), 'The constructive, destructive, and reconstructive power of social norms', *Psychology Science*, **18**(5): 429–434.

https://doi.org/10.1017/bpp.2024.50 Published online by Cambridge University Press

Slemrod, J., M. Blumenthal and C. Christian (2001), 'Taxpayer response to an increased probability of audit: evidence from a controlled experiment in Minnesota', *Journal of Public Economics*, **79**(3): 455–483.
Sunstein, C. R. (2015), 'Nudging and choice architecture: ethical considerations', *Yale Journal on Regulation*, **32**.
Sunstein, C. R. (2017), 'Nudges that fail', *Behaviour Public Policy*, **1**(1): 4–25.
Sunstein, C. R. and L. Reisch (2023). 'Do people like algorithms? A Research Strategy'. *Working Paper*. SSRN: 10.2139/ssrn.4544749
Thaler, R. H. and C. R. Sunstein (2009), *Nudge: Improving Decisions about Health, Wealth, and Happiness*, London, UK: Penguin.
Treasury, H. M. (2007), *The Magenta Book: Guidance Notes for Policy Evaluation and Analysis*, London:HM Treasury.
Tummers, L. (2019), 'Public policy and behavior change', *Public Administration Review*, **79**(6): 925–930.
Tummers, L. (2020), 'Behavioral public administration', *Oxford Research Encyclopedia of Politics*.
Wang, Y. and M. Kosinski (2018), 'Deep neural networks are more accurate than humans at detecting sexual orientation from facial images', *Journal of Personality and Social Psychology*, **114**(2): 246–257.
Webb, G. I., M. J. Pazzani and D. Billsus (2001), 'Machine learning for user modeling', *User Modeling and User-Adapted Interaction*, **11**: 19–29.
West, J., M. Bhattacharya and R. Islam (2015) 'Intelligent financial fraud detection practices: an investigation'. J. Tian, J. Jing and M. Srivatsa, eds. International Conference on Security and Privacy in Communication Networks. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. (Springer International Publishing, Cham), 186–203.
Wilkinson, M. (2013), 'Nudges manipulate, except when they don't', *LSE British Politics and Policy Blog,* London, UK.
Xiao, Z., O. Hauser, C. Kirkwood, D. Z. Li, B. Jones and S. Higgins (2024). 'Uncovering individualised treatment effect: evidence from educational trials'. *Scientific Reports*.
Yogama, E. A., D. J. Gray and M. D. Rablen (2024), 'Nudging for prompt tax penalty payment: evidence from a field experiment in Indonesia', *Journal of Economic Behavior & Organization*, **224**: 548–579.

Cite this article: Hauser OP, Greene M and DeCelles K (2025), 'Catch me if you can: Using machine learning and behavioral interventions to reduce unethical behavior', *Behavioural Public Policy*, 1–18. https://doi.org/10.1017/bpp.2024.50