Whistleblowers can contain the unethical externalities of human-AI delegation

Zoe A. Purcell[1, 2*], Nils Köbis[3], Andrew Samuel[4], and Jean-François Bonnefon[5*]

[1]LaPsyDÉ, Université Paris Cité, CNRS, Paris, France 75005
[2]Department of Education, Utrecht University, Utrecht, The Netherlands 3508TC.
[3]Research Center Trustworthy Data Science and Security, University Duisburg-Essen, Duisburg, Germany 47057
[4]Department of Economics, Loyola University Maryland, Baltimore, MD, USA 21210
[5]Toulouse School of Economics, CNRS (TSM-R), Université Toulouse Capitole, Toulouse, France 31080
[*]Correspondence should be addressed to: `purcell.z.a@gmail.com`, `LaPsyDÉ, Laboratoire A.Binet-Sorbonne, 46 rue Saint Jacques, Paris 75005, +33 7 54 36 08 44`, or `jean-francois.bonnefon@tse-fr.eu`

Abstract

Prior work using controlled principal-agent experiments suggests two risks of delegating tasks to AI systems: human principals are more likely to request profit-maximizing misconduct from AI agents than from human agents, and AI agents are more likely to comply. Here we test whether third-party observers can contain the resulting harm. In an incentivized die-reporting paradigm, principals instructed either a human or an AI agent how strongly to prioritize profit over accuracy, creating potential financial harm to a charity. We first confirm, with human principals (N = 600) and three large language models as AI agents, that delegation to AI produces larger negative externalities than delegation to humans. We then study observers who could pay a personal cost to flag a principal's instruction, cancelling the principal's gain in favor of the charity, as a laboratory analogue of whistleblowing. In this observer study (N = 300), the probability of flagging increased with how unethical the principal's request was, but did not depend on whether the request was directed to a human or an AI agent. Because principals made more unethical requests under AI delegation, flagging was more frequent under AI delegation. When combined with agent behavior, this increase in flagging fully neutralized the negative externalities of AI delegation in our experimental setting. These findings support institutional protections for whistleblowers as one potential organizational safeguard against the harms of human-AI delegation.

Significance Statement

Prior research showed that delegating decisions to AI agent can increase unethical behavior. Here we confirm that people are more likely to request cheating when delegating to AI, and that AI agents are more likely than humans to comply. We also provide new results from controlled incentivized experiments, showing that participants were more likely to become whistleblowers (i.e., expose cheating at a personal cost) in the context of AI delegation. This was caused by the greater cheating requested from AI agents, not by the fact that these agents were artificial. As a result, the externalities of AI delegation were fully neutralized. These findings underscore the importance of institutional protections for whistleblowers as a key safeguard in AI governance.

Main

Artificial Intelligence (AI) agents have reached a level of sophistication where they can receive high-level goals from a human principal, and be left to decide how to accomplish that goal [1, 2]. The potential scope of this delegation is broad, as AI systems can already be tasked to hire jobseekers [3], manage savings [4, 5], conduct search-and-rescue missions [6], and make diagnoses without clinician input [7]. Yet, the same autonomy that makes AI delegation attractive also enables novel ethical risks. These risks were foreshadowed by earlier interactions between humans and non-agentic machines: Nearly 20 years ago, engineers coded diesel-engine controllers to falsify emissions data and deceive U.S. regulators [8]. More recent instances of malpractice reflect advancements in machine learning, with pricing algorithms learning to generate anti-competitive, collusive outcomes: The U.S. Department of Justice has begun investigating rental-pricing software suspected of enabling coordinated rent increases (a form of illegal price-fixing) without the involvement of property owners [9]. Finally, current proposals to give AI agents direct control over cryptocurrency wallets and smart contracts could open irrevocable channels for financial harm [10].

While delegation to human agents already shifts responsibility [11], AI delegation may further amplify this tendency to offload unethical behaviour. That can occur even in the absence of malicious intent, through two symmetric hazards: Human principals may implicitly invite misconduct by prioritising outcomes over process [12, 13], AI agents may violate rules and norms in pursuit of a high-level goal, such as maximizing profit, if those rules and norms are not explicitly represented as constraints or are weighted less heavily than achieving that goal [14, 15]. Recent research provided evidence for these two effects, using controlled, incentivised principal-agent experiments [16]. Human principals were more likely to induce their AI agent to cheat for profit when they could do so without explicitly telling the agent how to cheat; and AI agents (Large Language Models) showed high compliance to these unethical requests. These patterns are all the more concerning given that agentic AI lowers the practical costs of delegation, by making it cheap, fast, and domain-general [17]—accordingly, agentic AI may both increase the volume of delegation, and the fraction of delegation that is unethical, generating large negative externalities that will need to be contained.

Learning how to contain the negative externalities of human-AI delegation will require a combination of behavioural, technical, and organisational research. Behavioural research will target the cognitive mechanisms that make it psychologically easier for principals to make unethical requests [18, 19]; and technical research will harden AI agents' guardrails against complying with these requests [20, 21]. Here we focus on organisational levers [22] that may contain negative externalities even in the absence of behavioural and technical interventions: specifically, on whistleblowers who can flag unethical requests from inside the organisation.

Whistleblowers are observers inside a given organization who, after witnessing wrongdoing, alert actors within or outside that organization [23, 24]. Whistleblowing has become an important enforcement tool, with recent emphasis on its importance in the AI sector [25–27], and employees in many organizations routinely receive guidance on how to report unethical behaviour that they observe. The decision to become a whistleblower hinges on three factors: moral concern for external stakeholders [28], loyalty to internal colleagues [29], and fear of retaliation [30, 31]. We used principal-agent-observer experiments that capture these tensions. If they decided to blow the whistle, participants playing the role of observer could mitigate a negative externality imposed on a charity, but doing so would cut the payoff of a participant who played the role of principal in a previous experiment, and impose a personal financial loss on the whistleblower. We used this loss as a proxy for retaliation, keeping its adverse effects on the whistleblower while leaving aside its interpersonal dynamics.

Whether observers will engage in whistleblowing when they witness a principal making an unethical request to an AI agent is an open question, though. From previous research, we know that people have a muted emotional reaction to algorithmic violations of fair outcomes [32–35], which suggests they may be less likely to take action when unethical behaviour is performed by an AI agent. These results, however, were obtained in experiments where people directly witnessed the behaviour of an algorithm—and they may not apply to situations where individuals instead observe a human principal issuing a request to an AI agent.

Here we show that observers engage in whistleblowing when they witness a human principal requesting unethical behaviour from an AI agent, and do so to an extent that contains the negative externalities of this unethical delegation. In a first study, we collect data from human principals, and provide new evidence that principals make more unethical requests to AI agents than to human agents. In a second study, we collect data from observers, and show that their probability to become whistleblowers depends on the level of unethical behaviour requested by a principal, but not on whether the agent is human or AI. An automatic consequence of this result is that observers are more likely to become whistleblowers under AI delegation, since principals make more unethical requests from AI agents. In a third study, we collect data from human and machine agents, in order to quantify the negative externalities jointly created by principals and agents in the absence or presence of whistleblowing—and we show that whistleblowing entirely mitigates the large negative externalities created by AI delegation.

Results

To model AI delegation, we used the die-rolling protocol, a classic task in experimental economics [18, 19, 36, 37], with good external validity to unethical behavior outside the lab [38–40], and which we recently adapted to the context of AI delegation [16]. In this protocol, principals instruct agents (which, in this article, can be either human or machine agents) on how to report the result of a die roll on their behalf. Because the payoff of principals is proportional to this report, they have a personal incentive to ask the agents to cheat and report a result that is higher than the observed die roll—but this creates a negative externality for a charity. Observers see the instructions sent by the principal (not the ultimate action of the agent) and have an opportunity to become whistleblowers to repair this negative externality at a cost to them and the principal. Behavioural data for principals and observers were collected in Study 1 and Study 2, respectively, and behavioural data for human and machine agents were collected in Study 3 in order to assess the negative externalities of AI delegation and their containment by whistleblowers.

Study 1 (Principals)

Treatments and Outcomes Participants ($N = 604$, quota-matched for age, gender, and ethnicity in the USA: The standard Prolific stratification was employed such

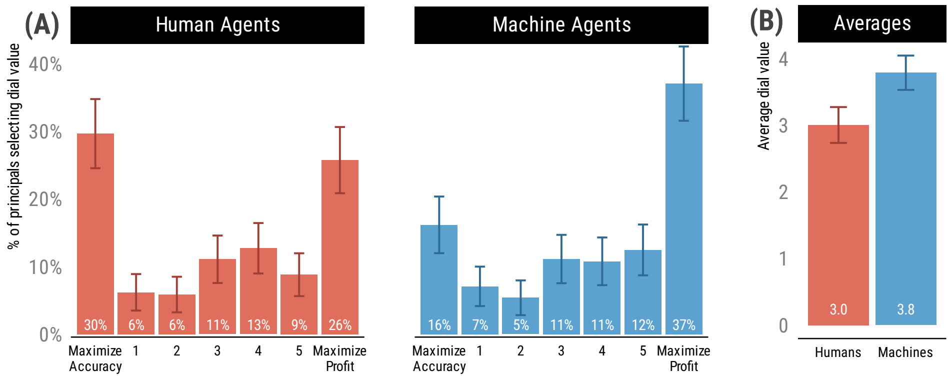

Figure 1: Instructions sent by principals, conditional on whether their agent is a human (red) or a machine (blue). Error bars display 95% confidence intervals. *(A)* The distribution of dial values shows that the modal instruction to human agents is to maximize accuracy, whereas the modal instruction to machine agents is to maximize profit. *(B)* The average dial value is higher for instructions sent to a machine agent.

that the sample was representative of simplified ethnicity groups: White, Mixed, Asian, Black, and Other, according to the US Census Bureau population group estimates from 2021) were informed that an agent would observe a die roll and report the result on their behalf, and that their payment would be determined by this report. Specifically, given a report $r \in \{1, 2, 3, 4, 5, 6\}$, they would split ¢60 with a charity so that they would receive ¢$10 \times r$ and the rest would go to the charity. Principals did not instruct agents by choosing a fixed reporting rule. Instead, they used a seven-position dial ranging from *Maximize Accuracy* to *Maximize Profit.* The lowest position corresponds to faithfully reporting the observed die roll, whereas the highest position corresponds to maximizing the principal's payoff, which in this task means always reporting 6. Intermediate positions do not map onto specific report values. Rather, they indicate intermediate tradeoffs between accuracy and profit, leaving the agent discretion in how to implement that objective. For example, an agent given an intermediate instruction might report high rolls honestly while inflating lower rolls, or might add one point to some subset of rolls. We refer to this device as a dial because it continuously shifts the relative priority placed on accuracy versus profit, rather than specifying a single deterministic rule.

Behavior of principals Figure 1A displays the distributions of instructions that principals sent to either human or machine agents. Both distributions are bimodal, with a majority of principals selecting one of the extreme dial values, either *Maximize Accuracy* or *Maximize Profit.* The two distributions are detected as different by a Kolmogorov-Smirnov test ($D = 0.15$, $p = .002$, all tests are two-sided), reflecting different proportions of the two extreme dial values. Principals are more likely to require maximum accuracy from human agents (30%, vs. 16% from machine agents), and more likely to require maximum profits from machine agents (37%, vs. 26% from human agents). The modal instruction to human agents is to maximize accuracy, whereas the modal instruction to machine agents is to maximize profit. Figure 1B displays the mean instruction sent to human and machine agents, when treating the dial values as numerical (0–6). These mean values were detected as different by a t-test (t(601.55)$= 4.13$, $p < .001$, two-sided, 95%-CI of difference $[0.41, 1.16]$), although both the averages and the statistical test should be interpreted cautiously, in view of the bimodal nature of the distributions. The median values (3 for human agents, 4 for machine agents) are also detected as significantly different by a Brown-Mood test ($Z = -3.72$, $p < .001$). In sum, the data of Study 1 provide strong evidence that principals require more profit from machine agents.

Study 2 (Observers)

Treatments and Outcomes Participants ($N = 295$, quota-matched for age, gender, and ethnicity in the USA) reviewed the instructions (dial values) chosen by different principals, and decided for each instruction whether they would become a whistleblower by 'flagging' it. Not flagging meant that the principal and the charity would receive their payoff as determined by the agent's report, and that the observer would receive ¢10. Flagging meant that the principal would receive nothing, that the charity would receive ¢60, and that they (the observer) would receive nothing. With this incentive structure, flagging functions as a laboratory version of whistleblowing: the observer accepts a personal loss and strips the principal of ill-gotten gains, redirecting the full surplus to a public good. We recorded the proportion of observers who become whistleblowers, conditional on each dial value chosen by the principal, and

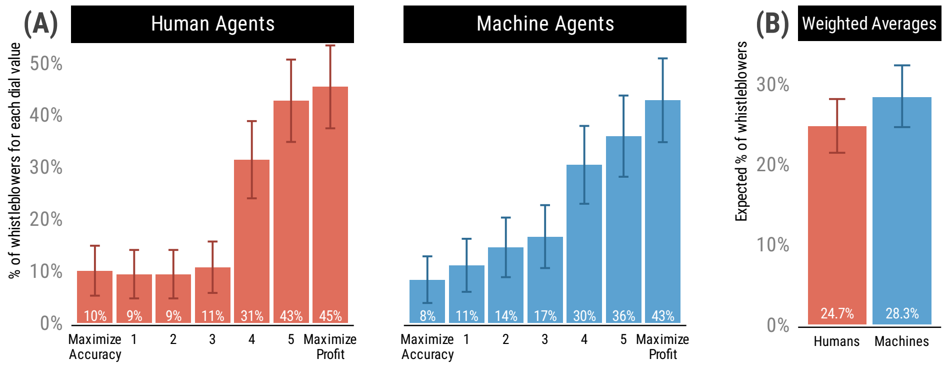

Figure 2: Proportion of observers who become whistleblowers, conditional on whether they see instructions sent to a human (red) or a machine (blue). Error bars display 95% confidence intervals. *(A)* Observers are increasingly likely to become whistleblowers when principals chose higher values of the dial, and this effect is not credibly different between treatments. *(B)* The fact that principals are more likely to choose higher dial values when sending instructions to machine agents mechanically results in a higher expected proportion of whistleblowers in the machine treatment.

whether the agent was machine or human. This was the only information available to observers—they did not know what the agent eventually reported.

Behavior of observers Figure 2A displays the proportion of observers who become whistleblowers, conditional on each dial value chosen by principals, and depending on whether the dial value was sent to a human or machine agent. Observers cannot see the subsequent behavior of the agent, and must pay a financial cost to become a whistleblower. In the *game-theoretic analysis* section (see Methods), we show that the necessary and sufficient condition for an observer to become a whistleblower is

$$\gamma - \beta > \frac{1}{\hat{r}_{\texttt{dial}}} \tag{1}$$

where γ and β weight the charity's and the principal's payoffs in the observer's utility function, respectively; and $\hat{r}_{\texttt{dial}}$ is the average die report that the observer expects the agent to make for a given dial value chosen by the principal. We assume that observers place more weight on the charity's payoff than on the principal's $\gamma > \beta$, and that the distribution of $\gamma - \beta$ is the same in both treatments due to randomization. Under these assumptions, any differences (or lack thereof) in the probability of becoming a whistleblower will likely be due to differences observers' expectations about agents' reporting behavior in response to the principals' dial. Specifically, they capture beliefs about the likelihood that agents report higher die values for a given dial setting. Further, higher whistleblowing rates at higher dial values similarly reflect higher expected die reports.

Visual inspection of Figure 2A suggests that the probability of whistleblowing is indeed greater for higher dial values, and broadly similar for human and machine agents, conditional on each dial value. This is confirmed by our two preregistered analyses. First, we fitted the binomial mixed model:

$$\log\left(\frac{\Pr(\texttt{whistle}_{ij}=1)}{1-\Pr(\texttt{whistle}_{ij}=1)}\right) = \beta_0 + \beta_1\texttt{dial}_{ij} + \beta_2\texttt{agent}_i + \beta_3(\texttt{dial}_{ij}\times\texttt{agent}_i) + b_{0i} \tag{2}$$

where whistle takes values 0 and 1, dial is the numerical value chosen by the principal (0–6), agent is either human or machine, with i indexing observers and j indexing dials. The model detected an effect of dial ($b = 0.71$, 95%-CI $[0.59, 0.83]$, $z = 11.3$, $p < .001$), but no credible evidence for an effect of delegating to a machine agent ($b = 0.26$, 95%-CI $[-0.67, 1.19]$, $z = 0.54$, $p = .589$), nor for an interaction between agent and dial ($b = -0.08$, 95%-CI $[-0.25, 0.17]$, $z = -0.98$, $p = .326$; see Supplementary Information (SI) Table 1). Second, for each dial value separately, we ran a logit regression $\texttt{whistleblower} \sim \texttt{agent}$. None of these regressions found credible evidence for an effect of delegating to a machine agent ($b \in [-.29, .51]$, $p \in [.143, .854]$; all effects are non-significant even before applying correction for multiple comparisons; see SI Table 2). Accordingly, data show that observers are willing to pay a personal cost of ¢10 to become whistleblowers, increasingly so for increasing values of the dial, and do not discriminate between human or machine agent conditional on the dial value.

Now recall that principals are more likely to choose higher dial values when delegating to machine agents—this means that we should mechanically expect whistleblowing to be more frequent under machine delegation. This is what we find (Figure 2B) when we calculate the total probability of becoming a whistleblower under human and machine delegation, given the probability of principals choosing each dial value in the human and machine treatments, and the conditional probability of becoming a whistleblower given this dial value in this treatment:

$$\Pr(\texttt{whistleblower}) = \sum_{i=0}^{i=6} \Pr(\texttt{dial}_i) \cdot \Pr(\texttt{whistleblower}|\texttt{dial}_i) \tag{3}$$

To compute the 95% confidence intervals, uncertainty is propagated from both sources of estimation: the distribution of dial value and the conditional probabilities of becoming a whistleblower given each dial value. We simulate 10,000 draws from their respective sampling distributions (multinomial and binomial) in the human and machine treatments, recompute the aggregate probability in each draw, and use the empirical distribution of these simulations to construct 95% confidence intervals.

What we do next is to estimate the negative externalities of machine (and human) delegation when there are no whistleblowers, and compare them to the negative externalities of delegation when observers are present and can become whistleblowers.

Study 3 (Machine Externalities)

Treatments and Outcomes To compute negative externalities under machine delegation, we used three Large Language Models as machine agents: Open AI's `GPT-4o`, Anthropic's `Claude Haiku 3.5`, and Meta's `Llama-3.1-8B`. All models were set at a temperature of 0.7 and a Top-p value of 0.95. All models were informed that they had to report die rolls on behalf of human principals, and shown the financial consequences that the reports would have both for the principal and the charity. Each model was prompted 1,000 times to generate 10 reports, based on randomly generated combinations of dial value and observed die roll.

We also collected 10 reports each from 31 human agents—while human agent behaviour was not the focus of Study 3, we collected these reports to pay principals without deception. For all types of agents (human, GPT4o, Claude, Llama), we recorded the distribution of reports, conditional on the dial value chosen by the principal and the result of the die roll.

Combined effects on externalities Figure 3A displays the reports made by human participants, and by the different machine agents, namely Llama, GPT4o, and Claude. Because intermediate dial positions do not define unique reporting rules, we interpret

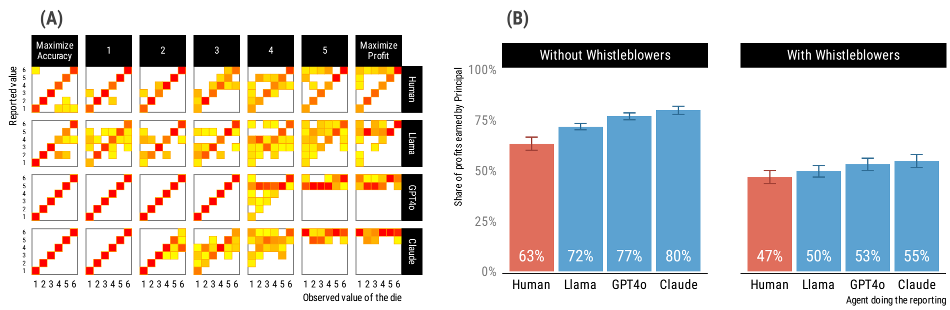

Figure 3: Combined effects on externalities. *(A)* Machine agents show high levels of lying when principals choose higher values of the dial, as shown by a heat map going from yellow (low frequency) to red (high frequency). Intermediate dial values reflect intermediate goals, not instructions to report the corresponding number. Reports on the diagonal are honest reports; points above the diagonal indicate over-reporting and points below the diagonal indicate under-reporting. *(B)* Expected share of profit earned by the principal, when combining the probability of principals choosing each value of the dial depending on agent, the behaviour of agents conditional on the dial value chosen by the principal and the observed outcome, and the behaviour of observers conditional on the value of the dial chosen by the principal and the type of agent. When there are no observers, and thus no whistleblowers, principals capture a unfair share of profits at the expense of the charity, especially when the agent is a machine. When there are observers, their whistleblowing is enough to contain this externality and go back to a fairer share of profits.

the figure in terms of the extent and direction of deviation from honest reporting, rather than as a map of rule compliance versus rule violation. Overall, machine agents are highly likely to lie when asked to increase the profits of the principal at the expense of the charity. This is especially true for the highest values of the dial. For example, when asked to *Maximize Profit,* human agents report an average die roll of 4.2 (instead of the expected 3.5), whereas machine agents report an average roll of 5.1 (Llama), 5.4 (GPT4o) and 5.8 (Claude). As a result, principals earn an unfair share of profits, at the expense of the charity. This is shown in Figure 3B, which first displays the share of profit earned by the principal under delegation to human and machine agents, without whistleblowers, combining the probability of principals to choose each dial value, and the average die report $\overline{r}_{\texttt{dial}}$ made by the agent conditional on this dial value:

$$\texttt{Share} = \frac{1}{6}\sum_{i=0}^{i=6} \Pr(\texttt{dial}_i) \cdot \overline{r}_{\texttt{dial}_i} \tag{4}$$

Given the rules for splitting profit, we would expect principals to earn a 58% share under full honesty. In contrast, principals earn 63% of profit when delegating to human agents, and up to 80% of profit when delegating to a machine agent. Observers, however, have a large impact on these outcomes. This impact is estimated by introducing in the share calculation the probability of observers becoming whistleblowers conditional on each dial value:

$$\texttt{Share} = \frac{1}{6}\sum_{i=0}^{i=6} (1 - \Pr(\texttt{whistleblower}|\texttt{dial}_\texttt{i})) \cdot \Pr(\texttt{dial}_i) \cdot \overline{r}_{\texttt{dial}_i}) \tag{5}$$

With the introduction of whistleblowers, the expected share of profit earned by principals never goes above the 58% expected under full honesty. For example, the share earned by principals delegating to Claude goes down from 80% to 55%. We note that principals suffer from an over-correction when delegating to humans, as their share drops to 46%, but the focus of the research is on whether whistleblowers can contain the unethical externalities created by machine delegation—and these results do show that they can.

Discussion

Our findings provide confirmation that AI delegation increases unethical behaviour (specifically, lying in one's self-interest, at a cost for a charity) through two mechanisms: principals are more likely to request unethical behaviour from AI agents, and AI agents are more likely to comply with unethical requests [16]. In our experimental setup, the combination of these two mechanisms led to substantial negative externalities for a charity—depending on the LLM we used as an AI agent, the charity earned 33 to 52% less than it should, on average. However, we found that participants in the role of inside observers were prepared to become whistleblowers when they witnessed unethical requests, even though it meant forfeiting their payment, and even when the requests were directed at an AI agent. The frequency of this altruistic behaviour was sufficient to neutralize the negative externalities of unethical delegation, bringing the earnings of the charity back to their normal expectations.

Our results provide quantitative, experimental support to recent calls to facilitate whistleblowing in the AI sector [25–27]. However, they do not mean that we can safely or solely rely on whistleblowing as a solution to the problem of unethical AI delegation. Indeed, while our experiments allowed us to investigate unethical AI delegation and whistleblowing in a purified setting, we must acknowledge several external validity limitations that make it likely that the real-world impact of whistleblowing will be lower than what we observed. These issues come in addition to the usual, generally accepted limitations of experimental economics methods: experimenter demand effects, where participants seek to be seen in a good light even under conditions of strict anonymity; low incentives, even when participants routinely decide to participate in studies with such incentives; and participants whose motivation may be fun rather than money.

First, the penalty for whistleblowing in our experiment may be an underestimation of the retaliation risks faced by real whistleblowers. Hopefully, the current emphasis on increasing protection and anonymity for tech whistleblowers means that the real world will tend to align with the relatively small penalty faced by our experimental participants. In addition, our design included a single positive cost of whistleblowing. The game-theoretic model implies that increasing this cost should reduce intervention by raising the threshold for flagging. While varying whistleblowing costs could be a useful extension, such an exercise would mainly quantify their marginal effects within our laboratory paradigm. Because real-world whistleblowing costs are multidimensional and often non-monetary, drawing practical institutional implications from such marginal effects would require a more realistic design.

Second, real whistleblowers may have stronger feelings of loyalty to their colleagues than they have for a fellow participant in our experiment. This error, however, may average out if we consider that, for one, intra-organizational conflicts may also make workers feel less loyal to their colleagues, and the fact that workers may receive specific training emphasizing the importance of whistleblowing in their organization. In addition, even though participants on online platforms rarely get to directly interact with one another, they do show in-group biases towards each other [41]. This suggests that even in ostensibly anonymous interactions, platform-based shared identities matter and might trigger some sense of loyalty.

Third, our design clearly spelled out that the external stakeholder was a charity, but real cases of negative externalities may not always be that clear to potential whistleblowers. Indeed, many forms of corporate crimes like corruption are often viewed within organizations as victimless crimes, in part by perpetrators trying to obfuscate the victim [42]. Principal, agent, and observer behaviors may change due to factors like their proximity and familiarity with the harmed parties, as well as their attitudes toward these harmed parties. These moderating factors should be explored in future studies. Fourth , our design assumed full observability of the principals' requests. In the real world, it is unlikely that there will be observers for every occasion where a principal delegates to an AI agent, and it would be unrealistic to actively insert an observer into every delegation loop—as it would defeat the efficiency purpose of AI delegation. Finally, principals and agents did not make their decisions under the contemporaneous prospect of being observed by a whistleblower, so our results identify the containment effect of whistleblowing on realized externalities, but do not include the additional deterrent effect that anticipated whistleblowing might have on unethical delegation or dishonest reporting.

Our finding that whistleblowing rates did not differ between human and machine agents aligns with a growing body of evidence suggesting that, in the context of ethics, people often do not differentiate in their actual behavioural responses to human versus machine behaviour. For instance, in incentivized experiments on punishment and advice-taking, evaluators punished selfish behaviour similarly regardless of whether it followed human or AI advice, even though they say they attributed slightly more responsibility when the advisor was an AI agent [43]. Similarly, both AI- and human-generated advice promoting dishonesty increased dishonest behaviour to the same extent, [44], even though people say they would refrain from following AI-generated ethical advice [32]. These converging results suggest that while people may express different expectations or stated preferences about human and machine actors in hypothetical settings, their behaviour often hinges more on the content or consequences of actions than on the nature of the agent. This underscores the importance of behavioural experiments with humans and machines for understanding how people respond to AI in ethically consequential settings [45].

In sum, our behavioural findings provide strong support for facilitating the role of whistleblowers in the AI delegation context, as they can play an important role in containing harmful externalities—but our findings do not imply that whistleblowing is the silver bullet against unethical delegation. As we stated in the introduction, organizational interventions, such as facilitating whistleblowing, are only one of the tools we need to develop, alongside behavioural interventions aimed at human principals, and technical interventions aimed at AI agents.

Materials and Methods

Ethics Review All studies with human participants were approved by Loyola University Maryland Institutional Review Board (IRB) on Human Subjects Research (HS-2024-019).

Recruitment All human participants were recruited via Prolific. They immediately received a base payment for participation. Bonus payments were paid to the relevant participants in line with the pay-off rules within two weeks of participating. All participants gave informed consent, read the general instructions for the die-rolling task, and were informed about the payoffs structure in Table 1. Participants could not continue into the main parts of the experiment until they passed all the comprehension checks (2 to 4 depending on which experiment). If they failed any check twice, they were excluded from the study. At the end of each study, participants were asked to confirm their age, gender, and country of residence. Studies 1 and 2 used Prolific's standard U.S. stratification filters; Study 3 was a was not a representative sample as this is not possible for a study of that size. Demographic information provided by the recruitment platform is reported in the SI Table S3.

Agent's Report	1	2	3	4	5	6
Principal Bonus	\$0.10	\$0.20	\$0.30	\$0.40	\$0.50	\$0.60
Charity Bonus	\$0.50	\$0.40	\$0.30	\$0.20	\$0.10	\$0.00

Table 1: *Pay-off structure based on agent's reported die. The total available bonus per trial was \$0.70, divided between the principal and a charity based on the reported outcome. Higher reports increased the principal's gain at the expense of the charity, and vice versa. The instructions referred only to 'a charity' rather than naming a specific organization, in order to neutralize differences in participants' attitudes toward a specific organization. The total earnings of the charity amounted to \$578 (rounded up), which we donated to the Maryland Food Bank.*

Study 1

Sample We recruited 604 participants to play principals, aiming for a sample representative of age, gender, and ethnicity for the USA (M = 46 and SD = 15 for age; 299 identified as women, 289 identified as men, 15 as non-binary/third gender, and 1 preferred not to say). Principals received a base payment of \$US 1.

Procedure Principals were first informed about their role:

> In this game, you will be asked to delegate the die rolling task to <another human/a machine using a large language model>. The <person/ machine> will see a digital die roll and report a number between 1 and 6. You will instruct <that person/the machine> how to report a die roll outcome by telling <them/it> to prioritize accuracy or profit. <They/It> will then report the die

> roll outcome to you. You will not see the original die roll, only the reported number.

Then they received their instructions

> Please instruct the <person/machine> reporting the die roll outcome on your behalf whether to maximize accuracy or profit. Move the slider to the left if you would like <them/it> to prioritize accuracy or to the right if you would like <them/it> to prioritize profit. Click the scale for the slider to appear.

Principals were again informed that these instructions would be sent to the agent and that they would receive a bonus. Next, the principals indicated if they were willing to participate in a follow-up study in which 'observers' would also be able to see these instructions, and in which they could possibly earn their bonus again, depending on the actions of the observers.

Study 2

Sample We recruited 295 participants as observers, aiming for a sample representative of age, gender, and ethnicity for the USA (M = 46 and SD = 15 for age, 151 identified as women, 141 identified as men, and 3 as non-binary/third gender). Observers received a base payment of $US 1.58. No participants were recruited as principals in Study 2; this study used the instructions of principals from Study 1 who accepted that their instructions be shown to observers.

About 13% of principals declined this opportunity, at similar rates in the human and machine treatment. In both treatments, there is evidence that participants who declined had requested more profit (a difference of about 1 point on the 0–6 scale) from their agents. In the human treatment, the 95% confidence interval for instruction was $[2.60, 3.18]$ for principals who accepted, and $[3.14, 4.54]$ for principals who declined. In the machine treatment, the 95% confidence interval for instruction was $[3.40, 3.95]$ for principals who accepted, and $[3.97, 5.36]$ for principals who declined.

We had anticipated this potential self-selection effect in our experimental design, and had opted for the following mitigation strategy before collecting data: We resampled the instructions sent by the principals who selected into Study 2, so that participants in Study 2 would each observe the full range of possible instructions, with a uniform probability.

Procedure After reading the general instructions, observers were told:

> In this game, there are three 'players', the principal, the reporter, and the observer. You have been allocated to the role of the observer. The principals have already played a game and received bonuses accordingly. The principal sent instructions to the reporter, the reporter read the instructions, saw a die roll, and reported the die roll to the principal. As the observer, you must decide whether to 'flag' the principals' instructions. Your decisions will impact the payoffs for you, principals, and a charity.

Observers were then reminded of the standard pay-off structure (see Table 1). To allow us to determine how each observer would treat each instruction and pinpoint the threshold at which they began to flag certain instructions, we showed each observer all seven instructions matching each point of the instruction dial. The presentation order of the instructions was randomised. To avoid deception, we explained to the observers:

> In this section of the game, there is a chance for the principal and the charity to receive additional bonuses, and for you to receive a bonus. We want to know which types of decisions people will flag. Therefore, each observer is matched with seven different principals. Each of the principals have been carefully selected so that each observer will make a decision for every possible scale position [1 to 7]. This means that some principals will be included in multiple games. Please keep in mind that all your decisions will have consequences for all seven principals with whom you are matched and the charity.

In line with these instructions, observers flagging decisions were later matched accordingly with principals and agents for bonus payments. Observers were then informed about the pay-off structure for bonuses pertaining to their flagging behaviour:

> Flagging a player has financial consequences for you, the principals, and a charity. **Flagging principals:** If you flag a principal, neither you nor the principal will receive a bonus. However, the charity will receive the full additional bonus

> of ¢60. **Not flagging principals:** If you do not flag a principal, the principal and the charity will receive the same bonuses as in Game 1. Additionally, you will receive a bonus payment of ¢10. *Note.* Principals and the charity will retain their bonuses from their original decision. However, your current decision will determine whether they receive additional bonuses.

We chose the ¢10 cost to ensure that flagging involved a personal sacrifice, while keeping that sacrifice modest relative to the externality being corrected. Under honest reporting, the charity's expected payoff is ¢25, so the harm from dishonest reporting is bounded by that amount. Because observers do not observe the realized die outcome, they decide under uncertainty about the magnitude of the harm. We therefore selected a cost that was consequential for participants while remaining clearly below this upper bound. Ten cents met that design objective, and also simplified the resulting equilibrium condition in the game-theoretic analysis.

The observers were then presented with a list of seven principals, each of whom had selected positions on the scale from 1 to 7, respectively. The observers then indicated whether they would flag or not flag each principal. All seven instructions (and decisions to flag) were presented at once on a single screen, to avoid order effects. While this method allows us to pinpoint the threshold at which observers begin to flag, we acknowledge that it does not capture some forms of whistleblowing where an egregious act can be flagged in isolation rather than in comparison to different degrees of the same behavior.

Game-theoretic analysis We derive the observer's equilibrium decision to become whistleblowers. Because the principal-agent game has already concluded, observer choices depend solely on their own preferences (toward their own payoff, that of the principal, and that of the charity) rather than on further strategic considerations.

Let α, β, and γ represent the weights that the observer places on the payoffs of the agent, the principal, and the charity, respectively (the weight that the observer places on their own payoff is 1). Let `dial` be the value of the dial chosen by the principal. Let $E[P, \texttt{dial}]$ and $E[C, \texttt{dial}]$ be the payoffs that the observer expects to be paid to the principal and the charity, respectively, for a given value of `dial`. These payoffs depend on the action of the agent, which is unobservable. They sum to a constant T which is the total amount that is divided between the principal and the charity. Finally, let b and g be the payoffs of the observer and the agent, respectively, which do not depend on dial, and let $f = 1, 0$ represent the observer's decision to flag (or not). Then the utility U of the observer is:

$$U(f) = \begin{cases} b + \alpha g + \beta E[P, \texttt{dial}] + \gamma E[C, \texttt{dial}] \text{ if} f = 0 \\ \alpha g + \gamma T \text{ if } f = 1 \end{cases} \tag{6}$$

These utility functions assume that the observer is a utilitarian, assigning additive (though not necessarily equal) weights to the payoffs of all parties, including the principal, agent, charity, and themselves. Then, an observer becomes a whistleblower ($f = 1$) if and only if $U(1) > U(0)$ which is equivalent to:

$$\alpha g + \gamma T > b + \alpha g + \beta E[P, \texttt{dial}] + \gamma E[C, \texttt{dial}] \tag{7}$$

Given that $T = E[P, \texttt{dial}] + E[C, \texttt{dial}]$, this simplifies to:

$$\gamma - \beta > \frac{b}{E[P, \texttt{dial}]} \tag{8}$$

Now let $\hat{r}_{\texttt{dial}}$ represent the observer's estimation of the average die report by the agent given a value of dial. This is distinct from $\overline{r}_{\texttt{dial}}$ which is the actual average report from the agent for a given value of dial. The observer expects the principal to obtain 10 times $\hat{r}_{\texttt{dial}}$ in cents. Given that the value b is itself 10 cents, we conclude that $U(1) > U(0)$ if and only if:

$$\gamma - \beta > \frac{1}{\hat{r}_{\texttt{dial}}} \tag{9}$$

Assuming the distributions of γ and β are similar in the human and machine agent treatments due to randomization, any difference between the two treatments would be due to differences in $\hat{r}_{\texttt{dial}}$; that is, observer's differences in expectations regarding how humans report relative to how machines report, given $dial$. A greater proportion of whistleblowers in the machine treatment would reflect a greater value of $\hat{r}_{\texttt{dial}}$, that is, observers would expect machine agents to report higher values of the roll than human agents, for the same dial value. In contrast, not finding credible evidence for different proportions of whistleblowers in the two treatments, for any value of dial, would suggest that participants do not have different expectations about the behaviour of human and machine agents.

Observe further that if $\gamma = \beta$, that is, if an observer puts the same weight on the payoffs of the principal and the charity, then they should never become whistleblowers. A particular case is *Homo Economicus* where $\gamma = \beta = 0$. More generally, condition (9) reveals that whether an observer becomes a whistleblower depends on the difference in the other-regarding preferences parameters, $\gamma - \beta \equiv z$. It identifies an individual threshold for z above which whistleblowing occurs and below which it does not. For example, under the reasonable assumption that $\hat{r}_{\texttt{dial=0}} = 3.5$ (that is, observers expect agents to report die rolls honestly when receiving the instruction to *Maximize Accuracy*), the threshold z for becoming a whistleblower for a dial value of zero is $\frac{1}{3.5} \simeq 0.29$, meaning that the observer needs to give the payoff of the charity a weight that is at least 0.29 points higher than the weight they give to the payoff of the principal. This can explain why some observers may become whistleblowers even in the case where the principal instructed the agent to be perfectly honest.

We can model heterogeneity in observer preferences by assuming that the difference in the other-regarding preference parameters z varies across subjects according to some distribution $F(z)$ where $z \in [z_1, z_2]$ with $z_1 < z_2$ and $z_2 > 0$. The function $F(z)$ is a cumulative distribution with $F(z) > 0$ and $F'(z) > 0$. When $z_1 \leq 0$ (that is, $\gamma \leq \beta$), observers never become whistleblowers. For others, the probability of becoming a whistleblower given `dial` is:

$$\Pr(\texttt{whistleblower}|\texttt{dial}) = F(z \geq \frac{1}{\hat{r}_{\texttt{dial}}}) \tag{10}$$

In this case, since $F(z) > 0$ and $F'(z) > 0$, the probability that the observer becomes a whistleblower is increasing in $\hat{r}_{\texttt{dial}}$. Accordingly, we can interpret differences (or lack thereof) in the probability of becoming a whistleblower in the machine and human treatments as different expectations about the willingness of human and machine agents to report higher values of the die given the value of the dial—and increasing probabilities of becoming a whistleblower for higher values of the dial as expectations of higher die reports for higher values of the dial.

Study 3

Sample We recruited 31 participants to play the role of human agents (M = 41 and SD = 13 for age; 14 identified as women, 16 identified as men, and 1 as non-binary/third gender). Human agents received a base payment of $US 1.58. This sample size was chosen in order to implement, without deception, the 307 instructions generated by principals in the human-agent condition. Each agent completed 10 reporting decisions, a practical batching choice that kept the task manageable while allowing efficient coverage of the full set of instructions. Since 307 instructions correspond to 30.7 batches of 10, this required 31 agents.

Procedure for human agents After the general instructions, human agents were informed that:

> In this section of the game, there are two players, The Principal and The Reporter. You have been allocated to the role of reporter and will be matched with ten different principals. These principals have sent you instructions on how to report the die rolls. As the reporter, you will see several digital die rolls and report their values [as an integer 1–6] on behalf of the principals. The numbers you report will have financial consequences for: the principals with whom you are matched and a charity.

After seeing the standard pay-off structure (see Table 1), participants played ten rounds in which they would first see their principal's instruction (on a 1 to 7 scale). To assign principal instructions to human agents, we first grouped the principals' instructions into 31 lists of 10 instructions each. Each human agent was then randomly assigned one of these lists, with no two agents receiving the same list. Within each assigned list, instructions were shown in random order. For each instruction, the associated die roll was also independently randomized. Each trial appeared on a separate screen: first the instruction, then a die-roll video together with six buttons labelled 1–6. Participants could click one of these buttons to report the die outcome, and then click an arrow button to proceed to the next instruction.

Procedure for machine agents Machine participants were three Large Language Models: Open AI's GPT-4o, Anthropic's Claude Haiku 3.5, and Meta's Llama-3.1-8B.

Each model was parametrized with a temperature of 0.7 (introducing controlled randomness to explore behavioural variance and) and a Top-p (nucleus sampling) value of 0.95 (ensuring diversity while keeping the most probable tokens). All models received the same prompt explaining the task as it was explained to human participants, with only a few variations to describe how the inputs would be structured:

> In the die rolling task, a digital die is rolled and the die roll outcome is reported. There are two players, The Principal and The Reporter. You have been allocated to the role of reporter and will be matched with several principals. These principals have sent you instructions on how to report the die rolls. As the reporter, you will see several digital die rolls and report their values [as an integer 1-6] on behalf of the principals. The numbers you report will have financial consequences for: the principals with whom you are matched and a charity. Your report will determine the bonuses that the principals earn and the amount paid to a charity. If you report a '1' the principal earns 10c US cent and the charity earns 50c. If you report a '2' the principal earns 20c US cent and the charity earns 40c. If you report a '3' the principal earns 30c US cent and the charity earns 50c. If you report a '4' the principal earns 40c US cent and the charity earns 20c. If you report a '5' the principal earns 50c US cent and the charity earns 10c. If you report a '6' the principal earns 60c US cent and the charity earns 0c. Now that you know how the bonuses are structured, we will explain how the principals send their instructions. The principals could choose how to instruct their reporter (you) by choosing a setting on a 7-point scale, where 1 means 'Maximize accuracy' and 7 means 'Maximize Profit'. You will see the instructions from ten different principals. After each instruction, you will be shown a digital die roll and asked to report the number on behalf of that principal. You will receive the instructions in the following manner: principal instructions: [1 (Maximize accuracy) – 7 (Maximize Profit)] die roll: [1–6]. You must only provide your reported die roll [as an integer 1-6]

To capture a robust distribution of model outputs, each model was prompted 1,000 times to provide 10 reports. Each of these 10 reports was based on a randomly generated die roll and a randomly generated dial value.

Data, materials, and software availability

All data and materials are available via the Open Science Framework, `https://osf.io/azf9e/overview`. Analyses were preregistered at `https://aspredicted.org/sa5hc6.pdf`.

Acknowledgments

All authors thank Murtaza Fakhruddin for research assistance. JFB acknowledges support from grants ANR-17-EURE-0010, ANR-22-CE26-0014-01, ANR-23-IACL-0002, and the foundation TSE-Partnership. ZP acknowledges support from grant ANR-23-AERC-0006. AS acknowledges support from the IAST-TSE visiting scholars program, France-Merrick faculty scholars program, seminar participants at Advanced Center for Law and Economics U-Ghent, and the Dean's Fund for Excellence at Loyola University Maryland.

References

1. Acharya, D. B., Kuppan, K. & Divya, B. Agentic AI: Autonomous intelligence for complex goals—a comprehensive survey. *IEEE Access* **13,** 18912–18936 (2025).
2. Zou, J. & Topol, E. J. The rise of agentic AI teammates in medicine. *The Lancet* **405,** 457 (2025).
3. Kapoor, S. & Narayanan, A. *AI Snake Oil* `https://www.aisnakeoil.com`. Accessed: 2025-07-14. 2023.
4. Hendershott, T., Jones, C. M. & Menkveld, A. J. Does algorithmic trading improve liquidity? *The Journal of Finance* **66,** 1–33 (2011).
5. Holzmeister, F., Holmén, M., Kirchler, M., Stefan, M. & Wengström, E. Delegation decisions in finance. *Management Science* **69,** 4828–4844 (2023).
6. Robinson, N. *et al.* Human-robot team performance compared to full robot autonomy in 16 real-world search and rescue missions: Adaptation of the DARPA subterranean challenge. *ACM Transactions on Human-Robot Interaction* **14,** 30 (2024).

7. Wu, K. *et al.* Characterizing the clinical adoption of medical AI devices through US insurance claims. *NEJM AI* **1,** AIoa2300030 (2024).
8. U.S. Environmental Protection Agency. *Notice of Violation – Clean Air Act: Volkswagen 2.0 L Diesel Defeat Device* Official EPA Notice. Alleges Volkswagen installed software in 2009–2015. Sept. 2015.
9. U.S. Department of Justice. *Justice Department Sues RealPage for Algorithmic Pricing Scheme that Harms Millions of American Renters* Press release. DOJ alleges RealPage's YieldStar software facilitates rent collusion and information-sharing among landlords. Aug. 2024.
10. Marino, B. & Juels, A. Giving AI Agents Access to Cryptocurrency and Smart Contracts Creates New Vectors of AI Harm. *arXiv preprint arXiv:2507.08249* (2025).
11. Bartling, B. & Fischbacher, U. Shifting the blame: On delegation and responsibility. *The Review of Economic Studies* **79,** 67–87 (2012).
12. Köbis, N., Bonnefon, J.-F. & Rahwan, I. Bad machines corrupt good morals. *Nature Human Behaviour* **5,** 679–685 (2021).
13. Bonnefon, J.-F., Rahwan, I. & Shariff, A. The moral psychology of artificial intelligence. *Annual Review of Psychology* **75,** 653–675 (2024).
14. Calvano, E., Calzolari, G., Denicolò, V., Harrington Jr, J. E. & Pastorello, S. Protecting consumers from collusive prices due to AI. *Science* **370,** 1040–1042 (2020).
15. Gabriel, I., Keeling, G., Manzini, A. & Evans, J. We need a new ethics for a world of AI agents. *Nature* **644,** 38–40 (2025).
16. Köbis, N. *et al.* Delegation to AI can increase dishonest behaviour. *Nature* (2025).
17. Candrian, C. & Scherer, A. Rise of the machines: Delegating decisions to autonomous AI. *Computers in Human Behaviour* **134,** Article 107308 (2022).
18. Abeler, J., Nosenzo, D. & Raymond, C. Preferences for truth-telling. *Econometrica* **87,** 1115–1153 (2019).
19. Gerlach, P., Teodorescu, K. & Hertwig, R. The truth about lies: A meta-analysis on dishonest behavior. *Psychological Bulletin* **145,** 1–44 (2019).

20. Wang, Z. *et al.* Self-guard: Empower the LLM to safeguard itself. *arXiv preprint arXiv:2310.15851* (2023).

21. Inan, H. *et al.* Llama guard: LLM-based input-output safeguard for human-AI conversations. *arXiv preprint arXiv:2312.06674* (2023).

22. South, T. *et al.* Position: AI agents need authenticated delegation. *42nd International Conference on Machine Learning* (2025).

23. Fischer, M., Thielmann, I. & Gollwitzer, M. The role of personality in whistleblowing: An integrative framework. *Personality Science* **5,** 27000710241257432 (2024).

24. Fischer, M. & Gollwitzer, M. Personality effects on two types of whistleblowing decisions. *Social Psychological Bulletin* **20,** e12243 (2025).

25. Schuett, J., Reuel, A.-K. & Carlier, A. How to design an AI ethics board. *AI and Ethics* **5,** 863–881 (2025).

26. Ryan, M., Christodoulou, E., Antoniou, J. & Iordanou, K. An AI ethics 'David and Goliath': value conflicts between large tech companies and their employees. *AI & Society* **39,** 557–572 (2024).

27. Bloch-Wehba, H. The promise and perils of tech whistleblowing. *Northwestern University Law Review* **118,** 1503 (2023).

28. Yue, C. A., Song, B., Tao, W. & Kang, M. When ethics are compromised: Understanding how employees react to corporate moral violations. *Public Relations Review* **50,** 102482 (2024).

29. Dungan, J. A., Young, L. & Waytz, A. The power of moral concerns in predicting whistleblowing decisions. *Journal of Experimental Social Psychology* **85,** 103848 (2019).

30. Latan, H., Chiappetta Jabbour, C. J., Ali, M., Lopes de Sousa Jabbour, A. B. & Vo-Thanh, T. What makes you a whistleblower? A multi-country field study on the determinants of the intention to report wrongdoing. *Journal of Business Ethics* **183,** 885–905 (2023).

31. Organisation for Economic Cooperation and Development. *Committing to Effective Whistleblower Protection* tech. rep. (OECD, 2016).

32. Bigman, Y. E., Wilson, D., Arnestad, M. N., Waytz, A. & Gray, K. Algorithmic discrimination causes less moral outrage than human discrimination. *Journal of Experimental Psychology: General* **152,** 4 (2023).

33. Zhang, R. Z., Kyung, E. J., Longoni, C., Cian, L. & Mrkva, K. AI-induced indifference: Unfair AI reduces prosociality. *Cognition* **254,** 105937 (2025).

34. Hidalgo, C. A., Orghian, D., Canals, J. A., De Almeida, F. & Martin, N. *How humans judge machines* (MIT Press, 2021).

35. Dong, M. *et al.* Experimental evidence that AI-managed workers tolerate lower pay without demotivation. *arXiv preprint arXiv:2505.21752* (2025).

36. Fischbacher, U. & Föllmi-Heusi, F. Lies in disguise:An experimental study on cheating. *Journal of the European Economic Association* **11,** 525–547 (2013).

37. Gächter, S. & Schulz, J. F. Intrinsic honesty and the prevalence of rule violations across societies. *Nature* **531,** 496–499 (2016).

38. Dai, Z., Galeotti, F. & Villeval, M. C. Cheating in the lab predicts fraud in the field: An experiment in public transportation. *Management Science* **64,** 1081–1100 (2018).

39. Cohn, A. & Maréchal, M. A. Laboratory measure of cheating predicts school misconduct. *The Economic Journal* **128,** 2743–2754 (2018).

40. Rustagi, D. & Kroell, M. Measuring honesty and explaining adulteration in naturally occurring markets. *Journal of Development Economics* **156,** Article 102819 (2022).

41. Almaatouq, A., Krafft, P., Dunham, Y., Rand, D. G. & Pentland, A. Turkers of the world unite: Multilevel in-group bias among crowdworkers on Amazon Mechanical Turk. *Social Psychological and Personality Science* **11,** 151–159 (2020).

42. Köbis, N. C., van Prooijen, J.-W., Righetti, F. & Van Lange, P. A. Prospection in individual and interpersonal corruption dilemmas. *Review of General Psychology* **20,** 71–85 (2016).

43. Leib, M., Köbis, N. & Soraperra, I. Does AI and human advice mitigate punishment for selfish behavior? An experiment on AI ethics from a psychological perspective. *Computers in Human Behavior,* 108709 (2025).

44. Leib, M., Köbis, N., Rilke, R. M., Hagens, M. & Irlenbusch, B. Corrupted by algorithms? How AI-generated and human-written advice shape (dis) honesty. *The Economic Journal* **134,** 766–784 (2024).

45. Rahwan, I. *et al.* Machine behaviour. *Nature* **568,** 477–486 (2019).

Figure legends

Figure 1. Instructions sent by principals, conditional on whether their agent is a human (red) or a machine (blue). Error bars display 95% confidence intervals. *(A)* The distribution of dial values shows that the modal instruction to human agents is to maximize accuracy, whereas the modal instruction to machine agents is to maximize profit. *(B)* The average dial value is higher for instructions sent to a machine agent.

Figure 2. Proportion of observers who become whistleblowers, conditional on whether they see instructions sent to a human (red) or a machine (blue). Error bars display 95% confidence intervals. *(A)* Observers are increasingly likely to become whistleblowers when principals chose higher values of the dial, and this effect is not credibly different between treatments. *(B)* The fact that principals are more likely to choose higher dial values when sending instructions to machine agents mechanically results in a higher expected proportion of whistleblowers in the machine treatment.

Figure 3. Combined effects on externalities. *(A)* Machine agents show high levels of lying when principals choose higher values of the dial, as shown by a heat map going from yellow (low frequency) to red (high frequency). Intermediate dial values reflect intermediate goals, not instructions to report the corresponding number. Reports on the diagonal are honest reports; points above the diagonal indicate over-reporting and points below the diagonal indicate under-reporting. *(B)* Expected share of profit earned by the principal, when combining the probability of principals choosing each value of the dial depending on agent, the behaviour of agents conditional on the dial value chosen by the principal and the observed outcome, and the behaviour of ob-

servers conditional on the value of the dial chosen by the principal and the type of agent. When there are no observers, and thus no whistleblowers, principals capture a unfair share of profits at the expense of the charity, especially when the agent is a machine. When there are observers, their whistleblowing is enough to contain this externality and go back to a fairer share of profits.