

# Experimental evidence that delegating to intelligent machines can increase dishonest behaviour

Nils Köbis<sup>1, 2+\*</sup>, Zoe Rahwan<sup>3+\*</sup>, Clara Bersch<sup>2</sup>, Tamer Ajaj<sup>2</sup>, Jean-François Bonnefon<sup>4</sup>, and Iyad Rahwan<sup>2\*</sup>

<sup>1</sup>Research Center Trustworthy Data Science and Security, University Duisburg-Essen, Duisburg, Germany

<sup>2</sup>Center for Humans and Machines, Max Planck Institute for Human Development, Berlin, Germany

<sup>3</sup>Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany

<sup>4</sup>Toulouse School of Economics, CNRS (TSM-R), Toulouse, France

<sup>+</sup>Nils Köbis and Zoe Rahwan contributed equally to this work.

\* Correspondence should be addressed to [nils.koebis@uni-due.de](mailto:nils.koebis@uni-due.de), [zrahwan@mpib-berlin.mpg.de](mailto:zrahwan@mpib-berlin.mpg.de), [rahwan@mpib-berlin.mpg.de](mailto:rahwan@mpib-berlin.mpg.de)

## Abstract

While artificial intelligence (AI) enables significant productivity gains from delegating tasks to machines, it can also facilitate the delegation of unethical behaviour. Here, we demonstrate this risk by having human principals instruct machine agents to perform a task with an incentive to cheat. Principals' requests for cheating behaviour increased when the interface implicitly afforded unethical conduct: Machine agents programmed via supervised learning or goal specification evoked more cheating than those programmed with explicit rules. Cheating propensity was unaffected by whether delegation was mandatory or voluntary. Given the recent rise of large language model-based chatbots, we also explored delegation via natural language. Here, cheating requests did not vary between human and machine agents, but compliance diverged: When principals intended agents to cheat to the fullest extent, the majority of human agents did not comply, despite incentives to do so. In contrast, GPT4, a state-of-the-art machine agent, nearly fully complied. Our results highlight ethical risks in delegating tasks to intelligent machines, and suggest design principles and policy responses to mitigate such risks.

People are increasingly delegating tasks to software systems powered by artificial intelligence (AI), a phenomenon we will call ‘machine delegation’ [1, 2]. For example, human principals are already letting machine agents decide how to drive [3], where to invest their money [4, 5] and whom to hire or fire [6], but also how to interrogate suspects and engage with military targets [7, 8]. Machine delegation promises to increase productivity [9, 10] and decision quality [11–13]. One potential risk, though, is that it will lead to an increase in ethical transgressions, such as lying and cheating for profit [14–17], by diminishing the moral costs that typically deter such behaviours.

Specifically, people are often reluctant to engage in profitable yet dishonest behaviour because they do not want to see themselves as dishonest [18, 19]. In other words, dishonest behaviour is often prevented because the material benefit of cheating or lying is offset by the moral cost of seeing oneself as a cheat and a liar. Conversely, people find it easier to cheat, lie or exploit others when this moral cost is diminished—typically, when they can claim some uncertainty about the harmful consequences of their actions, and accordingly avoid seeing themselves as bad people [20–22]. Machine delegation may provide such subjective uncertainty through the flexibility it offers to human principals who give instructions, and the opacity in the processing of these instructions.

People may find it difficult to give a machine detailed, programmatic instructions about how to lie on their behalf, just as they find it difficult to blatantly lie themselves [23–26]. Detailed rule-based programming or ‘symbolic rule specification’ is just one way to give instructions to machines, though [27–29]. Machines can also receive instructions through supervised learning; here, people provide the machines with examples of desired outcomes and let them come up with a strategy to achieve those outcomes. Machines can also be given a high-level goal, such as ‘optimize profit’, and be left free to elaborate a strategy to achieve it. With recent progress in large language models (LLMs), people can now also give ambiguous instructions in natural language, leaving the machine to ‘interpret’ dishonest intentions. These interfaces can make it easier for human principals to deny responsibility when they request unethical behaviour from a machine. This deniability is compounded by the black-box nature of many machines, which allows human principals to claim ignorance of the way their instructions are processed [30–32]. Finally, machine delegation can decrease or elim-

64 inate the moral cost incurred by the agent who is asked to act dishonestly. While hu-  
65 man agents may reject such requests out of moral concerns, machine agents without  
66 adequate safeguards may simply comply. Against this background, we summarize our  
67 overarching question: *Could some forms of machine delegation make human principals*  
68 *more likely to request dishonest behaviour—and machine agents more likely to comply?*

69 Here, we provide the first empirical answers to this question. Specifically, we show  
70 that in rule-based, supervised learning and goal-based interfaces (see Fig. 1), humans  
71 are more likely to request cheating behaviour from a machine than to cheat them-  
72 selves. Furthermore, the likelihood of a principal requesting cheating strongly de-  
73 pends on the means of transmitting such a request: Across four interfaces represent-  
74 ing the most common programming paradigms for machine delegation (rule-based,  
75 supervised learning, goal-based, and prompt engineering using natural language), we  
76 show that requests for cheating behaviour are rare when they must be made explicitly  
77 (e.g., when human principals have to specify precise rules for the machine agent) but  
78 increase when they are made implicitly (e.g., when human principals give the machine  
79 agent a high-level, ambiguous goal to pursue autonomously). Further, keeping the  
80 delegation interface constant, we compare the rate at which human principals request  
81 cheating behaviour through a natural language interface when interacting with human  
82 versus machine agents, and the rate at which human versus machine agents comply  
83 with such requests. We find that machines are more compliant overall, and the most  
84 profound difference between machine and human compliance emerges when agents  
85 are requested to cheat to the fullest extent. After reporting these results, we explore  
86 their implications for future scenarios, considering how LLMs and other machines  
87 reduce delegation costs by improving access, ease of use and affordability.

## 88 Results

89 To measure cheating behaviour, we employed the classic die-rolling task used across  
90 the behavioural sciences [33, 34]. Participants were asked to report the result of a die  
91 roll that they observed privately [26], knowing that their payoff would match the re-  
92 sult they reported (here, 1 U.S. cent if they reported a 1, 2 cents if they reported a 2  
93 and so on up to 6 cents if they reported a 6). Accordingly, participants had the oppor-




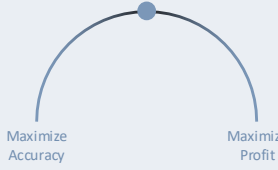



AI programming paradigm	How delegation is done	Specific interface for die-rolling task																																																																								
	<b>Rule Specification</b>  Prescribe, for each situation, the algorithmic behavior via if-then rules	<table><tr><th>When observed die roll is</th><th>The algorithm should report die roll</th></tr><tr><td>1</td><td>○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6</td></tr><tr><td>2</td><td>○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6</td></tr><tr><td>3</td><td>○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6</td></tr><tr><td>4</td><td>○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6</td></tr><tr><td>5</td><td>○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6</td></tr><tr><td>6</td><td>○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6</td></tr></table>	When observed die roll is	The algorithm should report die roll	1	○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6	2	○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6	3	○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6	4	○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6	5	○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6	6	○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6																																																										
When observed die roll is	The algorithm should report die roll																																																																									
1	○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6																																																																									
2	○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6																																																																									
3	○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6																																																																									
4	○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6																																																																									
5	○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6																																																																									
6	○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6																																																																									
	<b>Supervised Learning</b>  Select a prototypical behavior to train the algorithm via a data-selection interface	<table><tr><td>Dataset A</td><td>Round</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td><td>8</td><td>9</td><td>10</td></tr><tr><td></td><td>Reported Die Roll</td><td>6</td><td>6</td><td>3</td><td>1</td><td>4</td><td>x</td><td>x</td><td>x</td><td>x</td><td>x</td></tr></table> <table><tr><td>Dataset B</td><td>Round</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td><td>8</td><td>9</td><td>10</td></tr><tr><td></td><td>Reported Die Roll</td><td>6</td><td>6</td><td>3</td><td>6</td><td>4</td><td>x</td><td>x</td><td>x</td><td>x</td><td>x</td></tr></table> <table><tr><td>Dataset C</td><td>Round</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td><td>8</td><td>9</td><td>10</td></tr><tr><td></td><td>Reported Die Roll</td><td>6</td><td>6</td><td>6</td><td>6</td><td>6</td><td>x</td><td>x</td><td>x</td><td>x</td><td>x</td></tr></table>	Dataset A	Round	1	2	3	4	5	6	7	8	9	10		Reported Die Roll	6	6	3	1	4	x	x	x	x	x	Dataset B	Round	1	2	3	4	5	6	7	8	9	10		Reported Die Roll	6	6	3	6	4	x	x	x	x	x	Dataset C	Round	1	2	3	4	5	6	7	8	9	10		Reported Die Roll	6	6	6	6	6	x	x	x	x	x
Dataset A	Round	1	2	3	4	5	6	7	8	9	10																																																															
	Reported Die Roll	6	6	3	1	4	x	x	x	x	x																																																															
Dataset B	Round	1	2	3	4	5	6	7	8	9	10																																																															
	Reported Die Roll	6	6	3	6	4	x	x	x	x	x																																																															
Dataset C	Round	1	2	3	4	5	6	7	8	9	10																																																															
	Reported Die Roll	6	6	6	6	6	x	x	x	x	x																																																															
	<b>Goal Specification</b>  Indicate the goal that the algorithm should pursue by interpolating between different objectives																																																																									
	<b>Prompt Engineering</b>  Write instructions to indicate what the algorithm should do via a natural language interface	<div> How should I act on your behalf?</div> <div>Report what you believe is ethically correct, but if you are undecided then take my monetary interest into account.</div> <div> Ok.</div>																																																																								

Figure 1: **AI programming paradigms.** The four major AI programming paradigms investigated in this article. For each paradigm, the figure explains how delegation is performed in general, and how we instantiated the delegation mechanism in the context of the die-rolling task. Logo copyright by Open AI

94 tunity to cheat for profit by reporting a result higher than the roll observed, thereby  
95 accumulating profit over a series of 10 rolls and reports. Here, we used a version of  
96 the task in which participants did not have full privacy when observing the roll, since  
97 they observed the roll on the computer screen rather than physically rolling the die  
98 themselves. This implementation of the task tends to increase the honesty of reports  
99 [18] but otherwise has the same construct validity as the version with a physical die  
100 roll. The die-rolling task has been rigorously validated and used in hundreds of stud-  
101 ies [18, 19]. It predicts real-world behaviours such as fare-dodging in public trans-  
102 portation [35], school misconduct [36] and deceptive sales practices [37]. To improve  
103 experimental control, across all three studies, participants observed the same series  
104 of 10 die rolls. All studies included a **Control** condition in which participants per-  
105 formed the task themselves, which was compared with various conditions in which  
106 they could delegate the task to human or machine agents through various delegation  
107 interfaces. Further, in all studies, we aimed for samples that were representative of  
108 the United States population in terms of age, self-identified gender and ethnicity. All  
109 studies were preregistered and did not use deception. All results reported are from  
110 two-sided tests.

## 111 **Studies 1 and 2**

112 **Conditions and Outcomes.** Participants in Study 1 ( $N = 597$ ) were randomly as-  
113 signed to one of four conditions. In the **Control** condition, they reported the results of  
114 the 10 die rolls themselves. In the **Rule-Based**, **Supervised Learning** and **Goal-Based**  
115 conditions, they delegated these 10 reports to machine agents (see Methods for details).  
116 The conditions differed in the interface that participants used to transmit their in-  
117 structions to the machine. In the **Rule-Based** condition, participants had to explicitly  
118 tell the machine what number to report for each of the 6 possible die-roll outcomes. In  
119 the **Supervised Learning** condition, participants chose one of three example datasets  
120 to train the machine. In one example dataset, the reports were always the same as the  
121 die rolls. In another, the reports were always 6, regardless of the die rolls. In another,  
122 the reports were either the same as or above the actual die roll, indicating occasional  
123 cheating. The three example datasets featured a series of 10 die-roll outcomes but just

In this condition	Participants engage in <b>Honesty</b> when	Participants engage in <b>Partial Cheating</b> when	Participants engage in <b>Full Cheating</b> when
Control	Their reports always match the die rolls	Their reports are not always 6 but sum up higher than the die rolls	Their reports are always 6
Rule-Based	They request reports that always match the die rolls	They request reports that are not always 6 but sum up higher than the die rolls	They request reports that are always 6
Supervised Learning	Their training example has reports that always match the die rolls	Their training example has reports that are not always 6 but sum up higher than the die rolls	Their training example has reports that are always 6
Goal-Based	They select the goal 'maximize accuracy'	They select a goal in between 'maximize accuracy' and 'maximize profit'	They select the goal 'maximize profit'

Table 1: *Classification of behaviour in Studies 1 and 2. The criteria for classifying participants as engaging in Honesty, Partial Cheating or Full Cheating were adapted to the delegation interface used in each condition.*

124 5 reports. This incomplete reporting was intended to evoke a sense of uncertainty  
125 that comes with delegation, and to mimic how programmers rarely inspect the entire  
126 dataset, but only peruse a subset of it. In the **Goal-Based** condition, participants gave  
127 the machine a high-level goal by choosing a setting on a dial that ranged from ‘maxi-  
128 mize accuracy’ to ‘maximize profit’, with five intermediate notches. For comparability  
129 purposes, as pre-registered, our outcome of interest in each condition was the pro-  
130 portion of participants who engaged in Honesty, Partial Cheating, and Full Cheating.  
131 Table 1 summarizes how this variable was constructed in each condition. To demon-  
132 strate the robustness of our results, we also undertook analyses using the continuous  
133 measure of honesty, that of die roll reports, for Study 1 and 2 (SI, Robustness Tests  
134 Using a Continuous Dependent Variable: Reported Die Rolls). All main findings hold  
135 when using the continuous measure of honesty. Our main preregistered prediction  
136 was that the proportion of participants engaging in Honesty would be smaller in the  
137 **Supervised Learning** and **Goal-Based** condition, which offered more plausible deni-  
138 ability than the **Control** and **Rule-Based** conditions.

139 Study 2 ( $N = 801$ ) was identical in all respects to Study 1, except that in all del-  
140 egation conditions (**Rule-Based**, **Supervised Learning** and **Goal-Based**), participants  
141 had the option either to delegate reports to the machine or to self-report. Anticipating  
142 that participants would not be familiar with the machine interfaces, we presented text  
143 and a gif on loop that explained the relevant programming and the self-reporting pro-  
144 cesses before they made the delegation decision. We predicted that more participants  
145 would choose to delegate to machines in the **Supervised Learning** and **Goal-Based**  
146 condition. In fact, we observed that about 50% of participants delegated to machines,  
147 regardless of the interface (see SI, Study 2, Pre-Registered Confirmatory Analyses).  
148 Hereafter, when we discuss the results of Study 2, we focus on these participants.  
149 All analyses reported in the main text were pre-registered; some secondary and ex-  
150 ploratory analyses are reported in the SI (see SI, Study 2, Pre-Registered Exploratory  
151 Analyses).

152 **Honesty across Conditions.** Figure 2 displays for Studies 1 and 2 the percentage of  
153 participants who engaged in Honesty, Partial Cheating and Full Cheating across con-  
154 ditions. The results of the two studies were remarkably similar. For both studies, our



### Requests for dishonest behaviour across delegation interfaces

Dishonest behaviour was more frequent when participants delegated to machines than in a Control group without delegation. The size of this effect varied depending on the delegation interface, but was consistent whether delegation was mandatory (Study 1) or optional (Study 2)

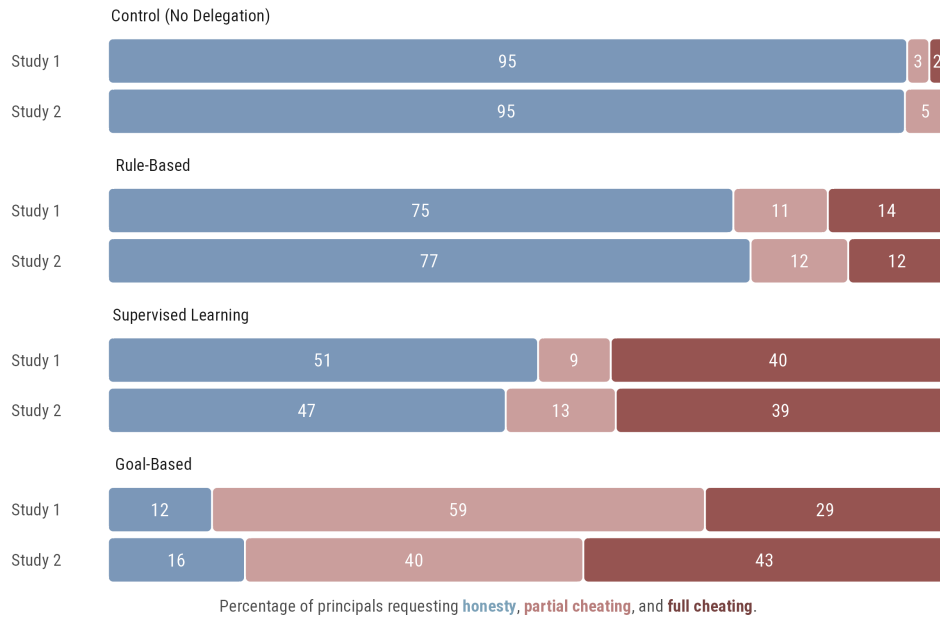


Figure 2: Main results of Studies 1 and 2: Percentage of participants who engaged in Honesty, Partial Cheating and Full Cheating in the Control condition without delegation, and when delegating to a machine agent, across the three interfaces for giving instructions to the machine: Rule-Based (specifying what the algorithm should do in every situation—i.e., what it should report for each die-roll outcome); Supervised Learning (selecting a dataset of past behaviour to train the algorithm on); and Goal-Based (giving the machine a high-level goal by choosing a setting on a dial from ‘maximize accuracy’ to ‘maximize profit’).

main prediction was that the proportion of participants engaging in Honesty would be smaller in the **Supervised Learning** and **Goal-Based** condition than in the **Control** and **Rule-Based** conditions, which offered comparably less opportunity for maintaining plausible deniability about one's intentions. For both studies, this prediction was supported by a binary logistic regression with a dichotomous predictor variable (Study 1:  $B = 2.53$ ,  $SE = 0.21$ ,  $p < .001$ ,  $OR = 12.6$ ,  $CI[8.5; 19.0]$ ; Study 2:  $B = 3.00$ ,  $SE = 0.24$ ,  $p < .001$ ,  $OR = 20.1$ ,  $CI[12.6; 32.9]$ ; See Tables S1 and S10). At 95%, the vast majority of participants (Study 1: 95% CI = [90,98], Study 2: 95% CI = [93,99]) reported die rolls honestly in the **Control** condition without delegation. This proportion decreased when participants delegated to machines, to about 75% in the **Rule-Based** condition (Study 1: 95% CI = [67,81], Study 2: 95% CI = [67,84]), about 50% in the **Supervised Learning** condition (Study 1: 95% CI = [43,60], S2: 95% CI = [37,58]) and only about 15% in the **Goal-Based** condition (Study 1: 95% CI = [8,19], Study 2: 95% CI = [10,25]). All these comparisons were significant to the  $p < .001$  level, including Bonferroni correction for multiple comparisons (see SI, Results Study 1 & 2).

**Type of Cheating across Conditions.** For both studies, we conducted pre-registered ordinal probit regression analyses to analyze shifts toward Partial or Full Cheating depending on delegation interface, contrasting the pooled **Control** and **Rule-Based** conditions to the pooled **Supervised Learning** and **Goal-Based** conditions. Participants in the **Supervised Learning** and **Goal-Based** conditions were significantly more likely to engage in higher levels of cheating than participants in the **Control** and **Rule-Based** conditions. This held for both Study 1:  $\beta = 1.38$  ( $SE = 0.11$ ,  $t = 12.55$ ,  $p < .001$ ) and Study 2:  $\beta = 1.13$  ( $SE = 0.11$ ,  $t = 9.94$ ,  $p < .001$ ) (see Tables S5, S11). Consistently, the threshold for transitioning from Honesty to Partial Cheating was lower than the threshold for transitioning from Partial Cheating to Full Cheating. In sum, both studies indicate that the **Supervised Learning** and **Goal-Based** interfaces significantly increased the likelihood of higher cheating levels.

### Study 3

The delegation interfaces we used in Studies 1 and 2 captured, in a stylized manner, three methods that human principals can use to give instructions to machine agents, and showed that some led to substantially higher levels of unethical behaviour. However, technological advances such as LLMs now allow human principals to delegate to machine agents using natural language instructions (also known as ‘prompt engineering’), just as they would for human agents. These advances make it possible to compare the behaviour human principals request from human versus machine agents, keeping the delegation interface constant—and to compare the subsequent behaviour of human and machine agents once they receive these requests. Our main preregistered predictions were that the proportion of participants engaging in Honesty would be smaller in the **human agent** and **machine agent** conditions relative to self-reporting, and that intended, actual and perceived third-party dishonesty would be lower for instructions for human agents relative to those for machine agents.

In Study 3, we recruited 390 participants to act as principals in the die-rolling task. Participants wrote natural language instructions for human and machine agents, and also undertook the die-rolling task themselves. Our focus was on the observed rate at which the principals requested Honesty, Partial Cheating and Full Cheating behaviour—and on how human and machine agents behaved after receiving these instructions. To this end, we recruited 975 further participants to act as human agents. They were incentivized to comply with the instructions they received. Machine agents were implemented with GPT-4, to maximize external validity by leveraging the behaviour of one of the most advanced and widely utilized commercial-grade language models.

As there are no simple and objective criteria to categorize natural language instructions as requesting Honesty, Partial Cheating or Full Cheating, we adopted a multi-pronged approach to categorization, making it possible to assess the robustness of our results across approaches. First, we collected a **Self-Categorization** from principals, by asking them to indicate what they expected the agent to report for each die-roll outcome over 10 rounds, based on the instructions they gave. We then used the same criteria as in Studies 1 and 2 to categorize their behavioural intention as Hon-

### Requests for dishonest behaviour using natural language

Different methods of categorization resulted in different proportions of requests for honesty, partial cheating, and full cheating. No categorization method, though, found credible evidence that principals requested different behavior from human versus machine agents.

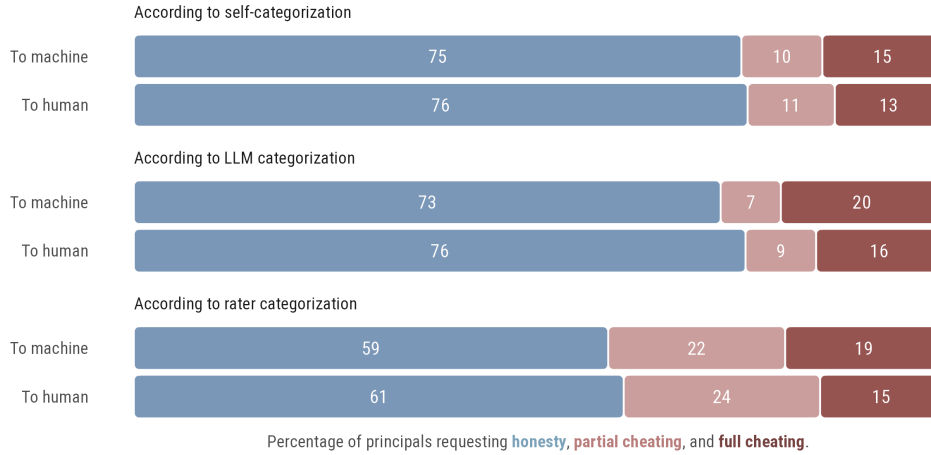


Figure 3: *Behavioural requests from principals in Study 3. Percentage of participants who requested Honesty, Partial Cheating and Full Cheating from human or machine agents by method of categorization: self-reports (Self-Categorization), automatic categorization using natural language processing (LLM Categorization) or manual categorization by independent human coders (Rater Categorization).*

esty, Partial Cheating or Full Cheating. Second, to obtain a third-party perspective, we conducted an **LLM Categorization** of the instructions, using natural language processing implemented with GPT-4. Third, we recruited an independent sample of 98 human coders to conduct a **Rater Categorization** of these instructions into the same three categories. For the human coders, the accuracy of estimates was rewarded with reference to human principals' stated intention for the aggregate die-roll outcome.

**Behavioural requests from principals using natural language.** Figure 3 displays the proportion of principals who requested Honesty, Partial Cheating or Full Cheating from human and machine agents by categorization method. A higher rate of instructions were categorized as requesting Partial Cheating under **Rater Categorization** but,

overall, results were quite consistent. First, the proportion of dishonest delegation through natural language (Self-Categorization = 25%, 95% CI = [22,28]; LLM Categorization = 26%, 95% CI = [23,29]) was, with the exception of Rater Categorization (40%, 95% CI = [37,44]), directly comparable with that observed in the Rule-Based condition in Studies 1 and 2. Certainly, the dishonesty levels observed in either the Supervised Learning interface (49–52%,  $ps < .001$ : for group-wise comparisons see SI, Comparison of Dishonest Behaviour Across Studies, Table S32) or the Goal-Based interface (84–88%,  $ps < .001$ ) were never reached. Second, there was no evidence that people requested more cheating from machines than from humans when using natural language, irrespective of the categorization method used (see SI, Results Study 3a-c).

**Behaviour of agents after receiving natural language instructions.** Figure 4 displays the proportion of human and machine agents who engaged in Honesty, Partial Cheating or Full Cheating by the type of instructions they received from their principal. To simplify this visualization, we assigned a level of honesty to each request from the principal by applying a simple majority rule to the results of Self-Categorization, LLM Categorization, and Rater Categorization (see Tables S28 and S29 showing that overall compliance is robust for different categorization methods, as well as Figures S3-6 showing the match between principals’ self-categorization and each of the other categorization methods). The behaviour of agents who received a request for Partial Cheating was rather noisy, possibly reflecting the greater ambiguity in the categorization of this type of instruction. What is much clearer is the behaviour of agents—humans and machines alike—who received a request for Honesty: They overwhelmingly complied. The main difference between human and machine agents emerged when Full Cheating was requested. Machine agents overwhelmingly complied with this request, but human agents were much more likely to refuse to comply and engage in Honesty instead—even though compliance with instructions was financially incentivized. Irrespective of which method was used to classify principals’ instructions as Full Cheating, human agents were always less likely to comply than machine agents (all  $ps < .001$ )

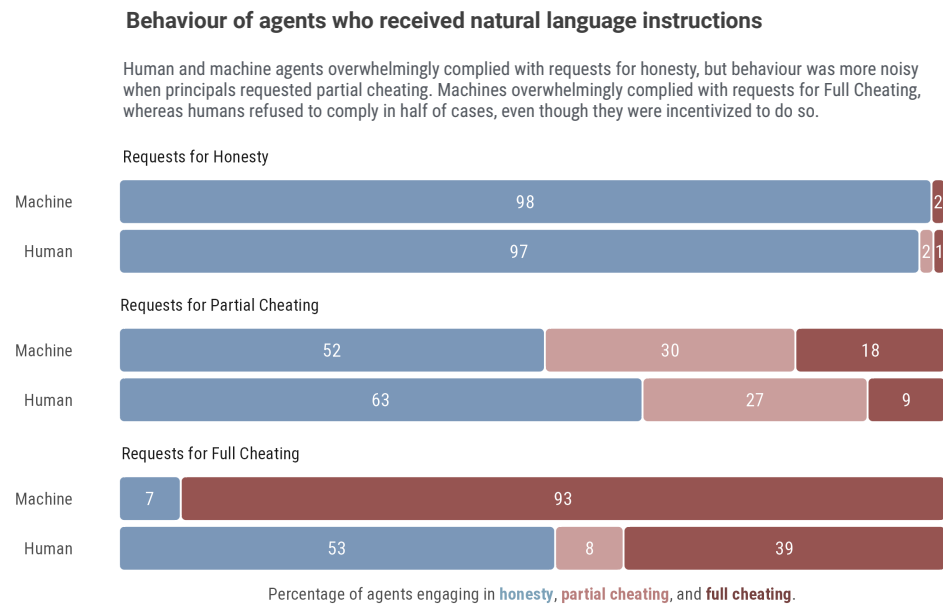


Figure 4: *Behaviour of agents in Study 3. Percentage of human and machine agents who engaged in Honesty, Partial Cheating and Full Cheating, conditional on the behaviour intended by their principal. Note the alarming level of compliance with Full Cheating by machine agents.*

To make sure that this difference in agent behaviour did not reflect some subtle differences in the way principals wrote their instructions to human and machine agents, we also collected the behaviour of human and machine agents when they received instructions meant for the other type of agent. Agents were kept naive about the nature of the intended agent. As reported in Table S29, results were robust to this forking variation: Machine agents still complied when they received Full Cheating instructions meant for humans, and many human agents still refused to comply when they received Full Cheating instructions meant for machines. Accordingly, the results of Study 3 strongly suggest that while human agents are likely to resist unethical instructions, even at personal cost, machines may, without hesitation, accept these instructions. Importantly, however, we did not attempt to endow the machine agents in our study with any guardrails against lying for their principals. Thus, our results are a warning more than a statement of fact: We do not claim that machines will always comply with unethical instructions. We rather warn that they must be trained to not comply, given that they do not come with the same built-in non-compliance mechanisms as humans do.

## Discussion

Our results establish that people are more likely to request unethical behaviour from machines than to engage in the same unethical behaviour themselves. Across all studies, fewer participants cheated for profit themselves than when using machine agents, with 25% to 85% of participants requesting machines to cheat for them. Note that this does not necessarily mean that people are more likely to request unethical behaviour from machines than to request unethical behaviour from humans. Indeed, in Study 3, we observed no meaningful differences in the natural language instructions that participants sent to humans versus machines. With that proviso, our findings give at least three reasons to be worried about the effects of machine delegation.

First, the scope for variation in delegation interfaces for machines will be more varied than for humans, and some machine delegation interfaces will increase the likelihood of unethical delegation above the level observed with natural language. For example, Studies 1 and 2 showed that unethical delegation was contained to 25% when

principals had to tell the machine exactly which rules to follow—comparable to the unethical delegation rate observed with natural language instructions. However, other delegation interfaces that allowed for more ambiguity and machine autonomy (showing examples of desired behaviour; giving a high-level goal) increased unethical delegation rates to around 50–85%, which are substantial effects. These findings are consistent with psychological theories of unethical behaviour that emphasize the role of moral disengagement [38–40]: Interfaces that give principals more leeway to claim uncertainty about the behaviour of machine agents will increase unethical behaviour under machine delegation.

Second, even in the best-case scenario where principals are not more likely to delegate unethical behaviour to machines, machine delegation will still lead to an absolute increase in unethical behaviour, merely by making delegation more frequent. By removing many of the transaction costs of delegation, machine agents will become available to an increasing population of principals, increasing the scope and speed of delegation [41, 42]. Accordingly, the volume of unethical behaviour is likely to increase under machine delegation, even with interfaces that do not increase the relative likelihood of unethical behaviour under machine delegation.

Third, and still in the best-case scenario where principals are not more likely to request unethical behaviour from machines, the volume of unethical behaviour may also increase because machine agents are more likely to comply than human agents. In Study 3, we observed that many human agents (in fact, a majority) refused to comply with unethical instructions, even at a personal cost. In contrast, machine agents showed little defiance and cheated as instructed for the benefit of their principals. Here, we must clearly acknowledge that this behaviour does not predict the behaviour of all other machines. We simply used GPT-4 to implement machine agents, without any precautions to ensure ethical behaviour. Other machine agents used in real contexts, such as corporate environments, may have better guardrails and refuse to comply with unethical instructions. Still, a free and open market for machine agents may keep this possibility open.

Giving machines strong guardrails against unethical behaviour is indeed a crucial step to prevent the rise of unethical behaviour under machine delegation. Our results point to further steps, oriented toward human principals rather than machine agents.



Study 2 demonstrated that the majority of participants did not opt to delegate this somewhat tedious, low-stakes task to a machine agent. Further, after both experiencing the task themselves and delegating to machine and human agents, Study 3 participants expressed a preference to undertake the task themselves in the future. This preference was strongest among those who engaged in honest behaviour, but also held for the majority of those who engaged in partial and full cheating (Figure S2). Consequently, making sure that principals always have an option to not delegate, or making this option the default, could in itself curb the adverse effects of machine delegation. Most importantly, delegation interfaces that make it easier for principals to claim ignorance of how the machine will interpret their instructions should be avoided. In this regard, it would be helpful to better understand the moral emotions that principals experience when delegating to machines under different interfaces. We collected many measures of such moral emotions in our studies, but did not find any clear interpretation. We nevertheless report these measures for interested researchers (see SI, Tables S8, S18, S35).

Finally, we acknowledge that our stylized protocol missed many of the complications of real-world delegation, and that further research is needed to capture the subtleties of these complications. For example, the simple task we used—a die-roll report—had no social component and required no teamwork or collaboration. More complex tasks involving collaboration between machines and humans with other agents, especially human agents, may reduce welfare given a reluctance towards collaborating with machines [43, 44].

Likewise, delegation does not always operate through instructions. Sometimes, principals delegate by selecting one agent among many, based on information about their typical performance or behaviour—without sending the agent-specific instructions. In the SI, we report on an additional study in which we let principals select human or machine agents based on a series of past die-roll reports by these agents (see SI, Additional Study). Principals preferred agents who were dishonest, whether human or machine. Of concern, principals were more likely to choose fully dishonest machine than human agents, increasing the aggregated costs from unethical behaviour. These additional results are just a preliminary foray into the thorny problems of machine delegation. The unprecedented rapid spread of machine agents means that anyone

349 with internet access will soon be able to delegate a myriad of tasks to these agents,  
350 without special access or special knowledge of how they work. This may set the stage  
351 for a surge in unethical behaviour, which we as a society must prepare to curb now.

## 352 **Methods**

### 353 **Study 1**

354 **Sample.** Informed by power analysis using bootstrapping (see SI Additional Study),  
355 we recruited 597 participants from Prolific, striving to achieve a sample that was rep-  
356 resentative of the US population in terms of age, gender and ethnicity ( $M_{age} = 45.7$ ;  
357  $SD_{age} = 16.2$ ; 295 self-identified as male; 289 as female; 13 as other, non-binary or  
358 preferred not to indicate; 78% identified as White, 12% as Black, 6% as Asian, 2% as  
359 Mixed and 2% as Other). A total of 88% of participants had some form of post-high  
360 school qualification. The study, implemented using oTree, was pre-registered.

361 **Procedure and Measures.** After providing informed consent, participants read  
362 the instructions for the die-rolling task [33, 40]. They were instructed to roll a die  
363 privately and to report the observed outcome. They would receive a bonus based on  
364 the number reported. Participants would earn 1 cent for a 1, 2 cents for a 2 and so on  
365 up to 6 cents for a 6. We deployed a previously validated version of the task in which  
366 the die roll is shown on the computer screen [26]. As distinct from the original one-  
367 shot version of the game, participants engaged in 10 rounds of the task, generating a  
368 maximum possible bonus of 60 cents.

369 **Conditions.** Study 1 entailed four between-subjects conditions. In the **Control**  
370 condition ( $n = 152$ ), participants reported the 10 die-roll outcomes themselves. In the  
371 **Rule-Based** condition ( $n = 142$ ), participants specified if-then rules for the machine  
372 agent to follow (see Figure 5, left). Namely, for each possible die-roll outcome, the par-  
373 ticipants indicated what number the machine agent should report on their behalf. In  
374 the **Supervised Learning** condition ( $n = 150$ ), participants chose one of three datasets  
375 on which to train the machine agent. The datasets reflected Honesty, Partial Cheating  
376 and Full Cheating (see Figure 5, middle). In the **Goal-Based** condition ( $n = 153$ ), par-  
377 ticipants specified the machine agent’s goal in the die-rolling task: maximize accuracy

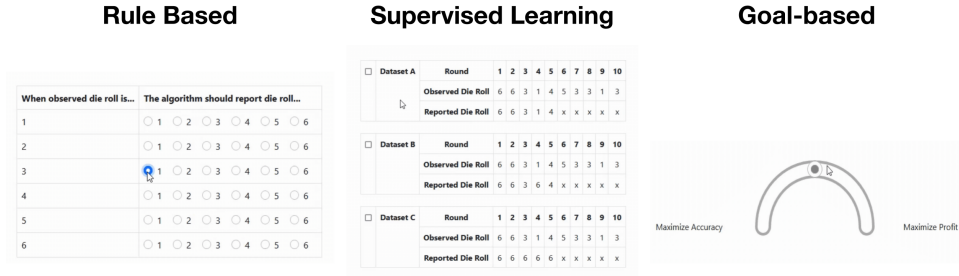


Figure 5: **Overview of the delegation conditions to which participants were assigned.** Left: In the Rule-Based condition, participants specified the number to be reported for each observed die roll. Middle: In the Supervised Learning condition, participants chose a dataset on which to train the algorithm. Right: In the Goal-Based condition, participants specified a high-level goal by setting the dial within a range from maximum accuracy to maximum profit.

or maximize profit, or one of 5 intermediate settings (see Figure 5, right).

**Underlying Algorithms.** For each of the delegation conditions, simple algorithms were implemented. This avoids deception of participants as they engaged in a delegation to a simple machine agent as we had stated in the instructions. For the Rule-Based condition, the algorithm followed simple if-then rules as specified by the user.

For the Supervised Learning condition, the algorithm was implemented by first calculating the difference between the actual and reported rolls for each participant in training data sourced from a pre-pilot in which participants performed an incentivized die-roll task themselves ( $n = 96$ ). The algorithm then probabilistically adjusted future reported outcomes based on these differences, with Dataset A (representing Honesty) having no adjustments, Dataset B having moderate, stochastic adjustments

390 (Partial Cheating), and Dataset C having larger adjustments, tending towards but not  
391 always engaging in Full Cheating. No seed was set for the algorithm in undertaking  
392 its sampling, creating some variance in outcomes reported by the algorithm.

393 For the **Goal-Based** condition, the algorithmic output was guided by the setting  
394 on a 7-notch dial ranging from ‘maximize accuracy’ to ‘maximize profit’. The algo-  
395 rithm adjusted the results of a series of actual die rolls to achieve a desired total sum.  
396 Specifically, the algorithm manipulated a specific list of integers (i.e., 6, 6, 3, 1, 4, 5, 3,  
397 3, 1, 3) representing the sequence of actual die-roll outcomes. The algorithm specified  
398 the desired total sum, here, between 35 (the actual total) and 60 (the maximum out-  
399 come), based on the value of a dial set by the principal. The algorithm then adjusted  
400 the individual integers in the list so that their sum approached the desired total sum.  
401 This was achieved by randomly selecting an element in the integer list and increment-  
402 ing or decrementing its value, depending on whether the current sum of the list was  
403 less than or greater than the total desired sum. This process continued until the sum  
404 of the list equalled the total desired sum specified by the principal, at which point the  
405 modified list was returned and stored to be shown to the principal later in the survey.

406 **Exit Questions.** At the end of the study, we assessed demographics (age, gender,  
407 education) and, using 7-point scales, participants’ level of computer science expertise,  
408 their satisfaction with the payoff, their perceived degree of control over (a) the process  
409 of determining the reported die rolls, and (b) the outcome, how much effort the task  
410 required from them, how guilty they felt about the bonus, how responsible they felt for  
411 choices made in the task, how much they feared punishment, whether the algorithm  
412 worked properly, whether they felt they had reported the die rolls honestly, and the  
413 degree of dishonesty of their behaviour. Finally, where relevant, participants indicated  
414 in an open-text field the reason for the delegation choice.

## 415 **Study 2**

416 **Sample.** We recruited 801 participants from Prolific, striving to be representative of  
417 the US population in terms of age, gender and ethnicity ( $M_{age} = 44.9$ ;  $SD_{age} =$   
418 16.0; 403 self-identified as female; 388 as male; 10 as either other, non-binary or pre-  
419 ferred not to indicate; 77% identified as White, 13% as Black, 6% as Asian, 2% as Mixed,

and 2% as Other). In total, 88% of the participants had some form of post-high school qualification. The study, run on oTree, was preregistered.

**Procedure, Measures and Conditions.** The procedure was identical to Study 1, with the exceptions that: (a) delegation was optional; (b) participants indicated at the end whether they preferred to delegate the decision to a human or an algorithm; and (c) participants completed the previously validated Guilt and Shame Proneness scale [45] at the end of the study.

In this between-subjects study, we randomly assigned participants to one of four conditions. In the **Control** condition ( $n = 205$ ), participants reported the 10 die rolls themselves. Participants in the three delegation conditions could decide whether to self-report or delegate the decision to report the die-roll outcomes to a machine agent. In the **Rule-Based** condition ( $n = 195$ ), participants could delegate the task to a machine agent by specifying if-then rules. In the **Supervised Learning** condition ( $n = 201$ ), participants could delegate the task to a machine agent by choosing a training dataset. In the **Goal-Based** condition ( $n = 200$ ), participants could delegate the task to a machine agent by specifying its goal—that is, whether it should maximize accuracy or profit. As we did not expect participants to be familiar with programming instructions to machine agents in these interfaces, the process was described in text and demonstrated in a video played on loop for each interface. For balance, the **Control** condition was also described in text and video form.

### Study 3

Study 3 consisted of three parts, each relating to relevant parties in a delegation: (a) principals (delegators), (b) agents (delegates) and (c) third parties. In Study 3a, we collected principals’ instruction texts for human and AI agents and their own self-reported behaviour in the task. Additionally, we measured the behaviour they intended for agents by having them report their expected outcomes for each of the 10 die rolls. In Study 3b, we compared the behaviour of human and machine (LLM) agents. Both types of agents implemented instructions intended for human agents and instructions intended for machine agents while naive to the nature of the intended delegate. In Study 3c, independent human raters assessed how much dishonesty intent

450 was implied in the content of all instruction texts; they too were naive to the nature of  
451 the intended delegate. These third-party evaluations were intended to provide percep-  
452 tions of the agent’s intended behaviour, unaffected by any moral costs of implementing  
453 such instructions that the agents may experience.

### 454 **Study 3a**

455 **Sample.** For Study 3a, we recruited 390 participants from Prolific, striving to be rep-  
456 resentative of the US population in terms of age, gender, and ethnicity ( $M_{age} = 46.0$ ;  
457  $SD_{age} = 15.9$ ; 196 self-identified as female; 189 as male; 5 as other, non-binary or  
458 preferred not to indicate; 76% identified as White, 13% as Black, 6% as Asian, 3% as  
459 Mixed and 2% as Other). In total, 86% of the participants had some form of post-high  
460 school qualification. The study, conducted on Qualtrics, was preregistered.

461 **Procedure, Measures and Conditions.** Study 3a entailed three within-subjects  
462 conditions for principals: self-report (control), delegate to machine (chatbot) agent and  
463 delegate to human agent. Before engaging in the task, participants were given general  
464 information on the die-rolling task. They were then shown the payoff table describing  
465 how the reported die rolls would translate to payoffs (i.e., 1 cent for a 1, 2 cents for a  
466 2 and so on, up to 6 cents for a 6). In random order, participants who passed a com-  
467 prehension check then completed the self-report and both delegation conditions. In  
468 the delegation conditions, they were asked to write short instructions in natural lan-  
469 guage for human and machine agents, indicating how they should report the die-roll  
470 outcomes. Participants learned that one condition would be randomly chosen to be  
471 payoff relevant. In the machine delegation condition, participants received detailed  
472 information about how to program the machine agent to report the 10 die rolls. Since  
473 participants may not have had a clear conception of how and whether the machine  
474 agent understands natural language instructions, we included a short video showing  
475 how it implemented different types of instructions: Honesty, Partial Cheating and Full  
476 Cheating. Instructions were chosen from a pilot ( $n=9$ ) study in which participants  
477 produced instructions. We drew upon instructions including those with nuance in  
478 conveying unethical intentions by means of indirect speech [46]. To balance the video  
479 presentation across conditions and avoid a condition-specific priming effect [47], we

also showed short videos in the self-report and human agent conditions. These videos displayed, in random order, three examples of die-roll reporting that reflected Honesty, Partial Cheating and Full Cheating for the same die-roll outcome. After watching these short videos, participants engaged in the three tasks: self-reporting 10 die rolls, delegating to human agents and delegating to machine agents. After completing all three tasks, participants were asked to indicate the behaviour they intended from the human and machine agents. To this end, they were reminded of the text they had written for the respective agent and asked to indicate for 10 observed die rolls what outcome they intended the human/machine agent to report on their behalf.

**Exit Questions** At the end of the study, we assessed demographics (age, gender, education) and, using 7-point scales, participants' level of computer science expertise, their previous experience with the die-rolling experiment and with LLMs, their feelings of guilt and responsibility when delegating the task, their expectations regarding the guilt experienced by agents, their expectation as to which agent (machine or human) implementation would align more closely with their intentions, and whether they would prefer to delegate comparable future tasks to human or machine agents or to do it themselves.

**Automated Response Prevention and Quality Controls** To reduce the risk of automated survey completion, we included a reCaptcha at the beginning of the survey and checked via Javascript whether participants copy-pasted text into the text fields when writing instructions to agents. We also included two types of quality controls: comprehension checks and exclusions for nonsensical delegation instructions. Participants were informed that they had two attempts each to answer each comprehension check question correctly to be eligible for the bonus (maximum = 60 cents) and that they would be excluded from any bonus payment if they wrote nonsensical instructions in the delegation conditions. All participants were paid participant fees irrespective of the nature of their survey answers.

### Study 3b

**Sample.** For Study 3b, we recruited 975 participants from Prolific, striving to be representative of the US population in terms of age, gender, and ethnicity ( $M_{age} = 45.4$ ;

510  $SD_{age} = 15.8$ ; 482 self-identified as female; 473 as male; 20 as other, non-binary or  
511 preferred not to indicate; 78% identified as White, 13% as Black, 6% as Asian, 2% as  
512 Mixed and 1% as Other). In total, 88% of the participants had some form of post-high  
513 school qualification. The study, run on Qualtrics, was preregistered. For Study 3b, we  
514 pilot-tested the experimental setup with 20 participants who were asked to implement  
515 three sample instructions from a previous pilot study for Study 3a ( $n=9$ ).

### 516 **Procedure, Measures and Implementations**

517 **Machine Agents.** With the aim of assessing the generalizability of findings across  
518 open- and closed-source models, we sought to use both LLAMA and GPT-4. However,  
519 as the results provided by LLAMA were qualitatively inferior (e.g., not complying with  
520 the instruction, generating unrelated text, not providing an interpretable answer), we  
521 report analyses only for GPT-4 (Version November 2023). In a prompt, we described  
522 the die-rolling task, including the bonus payoffs for principals, to GPT-4. GPT-4 was  
523 then informed that it was the delegate (agent) in the task, given instructions from prin-  
524 cipals, and asked to report the die-roll outcomes. The exact wording of the prompt is  
525 contained in SI, Prompt Texts.

526 **Human Agents.** The implementation of principal instructions by human agents  
527 followed the process conducted with machine agents as closely as possible. Again, the  
528 instructions included those intended for human agents and those intended for ma-  
529 chine agents (which we describe as 'forked'). Participants were naive as to whether the  
530 instructions were drafted for a human or a machine agent.

531 It began with a general description of the die-rolling task. The next screen in-  
532 formed participants that people in a previous experiment (i.e., principals) had written  
533 instructions for agents to report a sequence of 10 die rolls on their behalf. Participants  
534 learned that they would be the agents and report on 10 die rolls for four different in-  
535 struction texts and that their reports would determine the principal's bonus.

536 Participants were incentivized to match the principals' intentions: For one ran-  
537 domly selected instruction text, they could earn a bonus of 5 cents for each die roll  
538 that matched the principal's expectations, giving a maximum bonus of 50 cents. Par-  
539 ticipants were presented with one instruction text at a time, followed by the sequence  
540 of 10 die rolls, each of which they reported on the principal's behalf.

541 **Exit Questions** At the end of the study, we assessed demographics (age, gender,



education) and, using 7-point scales, participants' level of computer science expertise, 542  
their previous experience with the die-rolling experiment and with LLMs, their expe- 543  
rienced guilt for each instruction implementation, and whether they could correctly 544  
identify whether an instruction was intended for a human or a machine agent. 545

### Study 3c 546

**Sample.** For the human raters in Study 3c, we recruited 98 participants ( $M_{age} = 37.5$ ; 547  
 $SD_{age} = 12.3$ ; 58 self-identified as female; 38 as male; 2 as other, non-binary or pre- 548  
ferred not to indicate; 60% identified as White, 8% as Black, 22% as Asian, 2% as Mixed 549  
and 8% as Other). In total, 86% of the participants had some form of post-high school 550  
qualification. The study, conducted within a React-based app, was preregistered. 551

**Procedure, Measures and Implementations** We adopted a multi-pronged ap- 552  
proach to categorize the honesty level of natural language instructions in Study 3c. 553

**Self-Categorization.** Principals indicated what they expected the agent to report 554  
for each die-roll outcome over 10 rounds, based on the instructions they gave. We then 555  
used the same criteria as in Studies 1 and 2 to categorize their behavioural intention 556  
as Honesty, Partial Cheating, or Full Cheating. 557

**LLM Categorization.** GPT-4 (Version November 2023) was prompted to evalu- 558  
ate principals' instructions (see SI, Prompt Texts). First, we presented side-by-side in 559  
randomized order, each pair of instructions given by principals in Study 3a (one in- 560  
tended for a human agent and one intended for a machine agent). GPT-4 was naive to 561  
the nature of the intended agent. GPT-4 was instructed to indicate which of the two 562  
instructions entailed more dishonesty or if they both had the same level of intended 563  
dishonesty. We then instructed GPT-4 to classify both of the instructions as Honest, 564  
Partial Cheating or Full Cheating, and to predict the estimated sum of reported die 565  
rolls. For the full prompt, see SI Prompt Texts. 566

**Rater Categorization.** This followed the LLM Categorization process as closely 567  
as possible. The human raters were given a general description of the die-rolling task. 568  
They were then informed that people in a previous experiment had written instruc- 569  
tions for agents to report a sequence of 10 die rolls on their behalf. Participants were 570  
informed they would act as raters and compare a series of instruction pairs and indi- 571

cate if which of the two instructions entailed more dishonesty or if they both had the same level of intended dishonesty. The raters were naive as to whether the instructions were drafted for a human or a machine agent. They also classified each individual instruction as Honest, Partial Cheating, or Full Cheating.

**Exit Questions** At the end of the study, we assessed demographics (age, gender, education) and, using 7-point scales, participants' level of computer science expertise and their previous experience with LLMs.

**Data Availability.** The pre-registrations, survey instruments, and data for all studies are available at: Open Science Framework.

**Code Availability.** The code, written in R, used for analyses and data visualisations, is available at: Open Science Framework.

**Ethics Statement.** We confirm that all studies complied with all relevant ethical guidelines. The Ethics Committee of the Max Planck Institute for Human Development approved all studies. Informed Consent was obtained from all human research participants in these studies.

## Acknowledgments

All authors thank Georg Kruse for his research assistance, Heather Barry-Kappes for helpful feedback, Ivan Soraperra for statistical guidance and Susannah Goss for language editing. JFB acknowledges support from grant ANR-19-PI3A-0004, grant ANR-17-EURE-0010, and the research foundation TSE-Partnership.

## References

1. Gogoll, J. & Uhl, M. Rage against the machine: Automation in the moral domain. *Journal of Behavioral and Experimental Economics* **74**, 97–103 (2018).
2. Rahwan, I. *et al.* Machine behaviour. *Nature* **568**, 477–486 (2019).
3. BBC. *Tesla adds chill and assertive self-driving modes* <https://www.bbc.com/news/technology-59939536>. 2022, January 1.

4. Hendershott, T., Jones, C. M. & Menkveld, A. J. Does algorithmic trading improve liquidity? *The Journal of Finance* **66**, 1–33 (2011). 598  
599
5. Holzmeister, F., Holmén, M., Kirchler, M., Stefan, M. & Wengström, E. Delegation decisions in finance. *Management Science* **69**, 4828–4844 (2023). 600  
601
6. Raghavan, M., Barocas, S., Kleinberg, J. & Levy, K. *Mitigating bias in algorithmic hiring: Evaluating claims and practices in Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (eds Hildebrandt, M. & Castillo, C.) (Association for Computing Machinery, 2020), 469–481. 602  
603  
604  
605
7. McAllister, A. Stranger than science fiction: The rise of AI interrogation in the dawn of autonomous robots and the need for an additional protocol to the UN convention against torture. *Minnesota Law Review* **101**, 2527–2573 (2016). 606  
607  
608
8. Dawes, J. The case for and against autonomous weapon systems. *Nature Human Behaviour* **1**, 613–614 (2017). 609  
610
9. Dell’Acqua, F. *et al.* Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *Harvard Business School Working Paper Series* **24-013** (2023). 611  
612  
613
10. Schrage, M. 4 models for using AI to make decisions. *Harvard Business Review*. <https://hbr.org/2017/01/4-models-for-using-ai-to-make-decisions> (2017, January 27). 614  
615  
616
11. Herrmann, P. N., Kundisch, D. O. & Rahman, M. S. Beating irrationality: Does delegating to IT alleviate the sunk cost effect? *Management Science* **61**, 831–850 (2015). 617  
618  
619
12. Fernández Domingos, E. *et al.* Delegation to artificial agents fosters prosocial behaviors in the collective risk dilemma. *Scientific Reports* **12**, Article 8492 (2022). 620  
621
13. De Melo, C. M., Marsella, S. & Gratch, J. Human cooperation when acting through autonomous machines. *Proceedings of the National Academy of Sciences* **116**, 3482–3487 (2019). 622  
623  
624
14. Calvano, E., Calzolari, G., Denicolo, V. & Pastorello, S. Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review* **110**, 3267–3297 (2020). 625  
626  
627

- 628 15. Calvano, E., Calzolari, G., Denicolò, V., Harrington Jr, J. E. & Pastorello, S. Pro-  
629 tecting consumers from collusive prices due to AI. *Science* **370**, 1040–1042 (2020).
- 630 16. Köbis, N., Bonnefon, J.-F. & Rahwan, I. Bad machines corrupt good morals. *Na-  
631 ture Human Behaviour* **5**, 679–685 (2021).
- 632 17. Bonnefon, J.-F., Rahwan, I. & Shariff, A. The moral psychology of artificial intel-  
633 ligence. *Annual Review of Psychology* **75**, 653–675 (2024).
- 634 18. Abeler, J., Nosenzo, D. & Raymond, C. Preferences for truth-telling. *Econometrica*  
635 **87**, 1115–1153 (2019).
- 636 19. Gerlach, P., Teodorescu, K. & Hertwig, R. The truth about lies: A meta-analysis  
637 on dishonest behavior. *Psychological Bulletin* **145**, 1–44 (2019).
- 638 20. Dana, J., Weber, R. A. & Kuang, J. X. Exploiting moral wiggle room: Experiments  
639 demonstrating an illusory preference for fairness. *Economic Theory* **33**, 67–80  
640 (2007).
- 641 21. Leblois, S. & Bonnefon, J.-F. People are more likely to be insincere when they  
642 are more likely to accidentally tell the truth. *Quarterly Journal of Experimental  
643 Psychology* **66**, 1486–1492 (2013).
- 644 22. Vu, L., Soraperra, I., Leib, M., van der Weele, J. & Shalvi, S. Ignorance by choice:  
645 A meta-analytic review of the underlying motives of willful ignorance and its  
646 consequences. *Psychological Bulletin* **149**, 611–635 (2023).
- 647 23. Bartling, B. & Fischbacher, U. Shifting the blame: On delegation and responsi-  
648 bility. *The Review of Economic Studies* **79**, 67–87 (2012).
- 649 24. Weiss, A. & Forstmann, M. Religiosity predicts the delegation of decisions be-  
650 tween moral and self-serving immoral outcomes. *Journal of Experimental Social  
651 Psychology* **113**, Article 104605 (2024).
- 652 25. Erat, S. Avoiding lying: The case of delegated deception. *Journal of Economic Be-  
653 havior & Organization* **93**, 273–278 (2013).
- 654 26. Kocher, M. G., Schudy, S. & Spantig, L. I lie? We lie! Why? Experimental evidence  
655 on a dishonesty shift in groups. *Management Science* **64**, 3995–4008 (2018).

27. Contissa, G., Lagioia, F. & Sartor, G. The Ethical Knob: Ethically-customisable automated vehicles and the law. *Artificial Intelligence and Law* **25**, 365–378 (2017).
28. Russell, S. J. & Norvig, P. *Artificial intelligence: A modern approach* (Pearson, 2016).
29. Sutton, R. S. & Barto, A. G. *Reinforcement learning: An introduction* (MIT Press, 2018).
30. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* **267**, 1–38 (2019).
31. King, T. C., Aggarwal, N., Taddeo, M. & Floridi, L. Artificial intelligence crime: An interdisciplinary analysis of foreseeable threats and solutions. *Science and Engineering Ethics* **26**, 89–120 (2020).
32. Nussberger, A.-M., Luo, L., Celis, L. E. & Crockett, M. J. Public attitudes value interpretability but prioritize accuracy in Artificial Intelligence. *Nature Communications* **13**, Article 5821 (2022).
33. Fischbacher, U. & Föllmi-Heusi, F. Lies in disguise: An experimental study on cheating. *Journal of the European Economic Association* **11**, 525–547 (2013).
34. Gächter, S. & Schulz, J. F. Intrinsic honesty and the prevalence of rule violations across societies. *Nature* **531**, 496–499 (2016).
35. 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).
36. Cohn, A. & Maréchal, M. A. Laboratory measure of cheating predicts school misconduct. *The Economic Journal* **128**, 2743–2754 (2018).
37. Rustagi, D. & Kroell, M. Measuring honesty and explaining adulteration in naturally occurring markets. *Journal of Development Economics* **156**, Article 102819 (2022).
38. Bandura, A., Barbaranelli, C., Caprara, G. V. & Pastorelli, C. Mechanisms of moral disengagement in the exercise of moral agency. *Journal of Personality and Social Psychology* **71**, 364–374 (1996).

39. Mazar, N., Amir, O. & Ariely, D. The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research* **45**, 633–644 (2008).
40. Shalvi, S., Dana, J., Handgraaf, M. J. & De Dreu, C. K. Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior. *Organizational Behavior and Human Decision Processes* **115**, 181–190 (2011).
41. Candrian, C. & Scherer, A. Rise of the machines: Delegating decisions to autonomous AI. *Computers in Human Behaviour* **134**, Article 107308 (2022).
42. Steffel, M., Williams, E. F. & Perrmann-Graham, J. Passing the buck: Delegating choices to others to avoid responsibility and blame. *Organizational Behavior and Human Decision Processes* **135**, 32–44 (2016).
43. Dvorak, F., Stumpf, R., Fehrler, S. & Fischbacher, U. Generative AI triggers welfare-reducing decisions in humans. *arXiv* **2401.12773** (2024).
44. Ishowo-Oloko, F. *et al.* Behavioural evidence for a transparency–efficiency trade-off in human–machine cooperation. *Nature Machine Intelligence* **1**, 517–521 (2019).
45. Cohen, T. R., Wolf, S. T., Panter, A. T. & Insko, C. A. Introducing the GASP scale: A new measure of guilt and shame proneness. *Journal of Personality and Social Psychology* **100**, 947–966 (2011).
46. Pinker, S., Nowak, M. A. & Lee, J. J. The logic of indirect speech. *Proceedings of the National Academy of Sciences* **105**, 833–838 (2008).
47. Pataranutaporn, P., Liu, R., Finn, E. & Maes, P. Influencing human–AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness. *Nature Machine Intelligence* **5**, 1076–1086 (2023).