The hidden costs of automation*

Christina Strobel†

August 31, 2019

Automated processes are taking over more and more tasks from humans. This is also the case in performance-related pay systems. The aim of this paper is to examine whether the use of an *automated Performance Appraisal System* (*automated PAS*) influences human performance and whether it matters if the company or the direct line manager decides to use an *automated PAS*. We present a modified Principal-Agent Game in a real job market using Amazon Mechanical Turk (Amazon MTurk). Depending on the treatment, either the principal or the system chooses whether to use an *automated PAS* or a *manual Performance Appraisal System* (*manual PAS*). We find that performance is significantly lower under an *automated PAS* than under a *manual PAS*. However, performance does not differ significantly depending on whether the principal or the system decides to use an *automated PAS* or a *manual PAS*.

Keywords: Principal-Agent-Setting; Automation; Performance-related Pay; Performance
JEL classification: C91, D63, D80

*This document was created on August 31, 2019, with R version 3.4.1 (2017-06-30), on x86_64-w64-mingw32. We thank the Max Planck Society for financial support through the International Max Planck Research School on Adapting Behavior in a Fundamentally Uncertain World (IMPRS Uncertainty). We also thank the audience at IMPRS Uncertainty doctoral seminars and the members of the Junior Research Group "Ethics of Digitization" for their valuable feedback.

†FSU Jena, School of Economics, Bachstraße 18k, 07737 Jena, Christina.Strobel@uni-jena.de.

1. Introduction

Most wages and salaries of top managers and mid-level employees have a fixed and a variable compensation component. Thereby, further compensation payments such as bonus payments, company shares or non-monetary remuneration benefits enhance the employee's fixed compensation. By far the most commonplace form of variable payment is bonus payments, i.e. non-guaranteed direct payments to employees in addition to their base wage. Bonuses are usually paid in recognition of performance goal achievements, which can be at the corporate, team or individual level. While the company's and the team's success are determined on the basis of key performance indicators, the employee's individual performance goal achievements are usually assessed by their direct supervisor in a so-called *performance appraisal* – a systematic process to evaluate the employee's performance on the job.

Nowadays, algorithm-controlled systems take over more and more tasks from humans. This is not only the case in the manufacturing sector – where manufacturing robots replace assembly line workers – but also in the management sector, where the level of process automation continues to increase. An example of this development can be found in human resources management, where, as Kaur and Sood (2017) describe, automated processes are going to replace direct human-to-human interactions in employees' performance evaluation. For example, employees' performance ratings influence bonus payments in many companies. Therefore, colleagues and supervisors rate employees' performance. An employee's overall performance rating is then calculated from the sum of all performance feedback as well as department and company performance indicators. If a certain performance-point threshold is reached, the employee is entitled to receive a bonus. While in the past a supervisor would have decided about final bonus payments, bonus assessments are now automated, leaving only very limited room for action by the supervisor. Thus, the employee's bonus payment depends much less on an individual's assessment than on a predetermined algorithm.

We know from former research that performance is influenced by situational circumstances. For example, Falk and Kosfeld (2006) found that setting a minimum performance requirement has a negative effect on performance. Furthermore, Corgnet et al. (2019) show that humans perform better in a working environment with only humans than in a working environment in which they interact with a machine. Letting an algorithm instead of a human decide whether a bonus is paid or not means changing the situational circumstances of the employee's performance evaluation process and, hence, might also influence the worker's performance.

For example, an algorithm is based on standardized processes and is built upon predetermined rules that are programmed ex-ante. According to Stone (1971), using an algorithm generally makes a process less individualizable and more rigid. Furthermore, if implementing a more automated approach is interpreted as a decrease in appreciation of the individual's work performance, the use of such a system might lead to a decrease in the employee's satisfaction and, in turn, work performance. Another important factor might be whether the supervisor or the company decides to use a *manual Performance Appraisal System* (*manual PAS*) or an *automated Performance Appraisal System* (automated
PAS). We know from former research by Fehr et al. (1993) that people tend to reward kind

actions and to punish unkind ones in an experimental labor market.[1] Thus, when a person makes the decision about the *PAS*, workers might respond to a decision that they perceive as kind [unkind] with higher [lower] performance. However, if a system makes the decision about the *PAS*, there is no reason to do so.

Given the business world's preoccupation with more and more automation, we will inevitably see how employees react to an *automated PAS*. However, to the best of my knowledge, the influence of more or less automation in *PAS*s has not been investigated yet. Our aim is to analyze whether an *automated PAS* has the same influence on an employee's effort as a *manual PAS*. Therefore, I use a simple two-stage Principal-Agent setup where the agent chooses a productivity level and the principal chooses a threshold for a bonus payment.

We found that performance is significantly lower under an *automated PAS* than under a *manual PAS*. However, performance was not influenced by whether a person or the system decided to use a specific *PAS*.

The remainder of the paper is organized as follows: Section 2 provides a literature review focusing on experimental evidence from economics and social psychology research. In Section 3 we describe the basic experimental design. Then, in Section 4, we relate the experiment to the theoretical background and derive behavioral predictions. We present the results in Section 5. Section 6 concludes the paper by summarizing the main findings and discussing their implications as well as further research ideas.

2. Related literature

In this section, we present previous research on bonus payments, job satisfaction, and performance appraisals related to the experiment.

Bonus payments increase effort provision and seem to work better than fixed-wage contracts and contracts that include the possibility of punishing the employee for bad performance. Fehr and Schmidt (2007) show that a bonus contract offered by the principal in a chosen-effort Principal-Agent experiment leads to higher effort provision by the agent than a contract that fines the agent or a trust contract where the principal offers a fixed wage. Fehr et al. (2007) confirm this observation, finding bonus contracts that rely on fairness and trust as an enforcement device to be more efficient and more profitable than incentive contracts enforced by the courts. In their experiment, the principal was able to choose a mechanism to enforce a specific effort from the agent with the support of a third party or to announce a non-binding, voluntary bonus payment instead if the agent's effort was satisfactory. The results show that a non-binding, voluntary bonus payment leads to higher performance than an explicit incentive contract, which fines the agent for unsatisfactory performance.

The positive effect of bonus payments on effort provision seems to be enhanced if the bonus payment is combined with an individual performance appraisal. Kampkötter (2017) analyzes the relationship between performance appraisals and job satisfaction using data from the German Socio-Economic Panel (SOEP) Study. He finds that a monetary performance

[1] Reciprocal behavior is also confirmed by further experimental research (e.g., Fehr et al., 1993; Berg et al., 1995) and by theoretical research (e.g., Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006).

appraisal by the supervisor increases the employee's job satisfaction and leads to higher effort provision compared to a performance appraisal without a bonus payment. A literature review of over 300 papers done by Levy and Williams (2004) indicates that not only the monetary incentive but also the performance process itself increases job satisfaction and effort provision. The review shows that the supervisor's recognition and appreciation of the work performed are, aside from the monetary benefit, the main factor that causes an increase in an employee's job satisfaction. Ockenfels et al. (2015) confirm the observation that the process itself plays an important role when it comes to the influence of bonus payments on performance. Perhaps surprisingly, Ockenfels et al. find that transparency does not increase job satisfaction per se but can also amplify dissatisfaction. In their real-effort experiment, two agents received either a high or a low bonus payment for their performance. After the bonus payments were assigned, the agents played a Public Goods Game and a Dictator Game to transfer a part of their endowment to the principal in response to the bonus decision. Agents who received a higher bonus transferred 27.3% of their endowment while agents who received a lower bonus transferred 7.5% of their endowment. When the bonus payments were transparent (i.e. the agent got to know the percentage of the bonus budget (s)he received), the response toward the principal was significantly more negative. In the treatment where the bonus payments were not transparent, the transfers did not differ significantly. Experiments also indicate that employee effort is sensitive to the level of control. Fehr and Rockenbach (2003) and Fehr and List (2004) show that the principal's decision to use a punishment device leads to a decrease in effort provision by the agent in Trust Games. In a similar spirit, Falk and Kosfeld (2006) and Kajackaite and Werner (2015), find that controlling the agent has a negative influence on their performance. In a chosen-effort experiment conducted by Falk and Kosfeld, the agent had to choose a costly productivity activity that benefited the principal while the principal had the choice to either control (i.e. enforce a minimum effort) or trust the agent. The results show that the majority of the agents reduced their performance because most agents perceived control as a signal of distrust and low expectations by the principal. Kajackaite and Werner build upon the finding that control has a counterproductive effect on performance provision by showing that the principal's active decision to control affects the agent's kindness perception and triggers reciprocal responses. However, they find no significant change in the average output level in a real-effort experiment if the principal decides to implement a minimum performance requirement.

The experimental finding that effort provision is not only influenced by monetary incentives but also by situational factors such as trust, transparency, and control is also supported by theoretical work. The model by Ellingsen and Johannesson (2005) shows that esteem influences performance, as a generous and trusting contract can elicit better performance from agents than a contract with low pay and strong incentives. Thereby, the model is based on the assumption that the principal's behavior conveys expectations about the agent and that these expectations might influence how the principal will rate the agent's performance ex-post. If the principal decides not to control the agent, the principal signals trust in the agent. This makes it harder for the agent to justify poor performance, in the sense that poor performance is not consistent with acceptable esteem. If the principal decides to control the agent, the principal shows a pessimistic expectation. In this case, the principal signals that low performance will not surprise them, which makes it easier for the agent to show low

performance.

Research on the relationship between human performance and automated systems, however, is just emerging. Corgnet et al. (2019) look at performance in a sequential task where participants had to work together with either another human or a robot – calibrated to the performance of an average worker – to fill out a grid. The results show that human performance was significantly lower when the participants were matched with a robot compared to when matched with another human. To the best of my knowledge, there is no study that investigates the link between automated bonus payments and performance.

3. Experimental design

We implemented a modified two-stage Principal-Agent Game like the design used by Falk and Kosfeld (2006). The agent had an initial endowment of 120 Points, while the principal's initial endowment was 0 Points. 1 Point equaled \$0.01 USD. The agent chose a productive activity x, and the cost of the productive activity for the agent was $c(x) = x$. To generate an overall benefit, which is the aim of every market-based system, the marginal benefits need to exceed the marginal costs. Thus, the principal earned two times the agent's effort $p(x) = 2x$.[2] The principal then determined a threshold x_t for a 'very good transfer' that the agent had to reach to get a bonus of $b^* \in \{0, 120\}$. The experimenter paid the bonus. Depending on the treatment, either the principal (treatment *HUMAN*) or the system (treatment *SYSTEM*) decided whether to use the *manual PAS* or the *automated PAS* mechanism. If using a *manual PAS* mechanism, the principal knew the agent's effort before determining a minimal threshold x_t for the agent to get a bonus b^* and decided about the agent's performance *ex-post*. If using an *automated PAS* mechanism, the principal did not know the agent's effort before determining a certain threshold x_t that had to be reached by the agent to get a bonus b^* and therefore decided about the agent's performance *ex-ante*. The agent got a bonus if $x_t \geq x$. Thus, the payoff functions were $\prod_P = 2x$ for the principal and $\prod_A = 120 - x + b^*_{x_t}$ for the agent.

3.1. Treatments

We conducted two treatments: *HUMAN* and *SYSTEM*. In treatment *HUMAN*, the principal decided if (s)he wanted to use the *automated PAS* or the manual PAS. In treatment *SYSTEM*, a random mechanism called the system decided to use either the *automated PAS* or the *manual PAS* with a probability of 50%. In both treatments, the agents' efforts were elicited with the help of a strategy method. Thus, the agents had to state how much of their initial endowment they would like to transfer to the principal if a *manual PAS* or an *automated PAS* was used. The results by Falk and Kosfeld (2006) do not indicate a difference between using the strategy method or the specific response method, and Charness et al. (2018) find qualitatively similar results for real-effort and stated-effort designs in a meta-study on effort measures in economic experiments. We therefore waived conducting an extra specific response treatment.

[2] We used an $p(x) = 2x$ mechanism as it allows the agent to form beliefs about the threshold set by the principal more easily than a more complex mechanism.

3.2. Procedure

We collected the data online via Amazon Mechanical Turk (MTurk) using oTree (Chen et al., 2016). All experimental stimuli and instructions were presented through a computer interface. Participants received a show-up fee of \$0.50 USD. We used a between-subjects design so the data for all statistical tests are independent for the two treatments. The order of the *PAS*s was randomly alternated for each participant in both treatments to control for potential order effects.

At the beginning of the experiment, all participants had to pass a human test to ensure only human participants would participate in the experiment. Therefore, the participants had to add up two two-digit numbers and write the correct answer into an input field. Participants who passed the human test were randomly assigned to a group of two as well as to a role. Each group consisted of one agent (labeled participant A) and one principal (labeled participant B). Participants were first provided with experimental instructions.[3] After reading the instructions, they acted in four different stages:

In stage (i), all participants had to answer a set of control questions to ensure they had understood the instructions before proceeding.

In stage (ii), the productive activity, each agent had to decide how much of his/her initial endowment (s)he wanted to transfer to his/her principal.

In treatment HUMAN, before the principal decided about the threshold, each principal had to decide if (s)he wanted to use the *manual PAS* or the *automated PAS*. In treatment *SYSTEM*, the system decided whether to use the *manual PAS* or the *automated PAS*.

In stage (iii), the bonus threshold task, the principal had to decide about a minimum transfer threshold that had to be reached by the agent's transfer for him/her to get the bonus.

In stage (iv), the additional questions, the agent and the principal were asked to answer questions related to their expectations, their risk appetite and if they perceived the procedure to be fair. Participants were also asked about their age and gender. Furthermore, agents were asked about their thoughts on the minimal transfer requested by the principal. As we are not interested in the accuracy of subjective beliefs, we decided not to incentivize agents' beliefs about the minimal transfer for getting a bonus. This also avoids hedging between the decision about how much to transfer and the expected payoff as a result of a correct belief.

4. Behavioral predictions

According to standard economic theories, the agent strives to maximize his/her payoff without taking into account social or situational circumstances. To maximize his/her payoff the agent could choose the productive activity to be $x = 0$. In this case, the agent maximizes his/her payoff without any risk, while the principal's payoff would be $\pi_p = 2x = 0$. However, as we know from social preference theories (e.g., Bolton and Ockenfels, 2000; Charness and Rabin, 2002; Fehr and Schmidt, 1999), individuals do also have other-regarding preferences. Therefore, we expect a substantial fraction of agents to choose a productive activity $x > 0$.

[3] We provide the instructions in Appendix A.1.

Social preference	Preferred productive activity (x)
Equality	~ 80
Fair split	~ 60
Efficiency	~ 120

The table shows the possible social preferences of the principals and the corresponding preferred productive activities.

Table 1: Preferred productive activities (x) by principals.

The threshold set by the principal determines whether the agent receives the bonus payment or not. However, the agent does not know the threshold set by the principal, neither in the *manual PAS* nor in the *automated PAS*. Thus, the agent's productive activity depends on their expectation about the principal's desired productive activity. As shown in Table 1, different social preferences suggest that principals might prefer certain productive activities by the agent over others. For example, allocative fairness models (e.g., Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000) assume that an individual's utility is negatively influenced by an unequal outcome. Therefore, depending on the principal's degree of inequity aversion, the principal might prefer a more or less equal outcome and thus want the agent to choose a productive activity roughly around $x \approx 80$. On the other hand, the principal could attach importance to a fair split. In this case, the principal might demand an equal split of the agent's initial endowment and therefore prefer the agent to choose a productive activity of $x \approx 60$. If the principal has a strong preference to maximize overall social welfare, as for example found by Charness and Rabin (2002), the principal might want to maximize the overall outcome and thus prefer the agent to choose a productive activity of $x = 120$.

In an *automated PAS*, the principal sets the threshold before knowing the agent's performance. Thus, the agent is best off choosing the productive activity that (s)he expects the principal to prefer. In a *manual PAS*, however, the principal gets to know the agent's productive activity before setting the bonus threshold. This certainty about the agent's productive activity increases the principal's influence on the bonus payment. It allows the principal not only to decide about a threshold but also to directly decide whether to pay a bonus or not. From models of social image concerns (e.g., Bénabou and Tirole, 2006; Andreoni and Bernheim, 2009) and concepts of self-perception maintenance (e.g., Rabin, 1995; Beauvois and Joule, 1996) we know, that individuals perceive an unpleasant tension or disutility if their actions cause harm to their social-concept and /or self-concept of being a kind and fair individual. In this respect, deciding about a threshold that denies a bonus payment to the agent in the *manual PAS* causes an immediate disutility for the principal, while determining a threshold in an *automated PAS* that eventually denies the bonus payment does not. Furthermore, the identifiable victim effect by Jenni and Loewenstein (1997) proposes that individuals are more generous and offer greater aid the more they know about the other person.[4] Thus, knowing the agent's productive activity of the agent might be sufficient to increase the prin-

[4]Experimental studies by Hsee and Weber (1997), Kogut and Ritov (2005) and Small and Loewenstein (2003) support the identifiable victim effect. Small and Loewenstein even find that only the determination of the recipient without any personalized information leads to more giving in Dictator Games.

cipal's willingness of the principals to set a lower threshold in a *manual PAS* than they would in an *automated PAS*.

Based on the considerations above, agents might perceive a *manual PAS* to be less rigid than an *automated PAS*. Furthermore, agents might also appreciate the fact the decision about whether the bonus is payed or not is rather an individual than a standardized one and thus might feel more appreciation for their chosen productive activity in a *manual PAS* than in an *automated PAS*. Therefore, we expect the agents to choose a higher productive activity, i.e., transfer more points, in an *manual PAS* than in an *automated PAS* (Hypothesis 1).

Hypothesis 1 *More points are transferred in a manual PAS than in an automated PAS in*

(i) treatment HUMAN, and in

(ii) treatment SYSTEM.

In treatment *HUMAN*, the principal has to make two decisions. The principal first decides whether to use a *manual PAS* or an *automated PAS*, and then decides on a threshold for the bonus payment. From the theory on intention-based reciprocity by Rabin (1993), Dufwenberg and Kirchsteiger (2004), and Falk and Fischbacher (2006) we know that people tend to reward kind intentions and to punish unkind ones. By actively choosing an *automated PAS*, the principal signals that (s)he wants to avoid knowing the agent's productive activity before determining the threshold. Choosing an *automated PAS* also means actively abstaining from directly controlling the bonus payment. Thus, the principal's decision to use an *automated PAS* could be perceived by the agent as a lack of interest in or appreciation of the agent's productive activity. Therefore, the agent might reciprocate by choosing a lower productive activity if the principal decides to use an *automated PAS*. In treatment *SYSTEM*, however, the system and not the principal decides whether to use a *manual PAS* or an *automated PAS*. Hence, the agent should not show a behavioral reaction based on intention-based reciprocity. Thus, we expect the agents to choose a higher productive activity if the principal decides to use a *manual PAS* in treatment *HUMAN* compared to when the system chooses the *manual PAS* in treatment *SYSTEM*. Correspondingly, we expect the agents to choose a lower productive activity if the principal decides to use an *automated PAS* in treatment *HUMAN* compared to when the system chooses the *automated PAS* in treatment *SYSTEM* (Hypothesis 2).

Hypothesis 2 *In treatment HUMAN,*

(i) more points are transferred in a manual PAS, and

(ii) fewer points are transferred in an automated PAS

than in treatment SYSTEM.

5. Results

We ran all sessions in August and September 2018 on Amazon MTurk using workers within the United States of America. The workers had to have completed at least 100 so-called

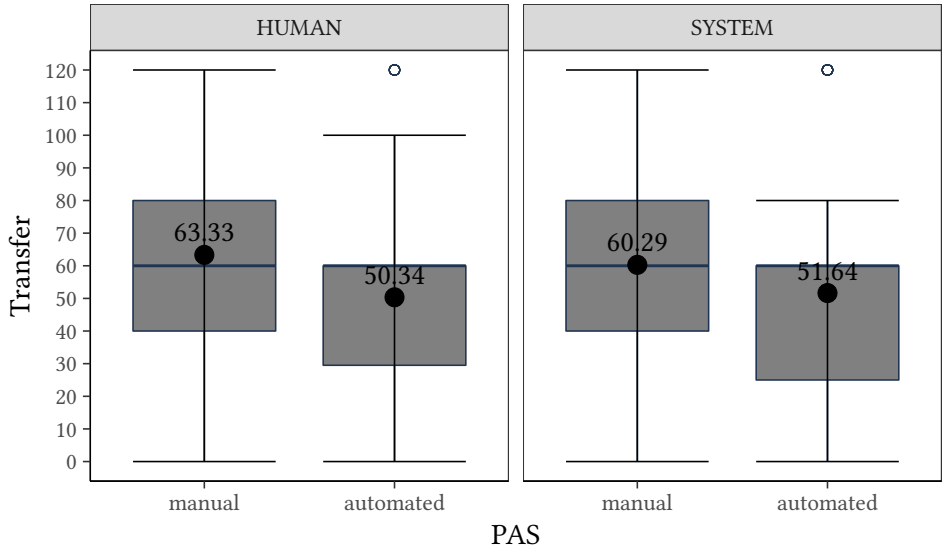
Human Intelligence Tasks (HITs) on Amazon MTurk and had to have an approval rate of 99% for their completed HITs to be able to take part in the experiment. Overall, 520 participants (44.4% female) participated in the study: 259 participants (42% female) in treatment *HUMAN* (131 agents and 128 principals), and 261 participants (46.7% female) in treatment *SYSTEM* (135 agents and 126 principals).[5] The participants were on average 37 years old. The study took about 10 minutes and the participants earned on average $1.8.

We used a between-subjects design for the treatments, hence the data for all statistical tests are independent for the different treatments. For the different *PAS*s we used a within-subjects design, i.e. agents were asked about their transfer in a *manual PAS* as well as in an *automated PAS* using a strategy method.

We first analyze the number of points transferred by the agent within each treatment before comparing the transferred points in the *manual PAS* and the *automated PAS* between treatment *HUMAN* and treatment *SYSTEM*. We also present further results on the agent's expectations about receiving a bonus and the threshold set by the principal.[6]

5.1. Hypothesis 1: *manual PAS* vs. *automated PAS*

According to Hypothesis 1, the agents should transfer more points to the principals under a *manual PAS* than under an *automated PAS* in both treatments.



Filled dots represent means, lines represent medians.

Figure 1: Box-and-whisker plots for the transferred points.

[5]The number of agents differs from the number of principals as some principals left the experiment before setting a threshold. In this case, the experimenter granted the bonus to the remaining agents independent of their productive activity.

[6]We present an analysis of the principals' behavior in Appendix A.4.

	HUMAN	*SYSTEM*
manual PAS - automated PAS	$\Delta = 12.99$	$\Delta = 8.65$
	(p =0.0000)	(p =0.0000)

The table shows differences between the *PAS*s ($\Delta = \ldots$) and p-values for a one-sided Wilcoxon signed-rank test. The tests resport no problem about the frequency of ties.

Table 2: Differences in the agents' transferred points between *PAS*s.

Indeed, as Figure 1 shows the agents transfer on average more to the principals in the *manual PAS* than in the *automated PAS* in both treatments.[7]

Table 2 shows the difference in the mean number of transferred points between the *manual PAS* and the *automated PAS* and provides the corresponding p-values for whether the means differ significantly within the treatments. The table confirms that the agents transfer significantly more to the principal in the *manual PAS* than in the *automated PAS* in both treatments.

Hence, we can confirm Hypothesis 1.(i) and Hypothesis 1.(ii), i.e. that more points are transferred in a *manual PAS* than in an *automated PAS* for treatment *HUMAN* as well as for treatment *SYSTEM*.

5.2. Hypothesis 2: treatment *HUMAN* vs. treatment *SYSTEM*

According to Hypothesis 2.(i), the agents should transfer more points in the *manual PAS* and, according to Hypothesis 2.(ii), the agents should transfer fewer points in the *automated PAS* in treatment *HUMAN* than in treatment *SYSTEM*. In fact, this is what we see in Figure 1.

Table 3 provides the difference in the mean number of transferred points for both *PAS*s and p-values for whether the means differ significantly between the treatments. Indeed, agents in treatment *HUMAN* transfer on average more points in the *manual PAS* and fewer points in the *automated PAS* than agents in treatment *SYSTEM*. However, the difference is not statistically significant in either the *manual PAS* or the *automated PAS*. Hence, we cannot confirm Hypothesis 2.(i) or Hypothesis 2.(ii).

	HUMAN - SYSTEM
manual PAS	$\Delta = 3.04$
	($p = 0.1358$)
automated PAS	Δ = -1.3
	(p =0.4819)

The table shows differences between the *PAS*s ($\Delta = \ldots$) and p-values for a one-sided Wilcoxon rank-sum test. The tests resport no problem about the frequency of ties.

Table 3: Differences in the transferred points in *manual PAS* and automated PAS between the treatments.

[7] We present an analysis of the frequency of the agents who choose fewer, more or the same in an *automated PAS* than in a *manual PAS* in Table 6 in Appendix A.3.

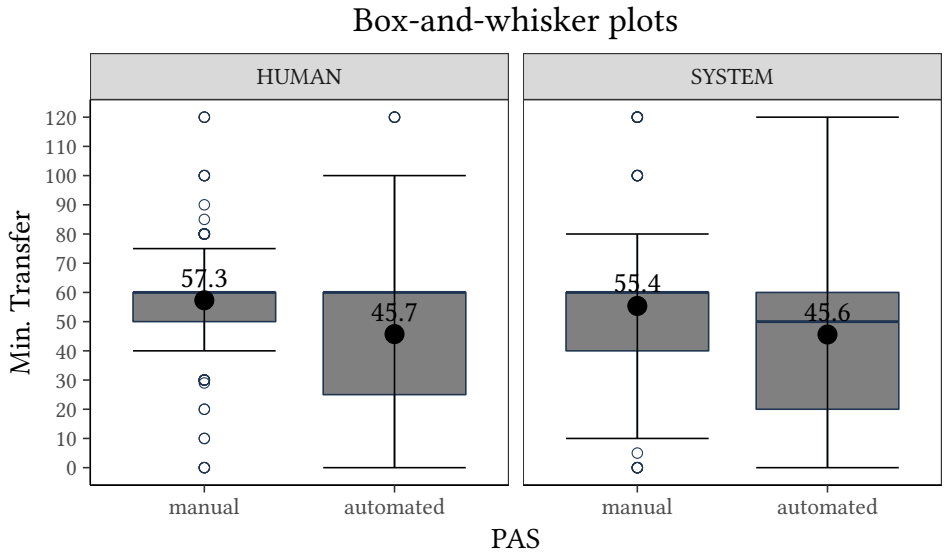
	HUMAN	SYSTEM
Strongly agree	9.20	17.20
Agree	77.30	65.60
Disagree	10.10	15.60
Strongly disagree	3.40	1.60

See Question 5 from Appendix A.2.

Table 4: Agents' beliefs about receiving a bonus [%].

5.3. Agents' expectations

We asked the agents about their expectations for getting a bonus. As Table 4 shows, most agents expected to get a bonus but agents in treatment *SYSTEM* tended to be overall less optimistic about receiving a bonus than agents in treatment *HUMAN*.



Filled dots represent means, lines represent medians.
See Question 3 and 4 from Appendix A.2.

Figure 2: Box-and-whisker plots for agents' beliefs about the threshold set by the principal.

Furthermore, as Figure 2 shows and Table 5 confirms, agents in both treatments expect the threshold to be significantly higher in the *manual PAS*, i.e. when the principal knows the transferred amount, than in the *automated PAS*, i.e. when the principal does not know the transferred amount.

	HUMAN	*SYSTEM*
manual PAS - automated PAS	$\Delta = 11.57$	$\Delta = 9.78$
	(p =0.0000)	(p =0.0001)

The table shows differences between the *PAS*s ($\Delta = \ldots$) and p-values for a two-sided Wilcoxon signed-rank test where this difference could be zero.

Table 5: Differences in the agents expectations about the threshold between *PAS*s.

6. Conclusion

The experiment studies whether automation leads to a decrease in employees' performance. For this purpose, we set up a Principal-Agent experiment in a real job market (Amazon MTurk) and compared performance under an *automated PAS* and a *manual PAS*. Furthermore, we are investigating whether it matters if the company (treatment *SYSTEM*) or the direct supervisor (treatment *HUMAN*) decides on what *PAS* to use.

We find that the agents' performance is significantly lower under an automated PAS than under a *manual PAS* in both treatments. Thus, we observe hidden costs of automation in the form of lower performance in an *automated PAS* than under a *manual PAS*. Under the assumption that employees do care about who decided which *PAS* to use, the performance in both *PAS*s should differ whether the principal or system (e.g. company) decided. We find, however, no significant difference in the performance if the principal decided to use the *manual PAS* [*automated PAS*] compared to when the system decided to use the *manual PAS* [*automated PAS*]. Hence, the costs of automation seem to be independent of whether the principal or the company decides to use an *automated PAS*.

A possible explanation for why the agents' performance is lower under an *automated PAS* than under a *manual PAS* might be that they perceive the *automated PAS* mechanism to be more rigid than the *manual PAS* mechanism, as the principal has to commit to a threshold before knowing the transferred amount. Therefore, an *automated PAS* might have a lower motivational effect on the agent than a *manual PAS*, where the principal does not have to commit to the threshold ex-ante.[8] Furthermore, performance under an *automated PAS* might be lower than under a *manual PAS* as the agents' expectations about the threshold differ. Indeed, this is what we found. Agents expect the threshold set by the principal to be lower in an *automated PAS* than in a *manual PAS*. Thus, agents might adjust to their lower expectation about the threshold in the *automated PAS* by showing more lenient effort provision. Principals, however, set more or less the same thresholds in a *manual PAS* and in an *automated PAS*. The discrepancy between the agents' expectations and the principals' actual behavior leaves some room for further exploration.

We are aware that the difference between a human decision and a decision made by a system is quite subtle in an online experiment as all interaction takes place via a computer interface. Thus, the divergence between them might not have been visible enough to the participant and this might explain why we did not find a difference in the performance whether

[8] A more or less rigid *PAS* can also be seen as a more or less rigid contracts. Therefore, your paper also offers a contribution to the economic literature on incomplete contracts.

the principal or the system decided about the *PAS*. Of course, this is only speculation and further research must be conducted to prove that claim.

In conclusion, our results show that the use of algorithms entails hidden costs that should be considered when replacing human-human interaction by automated processes. Besides the tremendous benefits automation generates, it might also have some downsides. When replacing human-human interactions in particular, they are not negligible. Superiors in charge of implementing an automated system which replaces a human-human interaction may benefit from communicating that the underlying parameters and demands of the automated system do not differ from the demands set by humans. In other words, open communication about the settings of an automated system might help to avoid a decrease in employees' performance in the course of automation.

References

Andreoni, J. and Bernheim, B. D. (2009). Social image and the 50-50 norm: A theoretical and experimental analysis of audience effects. *Econometrica*, 77(5):1607–1636.

Beauvois, J.-L. and Joule, R. (1996). *A radical dissonance theory*. Taylor & Francis, London,GB.

Bénabou, R. and Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review*, 96(5):1652–1678.

Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, 10(1):122–142.

Bolton, G. E. and Ockenfels, A. (2000). Erc: A theory of equity, reciprocity, and competition. *American Economic Review*, 90(1):166–193.

Charness, G., Gneezy, U., and Henderson, A. (2018). Experimental methods: Measuring effort in economics experiments. *Journal of Economic Behavior & Organization*, 149(3):74–87.

Charness, G. and Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, 117(3):817–869.

Chen, D. L., Schonger, M., and Wickens, C. (2016). otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97.

Corgnet, B., Hernán-Gonzalez, R., and Mateo, R. (2019). Rac(g)e against the machine? social incentives when humans meet robots.

Dufwenberg, M. and Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior*, 47(2):268–298.

Ellingsen, T. and Johannesson, M. (2005). Trust as an incentive.

Falk, A. and Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior*, 54(2):293–315.

Falk, A. and Kosfeld, M. (2006). The hidden costs of control. *American Economic Review*, 96(5):1611–1630.

Fehr, E., Kirchsteiger, G., and Riedl, A. (1993). Does fairness prevent market clearing? an experimental investigation. *The Quarterly Journal of Economics*, 108(2):437–459.

Fehr, E., Klein, A., and Schmidt, K. M. (2007). Fairness and contract design. *Econometrica*, 75(1):121–154.

Fehr, E. and List, J. A. (2004). The hidden costs and returns of incentives—trust and trustworthiness among ceos. *Journal of the European Economic Association*, 2(5):743–771.

Fehr, E. and Rockenbach, B. (2003). Detrimental effects of sanctions on human altruism. *Nature*, 422(6928):137–140.

Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3):817–868.

Fehr, E. and Schmidt, K. M. (2007). Adding a stick to the carrot? the interaction of bonuses and fines. *American Economic Review*, 97(2):177–181.

Hsee, C. K. and Weber, E. U. (1997). A fundamental prediction error: Self–others discrepancies in risk preference. *Journal of Experimental Psychology*, 126(1):45–53.

Jenni, K. and Loewenstein, G. (1997). Explaining the identifiable victim effect. *Journal of Risk and Uncertainty*, 14(3):235–257.

Kajackaite, A. and Werner, P. (2015). The incentive effects of performance requirements – a real effort experiment. *Journal of Economic Psychology*, 49:84–94.

Kampkötter, P. (2017). Performance appraisals and job satisfaction. *The International Journal of Human Resource Management*, 28(5):750–774.

Kaur, N. and Sood, S. K. (2017). A game theoretic approach for an iot-based automated employee performance evaluation. *IEEE Systems Journal*, 11(3):1385–1394.

Kogut, T. and Ritov, I. (2005). The "identified victim" effect: An identified group, or just a single individual? *Journal of Behavioral Decision Making*, 18(3):157–167.

Levy, P. E. and Williams, J. R. (2004). The social context of performance appraisal: A review and framework for the future. *Journal of Management*, 30(6):881–905.

Ockenfels, A., Sliwka, D., and Werner, P. (2015). Bonus payments and reference point violations. *Management Science*, 61(7):1496–1513.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American Economic Review*, 83(5):1281–1302.

Rabin, M. (1995). Moral preferences, moral constraints, and self-serving biases. *Department of Economics UCB (unpublished manuscript)*.

Small, D. A. and Loewenstein, G. (2003). Helping a victim or helping the victim: Altruism and identifiability. *Journal of Risk and Uncertainty*, 26(1):5–16.

Stone, H. S. (1971). *Introduction to computer organization and data structures*. McGraw-Hill, New York, NY.

A. Appendix

This section contains additional information on the interfaces and questions used in the treatments. We also present further analyses of the data we collected in addition to the data used to test your hypotheses. Data and methods are available online.

A.1. Instructions

This HIT *[Human Intelligence Task]* is an economic experiment. Please read the following instructions carefully. The instructions provide you with all the information required for participating in the experiment. You will receive $0.50 USD for participating in the experiment (paid only if you finish the experiment). Your final payoff is the $0.50 USD for participating in the experiment plus the amount earned during the experiment. You will earn at least the $0.50 USD for participating in the experiment. In the experiment, the currency used is points. Your points will be converted to USD at the end of the experiment using a conversion rate of **1 point = $0.01 USD**.

General setup
In this experiment, you are matched with another human participant. You will play in a group of two. All decisions are made anonymously. No participant knows with whom (s)he is matched. During the experiment, the members of the group are called ”participant A” and ”participant B”. The roles are randomly assigned.

The experiment
Participant A starts with 120 points at the beginning of the experiment. Participant B starts with no points. Each participant has to make a decision during the experiment. The decisions are explained below. Please read the explanations for both participants as both decisions will affect the number of points you will earn.

Participant A's decision:
Participant A has to decide how many points (s)he wants to transfer to participant B. The points transferred to participant B are doubled by the experimenter, meaning each point transferred to participant B reduces the points of participant A by one point but increases the points of participant B by two points.

Participant B's decision:
Before participant B knows what participant A transferred, participant B *[the system]* selects an approach. The two possible approaches (Approach BLUE or Approach GREEN) are explained below.

In each approach participant B has to decide if participant A should be given a **bonus of 120 points** for a **”very good transfer”**. The bonus is paid by the experimenter and does not reduce the points of participant B. The difference between the two approaches is how participant B determines the minimum amount (threshold) that participant A has to transfer to get a bonus.

In Approach BLUE, participant B **knows** the amount transferred by participant A when determining the threshold. The decision screen will look like this:

You *[the system]* decided to use Approach BLUE. **Participant A has transferred X of 120 points to you.** If participant A has transferred at least the threshold amount (s)he gets a bonus of 120 points (paid by the experimenter). Please indicate your threshold here: ……Points

In Approach GREEN, participant B **DOES NOT know** the amount transferred by participant A when determining the threshold. The decision screen will look like this:

You *[the system]* decided to use Approach GREEN. If participant A has transferred at least the threshold amount (s)he gets a bonus of 120 points (paid by the experimenter). Please indicate your threshold here: ……Points

Further note:
Participant A has different fields to enter amounts in case Approach BLUE or Approach GREEN is used.
Some examples:

- ***Example 1:*** Participant A transfers 0 points to participant B. Participant A will have 120 points (120 - 0) plus eventually a bonus of 120 points. Participant B will have 0 points (0 x 2). In addition, both participants receive $0.50 USD for participating.
- ***Example 2:*** Participant A transfers 40 points to participant B. Participant A will have 80 points (120 - 40) plus eventually a bonus of 120 points. Participant B will have 80 points (40 x 2). In addition, both participants receive $0.50 USD for participating.
- ***Example 3:*** Participant A transfers 80 points to participant B. Participant A will have 40 points (120 - 80) plus eventually a bonus of 120 points. Participant B will have 160 points (80 x 2). In addition, both participants receive $0.50 USD for participating.
- ***Example 4:*** Participant A transfers 120 points to participant B. Participant A will have 0 points (120 - 120) plus eventually a bonus of 120 points. Participant B will have 240 points (120 x 2). In addition, both participants receive $0.50 USD for participating.

Before clicking "Next" please make sure you have read and understood the instructions. After clicking "Next" we will match you with the next person starting the experiment. This might take some time.

A.2. Questions

All participants were asked to complete a questionnaire. The questions were asked right after the decision and before the final outcome was announced. The answer method used is presented in brackets. Apart from the first four questions, which were only presented to agents, all questions were asked to agents and principals.

1. Why did you choose to transfer the amount you have chosen to participant B in Approach BLUE (participant B **knows** how much you transferred)? [Open Question] (For the answers given see online data-set)

2. Why did you choose to transfer the amount you have chosen to participant B in Approach GREEN (participant B **does not know** how much you transferred)? [Open Question] (For the answers given see online data-set)

3. What do you think is the minimum amount you would have had to transfer to get the bonus if participant B decided to use Approach B (participant B **knows** how much you transferred)?"[Integer from 0 to 120 points] (For an analysis of the answers given see Section 5.1)

4. What do you think is the minimum amount you would have had to transfer to get the bonus if participant B decided to use Approach GREEN (participant B **does not know** how much you transferred)?[Integer from 0 to 120 points] (For an analysis of the answers given see Section 5.1)

5. How much do you agree with this statement: 'I think that I will get a bonus.'? [*"Strongly disagree"; "Disagree", "Agree"; "Strongly agree"*] (For an analysis of the answers given see Appendix 5.3)

6. Do you consider the procedure to get the bonus to be fair?[*"YES"; "NO"*] (For an analysis of the answers given see Appendix A.5)

7. How do you see yourself: Are you a person who is willing to take risks or do you try to avoid taking risks? Please select a number on a scale from 0 to 10. The value 0 means: 'not at all willing to take risks' and the value 10 means: 'very willing to take risks'. [scale 0 to 10] (For an analysis of the answers given see Appendix A.6)

8. Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with other people? Please select a number on a scale from 0 to 10. The value 0 means: 'can't be too careful' and the value 10 means: 'most people can be trusted'.[Scale 0 to 10] (For an analysis of the answers given see Appendix A.6)

9. What is your gender?[*"YES"; "NO"; "OTHER"*] (For an analysis of the answers given see Section 5)

10. What is your age [in years]?[Years] (For an analysis of the answers given see Section 5)

A.3. Relative frequency of the agents' transfer decision

	HUMAN	SYSTEM
less	42.70	30.40
more	6.90	6.70
same	50.40	63.00

The table shows the percentage of agents transferring the same, more, or less in an *automated PAS* than in a *manual PAS* by treatment.

Table 6: Agents' transfer decisions [%].

Table 6 shows the relative frequency of agents who transferred less, more or the same in an *automated PAS* than in a *manual PAS*. The table reveals that around half of the agents transferred the same in a *manual PAS* as in an *automated PAS* in both treatments. Nevertheless, around 40% of the agents transferred fewer points in an *automated PAS* than in a *manual PAS* in treatment *HUMAN* and slightly less than one-third of the agents did so in treatment *SYSTEM*.

A.4. Analysis of the principals' behavior

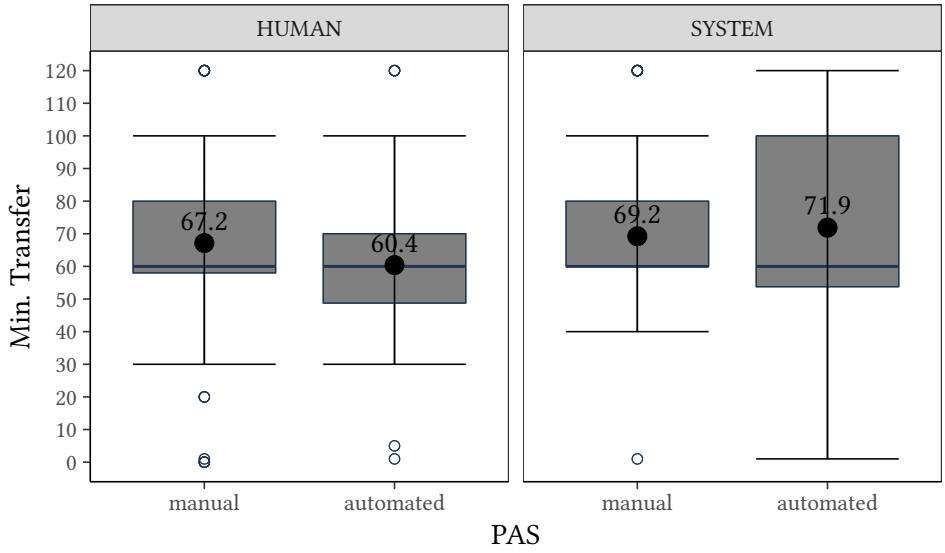
As Table 7 shows, the vast majority of the principals decided to use a *manual PAS* instead of an *automated PAS* in treatment *HUMAN*, where the principals were able to choose.

	HUMAN	SYSTEM
manual	78.10	49.20
automated	21.90	50.80

The table shows the percentage of principals choosing a *manual PAS* or *automated PAS*.

Table 7: *PAS* choices by principals and the system[%].

As suggested by Figure 3 and confirmed by Table 8, the threshold set by the principals in the *automated PAS* does not differ significantly from the threshold set in the *manual PAS* in both treatments.



Filled dots represent means, lines represent medians.

Figure 3: Box-and-whisker plots for thresholds set by principals.

	HUMAN	*SYSTEM*
manual PAS - automated PAS	Δ = 6.7571429	Δ = -2.6491935
	(p =0.1953)	(p =0.8511)

The table shows differences between the *PAS*s (Δ = ...) and p-values for a two-sided Wilcoxon rank-sum test where this difference could be zero.

Table 8: Differences in the principals' threshold set between *PAS*s.

As Table 9 shows, the threshold does not differ significantly between both treatments.

	HUMAN - SYSTEM
manual PAS	Δ = -2.0758065
	(p = 0.9565)
automated PAS	Δ = -11.4821429
	(p =0.2283)

The table shows differences between the *PAS*s (Δ = ...) and p-values for a two-sided Wilcoxon rank-sum test.

Table 9: Differences in the principals' threshold set in *manual PAS* and *automated PAS* between the treatments.

The thresholds set in *automated PAS*s in treatment *SYSTEM*, however, are more dispersed than in the other conditions. In summary, principals in treatment *HUMAN*, who decided on their own which approach to use, do not set a significantly different threshold than participants in treatment *SYSTEM*, where the system decided randomly which approach to use.

A.5. Perceived fairness of the procedure

We asked the participants if they perceive the procedure to result in a fair (see Question 6). In treatment *HUMAN*, 84.87% of the participants perceived the procedure to be fair. In treatment *SYSTEM*, the procedure was perceived as fair by 80.33% of the participants. As Table 10 shows, the assessment by the agents and the principals hardly differs.

	Agent	Principal
HUMAN	84.87	83.59
SYSTEM	80.33	83.33

Table 10: Assessment of the fairness of the procedure [%].

A.6. Participants' propensity for risk and trust

All participants were asked if they are a person who is willing to take risks or tries to avoid taking risks (see Question 7) and if they would say that most people can be trusted or that you cannot be too careful in dealing with other people (see Question 8). Willingness to take risks was measured by a continuous scale from 'not at all willing to take risks' (0) to 'very willing to take risks' (10). The level of trust was measured by a continuous scale from 'can't be too careful' (0) to 'most people can be trusted' (10). As Table 11 shows, agents and principals were slightly risk averse and somewhat concerned about the trustworthiness of other people.

	Agent	Principal
Risk	$\varnothing = 4.36$	$\varnothing = 4.53$
	(2.58)	(2.3)
Trust	$\varnothing = 4.9$	$\varnothing = 5.04$
	(2.55)	(2.41)

The table shows the means for risk and trust ($\varnothing = \ldots$) and the corresponding standard deviations (in brackets).

Table 11: Mean and standard deviation for levels of risk and trust.