Exploring condition in which people accept AI over human judgements on justified defection

Supplementary Information

Hitoshi Yamamoto[1*], Takahisa Suzuki[2]

1: Department of Business Administration, Rissho University, JAPAN

2: College of Policy Studies, Tsuda University, JAPAN

The Material of Study 1

Participants read the following two scenarios. In both scenarios, Bob defected against Alice, who has bad reputation.

> Consider that you are working in a restaurant.
>
> Alice and Bob are co-workers. In this restaurant, employees are assigned to work the night shift. Using a work management web application, each has a habit of asking other employees to take their place on the night shift when they are not available. In addition, bonuses and other assessments are made based on an employee's work attitude and how much they contribute to the restaurant's operation.
>
> Alice is not serious about her work and often takes time off from the night shift for various reasons. She often takes time off for selfish reasons and is not liked by her colleagues, including you. One day, Alice messaged Bob, asking him to cover for her on the night shift because she wanted to go to a concert of her favourite singer.
>
> Although Bob had plenty of time, he declined Alice's request.

After reading the scenario, participants rated how they assessed Bob's behaviour from three viewpoints using a 5-point scale (5: Agree, 4: Somewhat agree, 3: Neither agree nor disagree, 2: Somewhat disagree, 1: Disagree.): "Bob is a reliable person", "Do you like Bob?", and "Do you feel sympathetic toward Bob?" The evaluation scores for the donor's action were obtained by simply adding together the scores of these three statements. When the three items are summed, the total ranges from 3 to 15. The values were then normalised to fit within the range of 0 to 1 for readability. Normalisation makes it easier to understand that 0.0 means completely disagree, 1.0

means completely agree, and 0.5 means neutral. The normalisation procedure is identical to that used in the following experiments.

Then, participants read the second scenario as shown below. In this scenario, a manager or AI evaluated Bob's justified defection as good (or bad). Then participants evaluated the manager's (or AI's) judgment. We used a 2×2 between-subjects design (AI or manager, good evaluation or bad evaluation). Participants were assigned one of the four scenarios randomly.

> Continue to assume you are working in the same restaurant.
>
> Alice and Bob are co-workers. In this restaurant, employees are assigned to work the night shift. Using a work management web application, each has a habit of asking other employees to take their place on the night shift when they are not available. In addition, bonuses and other assessments are made based on an employee's work attitude and how much they contribute to the restaurant's operation.
>
> **[AI scenario]** The assessment is performed automatically by an artificial intelligence (AI) equipped with advanced algorithms. The AI is highly regarded for its ability to make fair decisions based on vast data.
>
> **[Manager scenario]** The assessment is performed by the manager in charge of the restaurant. This manager has been with the company for many years and is recognized as a fair judge.
>
> Alice is not serious about her work and often takes time off from the night shift for various reasons. She often takes time off for selfish reasons and is not liked by her colleagues, including you. One day, Alice messaged Bob, asking him to cover for her on the night shift because she wanted to go to a concert of her favourite singer.
>
> Although Bob had plenty of time, he declined Alice's request.
>
> After seeing this behaviour, the AI (or manager) judged Bob's behaviour as good (or bad) and raised (or lowered) his assessment by one point.

After reading the scenario, participants rated how they assessed AI's (or manager's) behaviour from five viewpoints using a 5-point scale(5: Agree, 4: Somewhat agree, 3: Neither agree nor disagree, 2: Somewhat disagree, 1: Disagree.): "I agree with the AI's decision", "The AI's

judgment is reasonable", "I trust the AI's judgment", "I empathize with the AI's decisions" and "The AI's judgment is wrong".

The Material of Study 2

The participants read two scenarios. This first scenario is same as that in Study 1. The second scenario is similar to the second scenario in Study 1. The difference is that there are two evaluators (AI and manager), each evaluating Bob's behaviour independently. We used a between-subjects design for the four cases below.

> Continue to assume you are working in the same restaurant.
>
> Alice and Bob are co-workers. In this restaurant, employees are assigned to work the night shift. Using a work management web application, each has a habit of asking other employees to take their place on the night shift when they are not available. In addition, bonuses and other assessments are made based on an employee's work attitude and how much they contribute to the restaurant's operation.
>
> **[Cases 0 and 1]** The assessment is performed by an artificial intelligence (AI) equipped with advanced algorithms and a manager in charge of the restaurant. The AI is highly regarded for its ability to make fair decisions based on vast data. The manager has been with the company for many years and is recognized as a fair judge.
>
> **[Case 2]** The assessment is performed by two managers in charge of the restaurant. Both managers have been with the company for many years and are recognized as fair judges.
>
> **[Case 3]** The assessment is performed automatically by two independent artificial intelligences (AIs) equipped with advanced algorithms. Both AIs are highly regarded for their ability to make fair decisions based on vast data.
>
> Alice is not serious about her work and often takes time off from the night shift for various reasons. She often takes time off for selfish reasons and is not liked by her colleagues, including you. One day, Alice messaged Bob, asking him to cover for her on the night shift because she wanted to

go to a concert of her favourite singer.

Although Bob had plenty of time, he declined Alice's request.

[Case 0] After observing Bob's behaviour, the AI decided that 'not helping someone who is not cooperating with other employees is a good thing for the workplace as a whole' and raised Bob's assessment by one point. On the other hand, the manager decided that 'not helping someone who is not cooperating with other employees is not good for the workplace as a whole' and reduced Bob's assessment by one point.

[Case 1] After observing Bob's behaviour, the manager decided that 'not helping someone who is not cooperating with other employees is a good thing for the workplace as a whole' and raised Bob's assessment by one point. On the other hand, the AI decided that 'not helping someone who is not cooperating with other employees is not good for the workplace as a whole' and reduced Bob's assessment by one point.

[Case 2] After observing Bob's behaviour, one manager decided that 'not helping someone who is not cooperating with other employees is a good thing for the workplace as a whole' and raised Bob's assessment by one point. On the other hand, another manager decided that 'not helping someone who is not cooperating with other employees is not good for the workplace as a whole' and reduced Bob's assessment by one point.

[Case 3] After observing Bob's behaviour, one AI decided that 'not helping someone who is not cooperating with other employees is a good thing for the workplace as a whole' and raised Bob's assessment by one point. On the other hand, another AI decided that 'not helping someone who is not cooperating with other employees is not good for the workplace as a whole' and reduced Bob's assessment by one point.

After reading the scenario, participants rated which of the two evaluator's judgments is more acceptable from eight viewpoints using a 5-point scale: (5:agree with "good" judgement, 4: somewhat agree with "good" judgement, 3: neither agree with "good" nor "bad" judgment, 2: somewhat agree with "bad" judgement, 1: agree with "bad" judgement). “Which decision do you agree with?”, “Which judgment do you trust?”, “Which judgment do you empathise with?”,

“Which decision do you think is more appropriate?”, “Which decision do you think is better for overall productivity in the workplace?”, “Which decision do you think is better for employee motivation?”, “Which decision do you think is better for employee relations?”, and “Overall, which decision do you think is better?”.

Descriptive statistics for the experiment

Study 1

Table S1: Descriptive statistics of the evaluations and acceptances for justified defection. Numbers in each cell are mean (sd).

		Evaluator		
		AI	Human	all
Evaluation	bad	0.42(0.27)	0.42(0.24)	0.42(0.26)
	good	0.50(0.26)	0.48(0.27)	0.49(0.26)
	all	0.46(0.27)	0.45(0.26)	0.45(0.26)

The evaluation of the donor by participants is mean=0.49, sd=0.20

Study 2

Table S2: Descriptive statistics of the evaluations and acceptances for justified defection.

	mean	sd
Case 0: AI Good and Human Bad	0.65	0.24
Case 1: AI Bad and Human Good	0.53	0.27
Case 2: Human Good and Human Bad	0.53	0.23
Case 3: AI Good and AI Bad	0.51	0.25
Evaluation by subjects	0.49	0.23

We conducted multiple comparisons using the Holm method, revealing significant differences between the following pairs: case 0 and case 1 ($p = 0.003$), case 0 and case 2 ($p = 0.004$), and case 0 and case 3 ($p < 0.001$).