Current Psychology
https://doi.org/10.1007/s12144-021-02559-w

When people are defeated by artificial intelligence in a competition task requiring logical thinking, how do they make causal attribution?

Ryosuke Yokoi[1,2] · **Kazuya Nakayachi**[1]

Accepted: 21 November 2021

Abstract
Given that artificial intelligence (AI) has been predicted to eventually take on human tasks demanding logical thinking, it makes sense that we should examine psychological responses of humans when their performance is inferior to AI. Research has demonstrated that after people fail a task, whether they reorient their behavior towards success depends on what they attribute the failure to. This study investigated the causal attributions people made in a competition task requiring such thinking. We also recorded whether they wanted to re-challenge the games after they were defeated by AI. Experiments 1 ($N=74$) and 2 ($N=788$) recruited Japanese participants, while Experiment 3 ($N=500$) comprised American participants. There were two conditions: in the first, participants competed against an AI opponent and in the other, they believed they were competing against a human. The results of the three experiments showed that participants attributed the loss to their own and their opponent's abilities more than any other factor, irrespective of the opponent type. The number of participants choosing to re-challenge the game did not differ significantly between the AI and human conditions in Experiments 1 and 3, although the number was lower in the AI condition than in the human condition in Experiment 2. Besides providing fresh insight on how people make causal attributions when competing against AI, our findings also predict how people will respond after their jobs are replaced by AI.

Keywords Artificial intelligence · Causal attribution · Behavioral response · Competition game · Self-effacing bias

The technology of autonomous systems, in particular artificial intelligence (AI), has been applied to several domains, such as transportation (Waldrop, 2015), medical diagnosis (Topol, 2019), and military uses (Dawes, 2017), and it is expected that the development of AI will be increasingly promoted. In the future, AI will have the ability to undertake jobs that have traditionally been performed by humans (Acemoglu & Restrepo, 2018; Huang & Rust, 2018). Frey and Osborne (2017) analyzed 702 jobs and estimated the probability of their being computerized; their findings revealed that professionals such as telemarketers, tailors, and mathematical technicians are likely to be replaced by AI. In a reported case of the performance of AI being superior to that of human experts, Topol (2019) found that medical AI can diagnose a disease more accurately and rapidly than human doctors. AI are growing increasingly more proficient at completing tasks that demand complex analysis and logical thinking, which humans can find difficult (Huang & Rust, 2018). As AI becomes increasingly prevalent in our society, research must examine the psychological responses of humans when their performance is inferior to that of AI. Therefore, this study focuses on human-AI interaction in a task requiring logical thinking and investigates how people make causal attributions after losing a game against an AI opponent.

The games used in our experiments required people to think logically. We did not examine a scenario where people's jobs are replaced by AI, as such a situation would involve various extraneous variables, such as types of jobs and individuals' past achievements that may influence causal attribution. The pattern of causal attribution was tested under two situations—where participants were defeated by AI and where they were defeated by a human—to examine whether there was a specific causal attribution when they

✉ Ryosuke Yokoi
cykd1004@mail2.doshisha.ac.jp

[1] Faculty of Psychology, Graduate School of Psychology, Doshisha University, 1-3, TataraMiyakodani, Kyotanabe-shi, Kyoto 610-0394, Japan

[2] Japan Society for the Promotion of Science Research Fellow, Tokyo, Japan

Published online: 14 January 2022

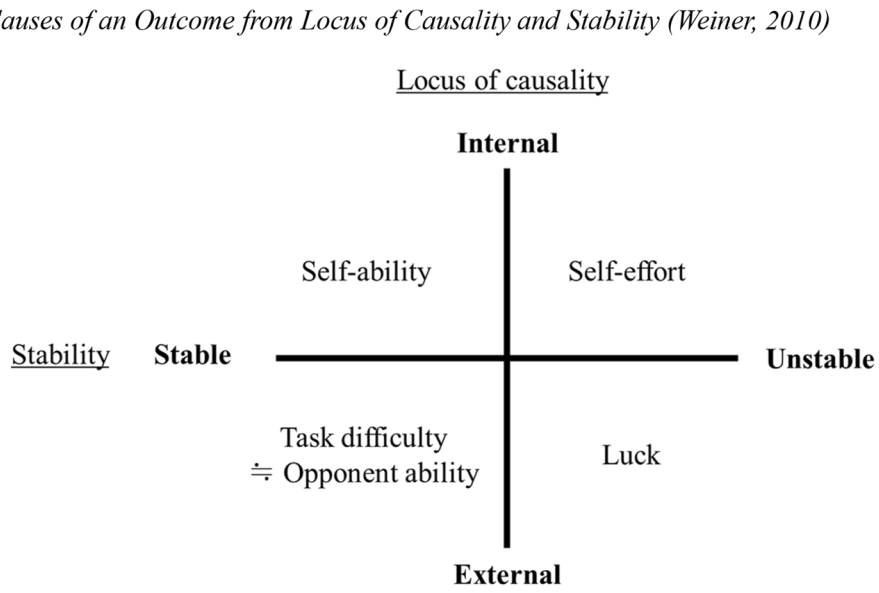

Fig. 1 Main Causes of an Outcome from Locus of Causality and Stability (Weiner, 2010). As our research used a competition task, opponent ability was adopted as an external and stable cause, rather than task difficulty

lost the game against AI. To the best of our knowledge, this is the first study to examine people's causal attribution in a situation where they compete against AI.

Causal Attribution

According to attribution theory, causal attribution is defined as the consideration of why a consequence arises (Weiner, 1979, 1985, 2010). Heider (1978), who discussed causal attribution in the 1950s, claimed that people interpret the relationship between an individual's action and the environment and attribute events to individual and environmental factors. Causal attribution was gradually adopted in the research on motivations in educational situations (Weiner, 1979), and most of this research focused on the causal attribution related to the success and failure of an individual's achievements and examined the relationship between causal attribution and motivation (Weiner, 1979).

Weiner (2010) proposed two dimensions of the perceived causes of success and failure: the locus of causality and stability. The locus of causality determines whether an outcome arises from internal or external causes, while stability refers to the extent to which the causes are temporally persistent. Weiner (2010) also described four representative causes of individual achievement based on these two dimensions (Fig. 1). Self-effort is internal and changeable depending on the circumstances, for example. Since our research used a task in which participants competed against AI, opponent ability was adopted instead of task difficulty as an external and stable cause. Although none of the literature has previously stated that opponent ability is an external and stable cause, in this case, the opponent is external and its ability is stable; therefore, opponent ability can be regarded as an external and stable factor. This study adopted four causes, including opponent ability, and investigated the pattern of causal attribution during a competition against AI. Although Weiner (1979) added a third dimension called controllability to his theory, thus proposing a total of eight attributed factors, as the present study is a novel attempt, we tried to keep our experimental design simple and adopted the earlier model of the attribution theory focusing on the four factors.

Causal attribution affects different variables, such as feelings, future behaviors, and expectations of success (Kelley & Michela, 1980). When people fail a task and attribute the consequence to stable factors, they may have a low expectation of success in the following task (McMahan, 1973; Weiner et al., 1976). Attribution to unstable and internal causes when a person fails can promote efforts for future success (Crittenden & Wiley, 1980; Mamonov & Koufaris, 2018; Rascle et al., 2015). It has been implied that attributional style is related to depression (Hymes & Akiyama, 1991; Sakamoto & Kambara, 1998), and explaining positive events due to internal factors can enhance subjective wellbeing (Titova & Sheldon, 2019). Thus, understanding how people make causal attributions when they are defeated by AI may be useful for predicting their feelings and future behaviors.

Self-Serving and Self-Effacing Attribution

In the typical pattern of causal attribution, people are more likely to attribute positive results to internal causes and negative results to external factors (Miller & Ross, 1975). This is called "self-serving attribution." Self-serving attribution is considered to protect self-esteem and decrease self-threat (Campbell & Sedikides, 1999; Zuckerman, 1979) and

tends to emerge when a task is important to an individual (Bradley, 1978). Research on causal attribution has observed self-serving attribution in various contexts, such as education (Wang & Hall, 2018), sports (Allen et al., 2020), business (Ford, 1985), and competition (De Michele et al., 1998; Grove et al., 1991; Polman et al., 2007). The tendency of attribution is robust and generalizable.

However, there are cultural differences in the patterns of causal attribution. Mezulis et al. (2004) reviewed studies on causal attribution and found that people in Western countries are more inclined to make self-serving attributions than those in Asia. They also demonstrated that the Japanese tend to make self-effacing, rather than self-serving, attributions; in other words, they attribute positive events to situational factors and luck, while attributing negative events to their own ability and effort. Many studies have examined the differences in causal attribution between Japanese and Western subjects (Chandler et al., 1981; Kashima & Triandis, 1986; Peterson et al., 2002; Yan & Gaier, 1994). For example, Kashima and Triandis (1986) found that when they failed a memory task, Japanese participants attributed the outcome to internal causes more than Americans. This Japanese causal attribution tendency is assumed to derive from their desire to gain positive evaluations from other people (Yoshida et al., 1982).

Research on Attribution during Human–Computer Interaction

Many studies on attribution have been conducted in the area of human–computer interaction, which have focused on how people make attributions of responsibility for robot and computer failures and who they blame for such failures (Awad et al., 2020; Belanche et al., 2020; Hong, 2020; Lee & Cranage, 2018; Leo & Huh, 2020; van der Woerdt & Haselager, 2019). Awad et al. (2020) investigated how attributions of responsibility were made for traffic accidents by human drivers and autonomous vehicles. They described various scenarios in which the primary driver of a shared-control car drove the vehicle and the secondary driver monitored the actions of the primary driver, with either a human or an autonomous system being allocated to one of the primary and secondary drivers. Their findings showed that, overall, the autonomous system was blamed less than the human driver.

Causal attribution has also been tested in situations where people either succeed or fail in tasks involving computers and robots (Brown et al., 2015; Hinds et al., 2004; Moon & Nass, 1998; Serenko, 2007). However, these studies have employed circumstances in which people cooperated with computers or where computers aided people's tasks. For example, Brown et al. (2015) examined people's causal attribution in circumstances where a navigation system helped a human driver. Their experiment manipulated the level of the system's autonomy and indicated that people attributed the driving performance to the system when the system's autonomy was high. Hence, although causal attribution has been investigated in tasks in which humans collaborate with computers and robots, to the best of our knowledge, no extant evidence has revealed the causal attributions people make when competing against AI.

Hypothesis and Rationale

We assume that the pattern of causal attribution differs depending on whether the opponent is human or AI. Madhavan and Wiegmann (2007) reviewed the literature on human–automation trust relationships and suggested that people have different expectations of the performance of automated systems to those they have of humans. The authors claimed that people expect automated systems to perform perfectly, while expecting humans to perform imperfectly. Many studies have indicated that automated systems are expected to perform at nearly perfect rates and make rational decisions (Dijkstra et al., 1998; Dzindolet et al., 2002; Lyons & Stokes, 2012; Sundar & Nass, 2000; Thurman et al., 2019). Salem et al. (2013) also revealed that when a humanoid robot's gestures were incongruent with speech, people evaluated the anthropomorphic nature of the robot highly. These findings imply that people expect automated systems to perform perfectly, whereas humans are perceived as agents who can fail. People may think that AI's ability strongly dictates the outcome of a competition against it, while they may also believe that luck is unrelated to the outcome. Given the above, we formulated the first hypothesis:

> H1: When people lose a game against AI, they are more likely to attribute the result to the opponent's ability than when they lose against a human. When losing against AI, people are less likely to attribute the result to luck than when they lose against a human.

Additionally, this study investigated people's behavioral responses when they were defeated by an AI or human opponent. Specifically, we measured whether participants want to compete against the same opponent again after experiencing a loss. H1 assumed that people would attribute losing a game to a stable cause when competing against AI and an unstable cause when competing against a human. Attribution to stable factors after a failure leads to low expectations of future success (McMahan, 1973; Weiner et al., 1976); thus, we hypothesized:

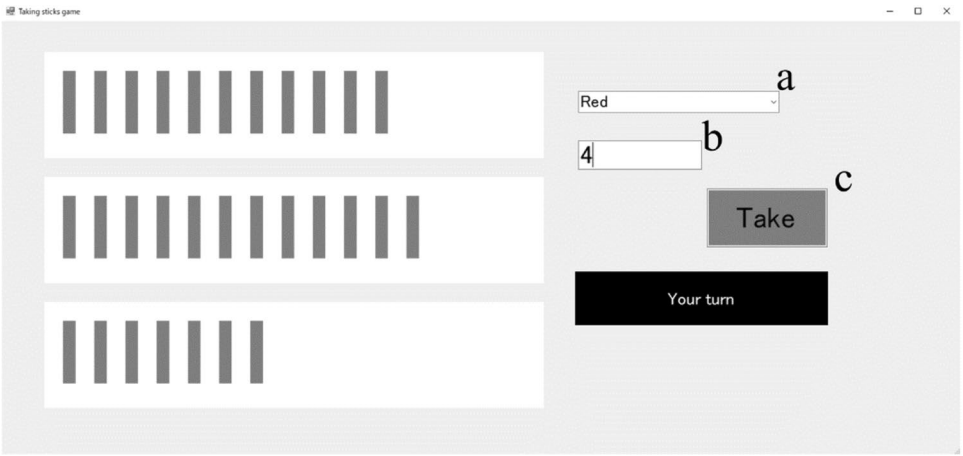

Fig. 2 Main screen of the taking sticks game. a indicates a drop-down list where participants selected the color of the stick they wanted to take. b is the text box in which participants input the number of the sticks they wanted to take. c is the button participants clicked to take the sticks

H2: The proportion of participants who want to re-challenge their opponent in the competition task is less if they competed against AI than if they competed against a human.

The present study contributes to the literature on causal attribution and human–AI interaction and presents meaningful findings that may help predict people's behavior in situations where they are defeated by AI. The results of this study are expected to contribute to the investigation of the psychological responses of humans whose jobs may be replaced by AI in the future.

Ethics Statement

All procedures used in this research were conducted in accordance with the guidelines of the Japanese Psychological Association. The experimental procedures were approved by the Research Ethics Review Committee regarding Human Subject Research of the Doshisha University (Faculty of Psychology), reference number 202016R2. None of the studies reported in this manuscript was preregistered.

Experiment 1

In Experiment 1, we conducted a laboratory experiment to clarify participants' causal attributions and behavior after losing a game against AI. An original competition task was used that was unfamiliar to participants, and only university students were recruited as participants. These procedures were adopted to control for individual variables such as the participants' initial skill level and age as much as possible and to increase internal validity. All the game screens and questionnaire wording used in Experiments 1 and 2 are available at the Open Science Framework (OSF; https://bit.ly/3diOsVe).

Method

Participants

Given the novel nature of this research, we could not identify an ideal sample size. A large sample size leads to correct estimations of average scores, differences in scores between groups, and effect size (Funder & Ozer, 2019). As the current research was novel, it was necessary to generate accurate and clear findings; thus, as large a sample that our resources allowed was selected. The COVID-19 pandemic reduced the number of students attending university; thus, we recruited fewer participants than expected. Ultimately, 74 students from a Japanese university (15 male, 59 female; age $M = 19.62$, $SD = 1.16$ years) participated in Experiment 1.

Design

Experiment 1 employed a 2 (opponent type: AI vs. human) × 4 (attributed factor: self-ability vs. self-effort vs. opponent ability vs. luck) design. The opponent type was manipulated as a between-participants factor, while the type of causal attribution was a within-participants factor. A total of 36 and 38 participants were assigned to the AI and human conditions, respectively.

Taking Sticks Game

A "taking sticks" game was created to allow the participants to compete with the AI opponent using Visual Studio 2017 (Version 15.9.30; Microsoft, 2020). The game used a one-to-one format. Two players chose between sticks of

three colors——red, yellow, and blue——in turn (Fig. 2).[1] The players could only take sticks of one color; however, they could take as many sticks as they wanted if the sticks were of the same color. The game did not allow players to take two or three colors at once. They could not end their turn without taking a stick. The player who took the last stick lost the game.

We invited two participants into the experimental room for each session. For experimental control, all participants competed against the AI opponent regardless of whether they were allocated to the AI or human condition, although those in the human condition were told that they would be competing against other participants. We partitioned the experimental room into two parts using a curtain, and the two participants were seated in front of computers arranged in each part of the room so that they could not see each other. When only one participant came to the experimental room, they were assigned to the AI condition. The participants were randomly assigned to take the first or second move before starting the game; the number of sticks in the three colors was then determined.

The game was programmed for the participants to be defeated. There was a definite way to win the game, depending on the turn and the number of sticks that were first determined. The winning method was as follows. Players were supposed to convert the number of sticks into binary numbers and take the sticks in such a way that the total of the binary numbers representing the sticks of the three colors became zero. For example, in a turn comprising four red sticks, two yellow sticks, and three blue sticks, the player needed to take three red sticks to make the total of the binary numbers representing the sticks equal zero. The game was programmed such that the AI opponent could adopt the winning strategy, while the participants could not. The total binary number was always zero in the participants' turns. A detailed explanation of the winning method is presented in an online supplemental file.

Measures

Participants answered questions on the computers after completing the taking sticks game. We recorded the participants' attributions of their defeat to self-ability ($\alpha = .82$), self-effort ($\alpha = .53$), opponent ability ($\alpha = .90$), and luck ($\alpha = .81$) as the dependent variables (Table 1). We also measured their perceived enjoyment of the game ($\alpha = .94$) and each individual's risk aversion ($\alpha = .44$) as the control variables (Table 1). These items were measured with a 7-point Likert scale ranging from "1 = do not agree at all" to "7 = strongly agree." We included a test item among these questions (e.g., "For this item, please indicate No. 5") to assess whether the respondents were concentrating while answering the questionnaire.

The participants were then asked to select whether they wanted to compete against the same opponent again; this selection was measured as a behavioral response. We checked whether participants understood the principles of the winning formula using the binary number and provided them with the number of sticks of the three colors (e.g., two reds, four yellows, and seven blues). The participants then selected the color and number of sticks that they wanted to take. If they understood the winning strategy, they would need to take the sticks so that the total of the binary numbers representing the sticks of the three colors became zero. When taking one blue stick, in the case of the above example, the total number of sticks became zero. Thereafter, we asked participants three questions related to the usage of the binary number. If a participant correctly answered these three questions, it was considered that they understood how to win the game. The participants were also required to answer whether they knew that there was a definite method through which to defeat their opponent in the game and that the game was designed to make them lose from the beginning. Additionally, in the human condition, participants were asked if they noticed that the real opponent was not another participant but a computer. These questions were used to determine participants' understanding of the winning strategy and the design of the game. We thus excluded data of participants who were aware of the winning strategy and the design of the game, as these understandings might affect the experimental results.

Procedure

The participants entered the experimental room and were seated in front of a computer. A researcher explained the experiment to them and that they would receive a prepaid card (JPY 500) as a reward for participating in the experiment as well as a gift as an additional reward. The participants were then asked for their informed consent to participate in the study. Thereafter, the researcher explained the rules of the game using an operation manual and that if participants defeated their opponent, they would receive an additional reward. As the participants could refer to the manual during the game, they were not given a training session in Experiment 1. We randomly assigned each participant to either the AI or human condition. After finishing the game, the participants were required to complete the questionnaire. Finally, they were told that the game was designed for them to lose and that they would be provided with an additional reward regardless of their win. They were also informed that they had actually competed

[1] Colored figures depicting the games used in this study are available in the online supplemental material (https://bit.ly/3diOsVe).

Table 1 Questionnaire Items in Experiments 1–3

Self-ability
• I lost because I lacked the logical thinking necessary for this game.
• I lost because I lacked the ability to see several moves ahead.
• I lost because I lacked the ability to find a path to victory.
Self-effort
• I lost because I did not seriously engage in the game.
• I lost because I did not intensively engage in the game.
• I lost because I did not use my maximum ability. [a]
• I lost because I did not engage in the game with all my might. [b]
Opponent ability
• I lost because the AI (or opponent) had the logical thinking necessary for this game.
• I lost because the AI (or opponent) had the ability to see several moves ahead.
• I lost because the AI (or opponent) had the ability to find a path to victory.
Luck
• I lost because I was unlucky.
• I lost because the order of the first move / second move was unfavorable.
• I lost because the number of three colored sticks was unfavorable. [a]
• I lost because the arrangement of the 10 coins was unfavorable. [b]
Perceived enjoyment [c]
• Playing the taking sticks game is enjoyable.
• The taking sticks game is fun.
• The taking sticks game is boring. [d]
Risk aversion [a]
• I prefer situations where I gain a foreseeable profit.
• I avoid situations that present the possibility of loss.
• I avoid situations where it is not certain how things will turn out.
Perceived stability of AI's (humans') performance [e]
• Humans (or AI) are always able to make the best choice in "the taking coin" game.
• Humans (or AI) are able to stably exert their maximum strength in "the taking coin" game.
• The strength of humans (or AI) in "the taking coin" game is stable.

AI = artificial intelligence. [a] These items were used in only Experiment 1. [b] These items were used in Experiments 2 and 3. [c] In Experiments 2 and 3, the taking sticks game was replaced with the taking coins game in these items. [d] Reverse scoring was used for this item. [e] These items were used in only Experiment 3

with the AI opponent in the human condition. The task used in this study was a computer game, and the AI in the game did not perform machine-learning or deep-learning. The AI simply performed according to the program to win the game. The term "computer" may, therefore, be more appropriate than "AI" to describe the opponent. In recent years, however, algorithm opponents have been described as AI in computer games such as Othello and UNO card game. As this study focused on a situation where human performance was inferior to that of AI, our experimental procedures aimed to make participants believe that their opponent was AI. Thus, we used the term "AI."

Results

The raw data, codebook, and R code for all studies are available at the OSF (https://bit.ly/3diOsVe). All participants completed the experiment and gave correct answers to the test question. None of them completely understood the specific method with which they could win the game. Some participants reported that they knew there was a winning strategy. However, as no participant gave correct answers to all the questions regarding the understanding of binary numbers, it was inferred that they did not actually know the winning method, even if aware of its existence. If they really knew the strategy, they would have given accurate answers to all the questions. Thus, we did not exclude the data of respondents who reported that they knew of the existence of a winning strategy. On the other hand, we excluded the data of participants who noticed that the game was created for them to lose from the beginning. In the human condition, respondents who realized that the opponent was an AI opponent were also eliminated. In total, 16 participants were eliminated from the analysis, resulting in a final sample of

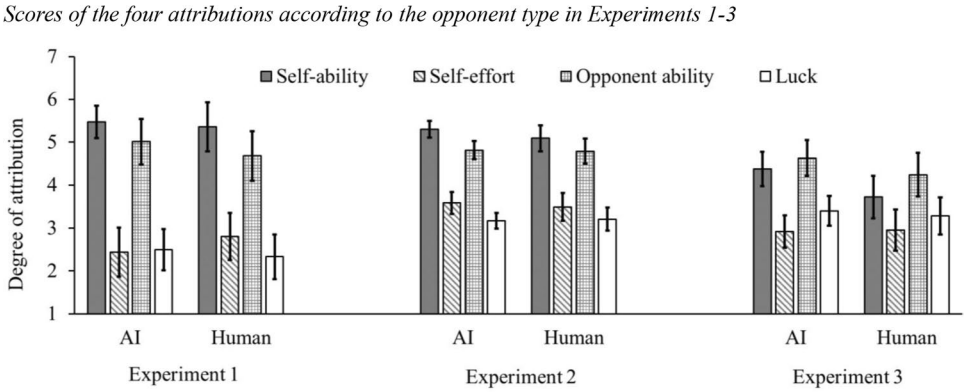

Fig. 3 Scores of the four attributions according to the opponent type in Experiments 1–3. Error bars represent 95% CIs

58 (AI condition = 31, human condition = 27). Additionally, the item assessing self-effort——"I lost because I did not use my maximum ability"——was not significantly correlated with the other two items ($r = .06$, $r = .18$) and was, therefore, excluded when the average self-effort score was calculated.[2] The two remaining items showed higher internal consistency ($\alpha = .79$) than that of all three combined.

Causal Attribution

We performed a 2 (opponent type: AI, human) × 4 (attributed factor: self-ability, self-effort, opponent ability, luck) mixed ANOVA with the last factor as repeated measures to test the difference in the participants' causal attributions in the AI and human conditions. The analysis found a significant effect of the attributed factor (F (3, 168) = 75.46, $p < .001$, $\eta^2_p = .574$), although the opponent type (F (1, 56) = .10, $p = .755$, $\eta^2_p = .002$) and two-way interaction (F (3, 168) = .71, $p = .548$, $\eta^2_p = .013$) had no significant effect. The descriptive statistics shown in Fig. 3 demonstrate that the pattern of causal attribution in the competition with the AI opponent was similar to the pattern when participants believed they were competing against a human. Table 2 shows the results of multiple comparison tests with a Bonferroni correction, which investigated the differences in attribution scores among the four attributed factors. The results indicated that participants attributed the loss of the game to self-ability more than any other factor.[3]

Behavioral Response

The effect of the opponent type on participants' choice to re-challenge them was tested using logistic regression. We also entered perceived enjoyment and risk aversion into the exploratory variables to control for their influence. Since the internal consistency of the three items of risk aversion was low, each item was individually added to the exploratory variables. A dummy variable was used for the opponent type as follows: 1 = AI condition and 0 = human condition. The results revealed that participants' choice to re-challenge their opponent was not significantly influenced by the opponent type (Table 3); that is, participants' behavior after losing the game did not significantly differ regardless of whether the opponent was AI or human. The proportion of participants who wanted to play the taking sticks game again was generally low in both the AI and human conditions (Table 4).

[2] We conducted a two-way ANOVA as a supplemental analysis using the average score of the three self-effort items in Experiment 1. The results were similar to those of the analysis using the two self-effort items. The supplemental analysis showed that the attributed factor had a significant effect on the degree of attribution (F (1, 168) = 82.79, $p < .001$, $\eta^2_p = .596$). The effects of opponent type (F (1, 56) = .26, $p = .614$, $\eta^2_p = .005$) and two-way interaction (F (1, 168) = .48, $p = .699$, $\eta^2_p = .008$) were not significant. Multiple comparison tests with a Bonferroni correction revealed that the attribution score for self-ability was higher than those for self-effort (t (56) = 12.35, $p < .001$, Cohen's $d = 2.11$), opponent ability (t (56) = 3.41, $p = .002$, Cohen's $d = .42$) and luck (t (56) = 12.11, $p < .001$, Cohen's $d = 2.32$). Opponent ability was evaluated as an attributional cause more frequently than self-effort (t (56) = 8.74, $p < .001$, Cohen's $d = 1.51$) and luck (t (56) = 9.01, $p < .001$, Cohen's $d = 1.73$). There was no significant difference in ratings between self-effort and luck (t (56) = 1.71, $p = .092$, Cohen's $d = .32$).

[3] As a supplemental analysis, in all three experiments, we conducted MANOVAs with four attribution scores of self-ability, self-effort, opponent ability, and luck as the dependent variables and the opponent type as the independent variable. The analysis found that the opponent type had no significant effect on causal attribution in Experiments 1 (F (1, 56) = .45, $p = .771$), 2 (F (1, 309) = .51, $p = .673$), and 3 (F (1, 123) = 1.58, $p = .198$). These results were congruent with those of the two-way ANOVAs.

Table 2 Results of the multiple comparison tests in Experiments 1–3

Pair	*T*-value	*P* value	Cohen's *d* (95%CI)
Experiment 1			
Self-ability vs. Self-effort	11.07	< .001	2.03 (1.58–2.47)
Self-ability vs. Opponent ability	3.41	.002	.42 (.05–.78)
Self-ability vs. Luck	12.11	< .001	2.32 (1.85–2.79)
Self-effort vs. Opponent ability	8.31	< .001	1.50 (1.09–1.91)
Self-effort vs. Luck	.74	.462	.14 (−.22–.51)
Opponent ability vs. Luck	9.01	< .001	1.73 (1.31–2.16)
Experiment 2			
Self-ability vs. Self-effort	13.24	< .001	.98 (.81–1.14)
Self-ability vs. Opponent ability	4.43	< .001	.25 (.09–0.41)
Self-ability vs. Luck	15.54	< .001	1.34 (1.16–1.51)
Self-effort vs. Opponent ability	9.50	< .001	.74 (.58–.90)
Self-effort vs. Luck	2.75	.038	.21 (.05–.37)
Opponent ability vs. Luck	13.13	< .001	1.06 (.90–1.23)
Experiment 3			
Self-ability vs. Self-effort	5.81	< .001	.64 (.38–.89)
Self-ability vs. Opponent ability	2.41	.035	.22 (−.03–.46)
Self-ability vs. Luck	3.43	.003	.43 (.18–.68)
Self-effort vs. Opponent ability	7.04	< .001	.85 (.59–1.10)
Self-effort vs. Luck	1.99	.049	.25 (.03–.50)
Opponent ability vs. Luck	5.32	< .001	.65 (.40–.90)

P values were calculated with a Bonferroni correction

Discussion

Experiment 1 showed that participants made self-effacing attributions when they were defeated in the taking sticks game, irrespective of whether their opponent was AI or human; that is, they attributed the loss to their own abilities more than any other cause. The proportion of participants who wanted to try the game again did not significantly differ between the AI and human conditions. However, we could not propose a conclusion based on only one experiment, and accumulated evidence was required to understand the patterns of causal attribution when people are defeated by AI. Experiment 2 was thus performed to examine whether the results of Experiment 1 could be replicated.

Experiment 2

Experiment 2 investigated whether the findings of Experiment 1 could be reproduced and tested both replicability and generalizability. Experiment 2 attempted to extend the findings of Experiment 1 to another competition task and extended the sample to graduates, rather than university students.

Table 3 Logistic regression results of participant's choice to re-challenge their opponent in Experiments 1–3

	Unstandardized coefficient	95%CI	Z-value	P value
Experiment 1				
Intercept	1.77	−2.38–6.08	.84	.402
Opponent type	.05	−1.17–1.29	.08	.934
Perceived enjoyment	.15	−.34–.65	.59	.552
Risk aversion 1	−.25	−.68–.16	1.19	.236
Risk aversion 2	−.52	−1.03– −.06	2.10	.036
Risk aversion 3	.06	−.38–.54	.26	.793
Experiment 2				
Intercept	−2.19	−3.01– −1.43	5.47	< .001
Opponent type	−.56	−1.08– −.04	2.12	.035
Perceived enjoyment	.47	.31– −.64	5.58	< .001
Experiment 3				
Intercept	−1.12	−2.41–.10	1.76	.079
Opponent type	.34	−.38–1.07	.92	.357
Perceived enjoyment	.17	−.04–.40	1.57	.117

Method

Participants

COVID-19 restrictions made it difficult to conduct an experiment in the experimental room; therefore, Experiment 2 was carried out online. As in Experiment 1, we tried to recruit as many participants as possible via a Japanese

Table 4 Number of participants who did or did not want to compete against the same opponent again in Experiments 1–3 by Condition

	Re-challenge	Do not re-challenge
Experiment 1		
AI condition	11 (35.5%)	20 (64.5%)
Human condition	9 (33.3%)	18 (66.7%)
Experiment 2		
AI condition	69 (34.2%)	133 (65.8%)
Human condition	46 (42.2%)	63 (57.8%)
Experiment 3		
AI condition	39 (53.4%)	34 (46.6%)
Human condition	23 (55.8%)	29 (44.2%)

Values in brackets are proportions of each condition

Fig. 4 Arrangement of the 10 Japanese coins at the beginning of the taking coins game

survey company, Cross Marketing (https://www.cross-m.co.jp/en/). A total of 788 adults (402 male, 386 female; age $M = 40.73$, $SD = 10.75$ years) volunteered to participate. Participants were recruited by e-mail and accessed a website to participate in the experiment. They earned points that could be exchanged for cash or prepaid cards.

Design

The experimental design remained identical to that of Experiment 1 to test the generalizability of the findings of Experiment 1. The AI and human conditions included 407 and 381 participants, respectively.

Taking Coins Game

In Experiment 2, a "taking coins" game was used with the online survey system Qualtrics. The game had a one-to-one format. Ten Japanese yen coins were arranged in a line (Fig. 4), and the players took one coin per turn. They had to take a coin from either the leftmost or rightmost side and could not complete their turn without taking a coin. The player whose total coin value was more than their opponent's at the end of the game was the winner.

Although the participants competed with the AI opponent in both the AI and the human conditions, as in Experiment 1, those in the human condition were informed that they would compete with another participant and that the turns and arrangement of the coins were randomly determined. However, participants were deliberately allocated to take the second turn at the beginning of the game and to the arrangement of the 10 coins shown in Fig. 4, which meant they would lose the game.

In the taking coins game, the first mover could win the game if they knew how to defeat the opponent. The winning method was as follows. The first mover totaled the value of the coins of every other coin from the left-hand side as well as from the right-hand side. If the total calculated from the left was more than that calculated from the right, they needed to take the leftmost coin in the first turn. Conversely, if the total calculated from the right was higher, they needed to tale the rightmost coin in the first turn. For example, referring to the 10 coins shown in Fig. 4 (5 yen, 10 yen, 50 yen, 100 yen, 5 yen, 100 yen, 10 yen, 1 yen, 100 yen, and 1 yen), the total value was 170 when summing every other coin from the left-hand side ($5 + 50 + 5 + 10 + 100 = 170$); the total was 212 when summing up every other coin from the right-hand side ($1 + 1 + 100 + 100 + 10 = 212$). To win the game, the player who made the first move had to take the rightmost coin in the first turn; thereafter, they needed to take the coins in congruence with the decision their opponent made. For instance, if the opponent took the rightmost coin, the first mover needed to take the next rightmost coin in their turn; if the opponent took the leftmost coin, the first mover needed to take the next leftmost coin. This strategy would enable the first mover to win the game; thus, we intentionally allocated the first move to the AI opponent and the second to the participants.

Measures

The questionnaire used in Experiment 1 was also used in Experiment 2 to assess causal attributions and perceived enjoyment, and the same test questions were included. The self-effort items improved from those of Experiment 1 due to low internal consistency (Table 1). We could not provide participants with an additional reward because Experiment 2 was conducted online; therefore, when they opted whether to play again, we could not state that the additional reward depended on their performance. The items of risk aversion were thus excluded from the questionnaire.

We checked whether participants understood that there was a particular way to win the game by asking them, "There was a definite way to win the taking coins game, depending on whether you had the first or second move. Who do you think could use the winning strategy, the first mover or the second mover?" They were then provided with 10 coins and asked, "If you had the first move, which coin would you take, the leftmost or rightmost?" Thereafter, we asked them three questions and presented them with 10 coins. If a participant correctly answered these three questions and realized that the player with the first move could use the winning method, we considered that they understood how to defeat an opponent in the taking coins game. Finally, all participants

were asked whether they knew that the game was designed so that they would lose, while those in the human condition were also asked if they noticed that their opponent was not another participant but a computer.

Procedures

Participants accessed the experiment via a website. Electronic informed consent was obtained from all participants by Cross Marketing. Participants were randomly assigned to either the AI or human condition and the rules of the game were explained to them. We allowed them to undertake adaptive training with a line of four coins before the formal game. After the training session, the arrangement of the 10 coins and the order of the players were determined; to ensure that they believed they were competing against another participant in the human condition, they were shown a screen stating that they were being synchronized to another participant in real time. After finishing the formal game, the participants were asked to complete the questionnaire. Finally, they were told that the game was designed for them to lose and that they had competed with an AI opponent in the human condition.

Results

Data from the 477 participants who gave incorrect answers to the item assessing their concentration, who completely understood how to win the game, who noticed that the game was designed for them to lose, or who realized that their opponent was not human (in the human condition) were removed. As this study used multiple criteria to decide whether to exclude certain data, many participant responses were eliminated before data analysis. This resulted in a final sample of 311 (AI condition = 202, human condition = 109). The data of only those participants who completed the experiment were provided by the survey company. Thus, we did not know how many people quit the experiment halfway. All questionnaire items had adequate internal consistency (self-ability, $\alpha = .90$; self-effort, $\alpha = .94$; opponent ability, $\alpha = .90$; luck, $\alpha = .71$; perceived enjoyment, $\alpha = .94$).

Causal Attribution

As in Experiment 1, a two-way mixed ANOVA was performed to examine the causal attribution in both the AI and human conditions. The results clarified that the attributed factor had a significant effect on the degree of attribution (F (3, 927) = 126.61, $p < .001$, $\eta^2_p = .291$). The effects of the opponent type (F (1, 309) = .48, $p = .488$, $\eta^2_p = .002$) and two-way interaction (F (3, 927) = .41, $p = .748$, $\eta^2_p = .001$) were not significant. The descriptive statistics showed that the pattern of causal attribution in the AI condition did not significantly differ from that in the human condition (Fig. 3). Table 2 describes the results of the multiple comparison tests with a Bonferroni correction, which investigated differences in the attribution scores among the four attributed factors. The analysis found that self-ability was attributed as factor of participants' loss more than any other factor. The participants also attributed their loss to self-effort more than luck.

Behavioral Response

We tested whether the opponent type affected whether participants chose to play again using logistic regression analysis. Perceived enjoyment was also entered into the exploratory variables to control for its influence. The results demonstrated that the opponent type had a significant effect on the choice to re-challenge in the game (Table 3), suggesting that when the opponent was AI, the proportion of participants who wanted to play the game again decreased (Table 4). Perceived enjoyment positively influenced participants' choice to play again.

Discussion

The results of Experiment 2 were partially in line with those of Experiment 1. The subjective evaluation of causal attribution was replicated. Participants attributed the loss of the game to their own ability more than any other cause in both the AI and human conditions. We increased the sample size in Experiment 2 compared with that in Experiment 1, which enhanced the power of the test. The participants' causal attribution did not differ significantly between the AI and human conditions; rather, they tended to make self-effacing attributions, regardless of the condition. The causal attribution pattern in the AI condition may be robust. Experiments 1 and 2 recruited only Japanese participants, however, and there is a cultural difference in causal attribution in that Western people are more inclined to display self-serving attributions than Japanese people (Mezulis et al., 2004). It is thus necessary to test the replicability of our findings in a sample of Western people.

On the contrary, the behavioral response result was not reproduced. The number of participants who wanted to play the taking coins game again was higher in the human condition than in the AI condition. It was considered that the perceived stability of the opponent's performance would influence the decision to re-challenge. The participants might think that AI could stably exert its maximum strength and they would be defeated by AI, even if they competed again. They might thus be less likely to compete against AI than against humans.

Experiment 3

Experiment 3 tested whether the findings of Experiments 1 and 2 could be reproduced with American participants. We also aimed to explain why people were reluctant to play the taking coins game again in the AI condition more than in the human condition, focusing on the perceived stability of AI's (or humans') performance. The procedure of Experiment 3 was consistent with that of Experiment 2, except for the types of participants and items in the questionnaire section.

Method

Participants

This study tried to make the sample as large as resources would allow. American adults were recruited via a survey company, Qualtrics (https://www.qualtrics.com/research-services/online-sample/). A total of 500 adults who had already graduated (233 males, 261 females; age $M = 40.01$, $SD = 10.77$ years) volunteered to participate for a compensation. Six participants did not indicate their gender and one participant did not reveal their age. The AI and human conditions comprised 256 and 244 participants, respectively.

Measure

In addition to the items used in Experiment 2, we recorded the perceived stability of AI's (or humans') performance (Table 1).

Results

We eliminated respondents who gave incorrect answers to the item assessing their concentration, completely understood how to win the game, noticed that the game was designed for them to lose, or realized that their opponent was not human (in the human condition), resulting in a final sample of 125 (AI condition = 73, human condition = 52). Similar to Experiment 2, many respondents who did not pass the multiple check items were excluded. As in Experiment 2, the data of only those participants who completed the experiment were provided by the survey company, and we did not know whether anyone quit the experiment halfway. All questionnaire items had adequate internal consistency (self-ability, $\alpha = .81$; self-effort, $\alpha = .84$; opponent ability, $\alpha = .87$; luck, $\alpha = .64$; perceived enjoyment, $\alpha = .91$; perceived stability of AI's or humans' performance, $\alpha = .75$).

Causal Attribution

A two-way mixed ANOVA was carried out to investigate the causal attribution in both the AI and human conditions. The results demonstrated that the attributed factor significantly influenced the degree of attribution (F (3, 369) = 23.47, $p < .001$, $\eta^2_p = .160$). The effects of the opponent type (F (1, 123) = 2.35, $p = .128$, $\eta^2_p = .019$) and two-way interaction (F (3, 369) = 1.17, $p = .319$, $\eta^2_p = .009$) were not significant. The descriptive statistics indicated that the pattern of causal attribution in the AI condition was similar to that in the human condition (Fig. 3). Table 2 shows the results of the multiple comparison tests with a Bonferroni correction that examined differences in the attribution scores among the four attributed factors. The analysis revealed that self-ability and opponent ability were attributed as factors behind participants' loss more than any other factor. Specifically, the participants most attributed the loss to opponent ability; luck was also attributed more than self-effort.

Behavioral Response

This study investigated the effect of the opponent type on trying the game again using logistic regression analysis. Perceived enjoyment was also entered into the exploratory variables to control for its influence. The results showed that neither the opponent type nor perceived enjoyment significantly affected the choice to re-challenge in the game (Table 3). The proportions of participants who wanted to play the taking coins game again was similar between the AI and human conditions (Table 4).

Supplemental Analysis

It was assumed that people perceived AI's performance as more stable than humans' performance. After they lost the game, therefore, they might be more reluctant to compete with AI than with a human. We tested the difference in perceived stability between the AI and human conditions, as a supplemental analysis, although the proportion of re-challenges did not significantly depend on the opponent type. It was found that the stability of AI's performance ($M = 5.16$) was significantly rated more than that of humans' performance ($M = 4.58$) using an independent samples t-test (t (123) = 2.52, $p = .013$, Cohen's $d = .46$).

Discussion

Experiment 3 adopted the taking coins game as in Experiment 2 and tested Americans' causal attribution after they lost the game against AI. The results of Experiment 3 were consistent with those of Experiments 1 and 2, not supporting our hypotheses. Americans tended to attribute their loss to their and their opponent's abilities irrespective of whether the opponent was AI or human. The ratings of attribution to their effort and luck were low. Americans thought that

opponent ability most affected their loss, which differed from the results of Experiments 1 and 2, wherein Japanese participants attributed their loss to self-ability more than any other factor. Moreover, the attribution score of luck was higher than that of self-effort in Experiment 3, although this difference was reversed in Experiment 2. The proportion of participants who wanted to re-challenge the game was similar between the AI and human conditions. This result was different from that of Experiment 2.

General Discussion

The current research investigated causal attributions and behaviors when people lose a competition against AI. We created two competition tasks that required people to display logical thinking in the experiments, as it is expected that AI might replace human jobs requiring such thinking in the future (Huang & Rust, 2018). The results related to causal attributions were consistent between the three experiments, while those of the behavioral responses were mixed. Participants generally made self-effacing attributions for their loss; that is, they attributed the outcome to internal and stable factors. Participants rarely attributed unstable factors——their own effort and luck——to their loss. As the same pattern of causal attribution was shown in three experiments, the findings may be considered robust. In terms of whether the participants wanted to play the game again after they lost (i.e., their behavioral response), the results implied that such behavior changes depend on the experimental environment.

Causal Attribution

The three experiments indicated that people attributed their loss against the AI opponent to their own and their opponent's abilities more than any other cause. Self-effacing attribution was also found in the condition where participants believed that they had lost the game to another human, that is, people make similar causal attributions when they lose to AI as when they lose to a human. The results do not support H1 that people are more likely to attribute the loss to the opponent's ability and are less likely to attribute it to luck when they lose the game against AI than when they lose against a human. Even though the task (i.e., the taking sticks game vs. the taking coins game) and participants' characteristics (university students vs. adults who had already graduated; Japanese vs. Americans) changed among the three experiments, self-effacing attributions were found in the AI and human conditions. No significant interaction (opponent type × attributed factors) was observed in Experiment 2 where we drastically increased the number of participants, additionally implying that the pattern of causal attribution does not greatly change regardless of whether the opponent is AI or human.

There are two explanations for our findings related to causal attribution. First, this study used competition tasks demanding participants to make logical decisions and see several moves ahead. Based on such tasks, the participants might have thought that the outcome of the games depended on their own and their opponent's logical thinking ability. Although research on causal attribution has demonstrated that Western populations tend to attribute failures to external rather than internal factors (Mezulis et al., 2004), the results of Experiment 3 found that Americans attributed their loss to their ability more than luck. The results of Experiments 1 and 2 are in line with prior research that Japanese are likely to make self-effacing attributions. High attribution to self-ability and low attribution to luck might have been brought about by the characteristics of our games.

The second possible account is as follows: participants responded to their interaction with the AI opponent in the same way as when they believed they were interacting with a human, thereby demonstrating the same causal attribution between the AI and human conditions. According to Reeves and Nass' (2001) media equation theory, people are inclined to respond and behave similarly to computers and robots as they are to humans. This theory has been supported by several findings, and Nass and Moon (2000) reviewed numerous studies that contributed to it. For instance, it was observed that people show ingroup bias toward computers and exhibit politeness when interacting with robots. Since the participants adapted the same responses and behaviors to AI as to humans, they might have made the same attribution in both the AI and the human conditions.

In Experiments 1 and 2, the Japanese considered that self-ability most influenced their loss. They also attributed their loss to self-effort more than luck in Experiment 2. In Experiment 3, on the other hand, Americans reported that their loss most depended on opponent ability and attributed to luck more than self-effort. These results can be explained by literature on cultural differences: For instance, the results of a meta-analysis by Mezulis et al. (2004) revealed that individuals from Western cultures tend to attribute their failure to external factors while Japanese tend to attribute it to internal ones. This cultural background might have influenced the evaluation of causal attribution in our experiments. Although self-ability and opponent ability were attributed more than self-effort and luck, overall, in both Japanese and American samples, this study observed slight cultural differences in causal attribution.

Behavioral Response

In terms of whether participants wanted to re-challenge the game after they were defeated by the AI opponent or believed they had been defeated by a human, the findings were mixed. Only the results of Experiment 2 supported H2 that people are more averse to re-challenging the competition task against AI than against a human. In contrast, the hypothesis was not supported in Experiments 1 and 3. These mixed results might have been brought about by the characteristics of our games.

The number of participants who opted to play again did not significantly differ between the AI and human conditions in Experiment 1. Overall, few participants chose to re-challenge the game, resulting in a floor effect. We offer two potential explanations for these results. First, the difficulty of the taking sticks game might have influenced participants' decisions. This game required players to choose the color and number of sticks that they wanted to take. Thus, they needed to plan several moves ahead and think logically to win the game, although the game was programmed for them to lose. Since the game was difficult for participants, they might have been averse to re-challenging their opponent. The second explanation is related to loss aversion, which posits that people are more sensitive to losses than to gains (Tversky & Kahneman, 1991). We told the participants that one reward would be removed if they lost the second game; thus, they may have wanted to avoid losing their rewards, which might have led to few participants deciding to try the game again.

In Experiment 2, the number of participants who wanted to play again was less in the AI condition than in the human condition. In the taking coins game, participants were asked to take one of only two coins on the left- or right-hand side, and the rules of the taking coins game might have been easier for them to understand than those of the taking sticks game. Additionally, in Experiment 2, the reward amount did not depend on participants' decision to re-challenge the game; the environment of the second experiment was not a situation in which loss aversion was invoked. Therefore, the participants might not focus on defeating their opponents, as there was no influence of the reward. It is also possible that participants wanted to know the programing of the game and explore other ways of taking coins. Thus, they might select whether to re-challenge based not on competition motivation but learning motivation. These procedures might have prevented a floor effect related to the number of participants who wanted to try the game again. The proportion of participants selecting to re-challenge increased overall in Experiment 3, where the taking coins game was used, compared with Experiment 1. However, the results of Experiment 3 did not support H2. About half of the participants wanted to try the taking coins game again in both the AI and human conditions. It is thus difficult to conclude about the behavioral response based only on our findings. Future studies should further investigate people's behavior after they are defeated by AI.

In interpreting the differences in behavioral response between the AI and human conditions in Experiment 2, we inferred that perceived stability of opponent's performance would affect participant behavior. Experiment 3 thus recorded participants' perception of stability to test its effect on the selection of re-challenge. In Experiment 3, however, the number of the participants who wanted to try the game again did not differ between the AI and human conditions. Therefore, we could not examine whether the behavioral response of re-challenge depended on the perceived stability of opponent's performance. The relation between the perception of stability and re-challenge selection should be tested in future research.

Theoretical and Practical Contributions

This study offers the first evidence on how people make causal attributions when competing against AI in a game requiring logical thinking. Psychological research on causal attribution has focused on individual achievement tasks and human–human competitions (Allen et al., 2020; Wang & Hall, 2018); thus, the present study extends the literature on causal attribution by adopting a situation in which people compete against AI. Our three experiments revealed that the pattern of causal attribution was similar regardless of whether participants played a game against AI or believed they were playing a game against a human. This finding could be accurate because the sample size of Experiment 2 was sufficiently large. In the future, AI may take over jobs that are currently performed by humans (Acemoglu & Restrepo, 2018; Huang & Rust, 2018); therefore, future studies should examine causal attribution in circumstances where people are defeated by AI. Our findings may be fundamental for the further investigation of causal attribution in competitions against AI.

Our study also contributes to human–computer interaction research. Much research has already examined attribution in interactions between humans and robots, computers, and autonomous systems. These studies have focused on attributions of responsibility (Awad et al., 2020; Belanche et al., 2020; Hong, 2020; Lee & Cranage, 2018; Leo & Huh, 2020; van der Woerdt & Haselager, 2019); for example, whether people attribute responsibility to either the user or robot when the robot service fails (Belanche et al., 2020; Lee & Cranage, 2018; Leo & Huh, 2020). We add new evidence to the literature on attribution by using a game in which humans competed against AI. Additionally, the present study's finding suggest that people made self-effacing attributions in both competitions against AI and humans,

indicating the generalizability of media equation theory. Our research could integrate this theory into situations of competition against AI.

We also make a practical contribution. Our results indicated that participants attributed their losses against the AI opponent to internal and stable factors and their own ability. Research on causal attribution has found that when people fail a task and attribute the fault to internal and stable factors, they do not expect future success or make an effort to achieve such success (Crittenden & Wiley, 1980; Mamonov & Koufaris, 2018; McMahan, 1973; Rascle et al., 2015; Weiner et al., 1976). People may have an undesirable outlook when their jobs are replaced by AI; thus, interventions may be needed to promote these employees to aim for success. For example, when people are told that ability is important for the achievement of a task and is changeable, they prefer to receive accurate feedback about themselves, irrespective of whether this feedback is positive or negative (Dunning, 1995). Psychological interventions may be required after people lose a competition against AI. On the other hand, job replacement by AI is not invoked only due to differences in logical thinking between AI and human. AI may be adopted to reduce labor costs and increase efficiency of work. Such job replacement differs from losing a competition requiring logical thinking. Future studies should investigate whether the patterns of causal attribution found in our experiments can be generalized to actual scenarios involving job replacement by AI.

Limitations

The limitations of our research should be mentioned. First, we used only original games that required players to have the ability to calculate and think logically. Our experiments are the first to investigate causal attribution when people lost a game against AI; since the study needed to control for extraneous variables, we used original tasks to promote internal validity. However, the pattern of causal attribution remains unclear in tasks that require other abilities, such as artistic skills and creativity. For instance, the Remote Associates Test developed by Mednick (1962) is frequently used to measure creativity, while artistic skills can be evaluated through drawings. The present study did not observe causal attribution when people were defeated by AI in such tasks. Logical tasks, such as the computer games used in this study, which AI can perform perfectly each time, might be perceived as difficult by participants, influencing their causal attributions and decision to re-challenge. To expand the scope of this study, tasks showing diverse aspects of AI, such as learning and prediction, should be used.

The second limitation is related to assessing the behavioral response and the order of our questions. Since this study focused on causal attribution, participants were first asked to answer the question about causal attribution and then evaluate the extent to which the competition tasks were enjoyable. After completing the questionnaire, they selected whether they wanted to play the game again. The order of the questions might then have influenced their decision to re-challenge. If future studies accurately investigate people's behavioral responses after they are defeated by AI, they should measure their behavior at the beginning instead. Additionally, the decision to re-challenge could be affected by diverse extraneous variables, such as participants' fatigue and mood. Expectancy factors, such as the subjective likelihood of winning, should be recorded in addition to the participants' behavior.

Third, as this study mainly focused on causal attribution, some critical variables were not recorded. For example, we did not measure individual difference variables, such as participants' interest, computer literacy, and gaming skills. These variables might influence the decision to try the game again. It was found, for example, that attitude toward playing online games is correlated with intention to play online games (Wu & Liu, 2007). Hancock et al. (2011) also demonstrated that high expertise leads to a positive attitude toward an interaction with a computer. Emotions such as shame and regret should be assessed because research on causal attribution has clarified that emotions are correlated with causal attribution (Weiner, 1979, 1985, 2010). Future studies need to measure these variables and statistically control their effects to understand the impact of the opponent type on the causal attribution and the decision to re-challenge.

Finally, the experimental design used in the current research comprised only a situation where participants lost the game. In addition to the losing condition, including a winning condition might lead to a more appropriate understanding of participants' causal attribution when they are defeated by AI. Such design would enable researchers to investigate the extent to which people make a self-effacing (or self-serving) attribution in a situation where they lose competition against AI, compared to when they win.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12144-021-02559-w.

Acknowledgments We thank T. Ozaki and Y. Shibata for their valuable comments and Editage (www.editage.com) for English language editing.

Funding This research was partially supported by a Grant-in-aid from the Japan Society for the Promotion of Science, Grant Number 20J11786.

Availability of Data and Material De-identified data with a codebook and method file for all studies are posted at https://bit.ly/3diOsVe.

Code Availability The R code for analysis is available at OSF (https://bit.ly/3diOsVe).

Declarations

Conflicts of Interest/Competing Interests The authors declare no conflicts of interest with respect to the authorship or the publication of this article.

References

Acemoglu, D., & Restrepo, P. (2018). The race between man and machine: Implications of technology for growth, factor shares, and employment. *American Economic Review, 108*(6), 1488–1542. https://doi.org/10.1257/aer.20160696

Allen, M. S., Robson, D. A., Martin, L. J., & Laborde, S. (2020). Systematic review and meta-analysis of self-serving attribution biases in the competitive context of organized sport. *Personality and Social Psychology Bulletin*, *46*(7), 1027–1043. https://doi.org/10.1177/0146167219893995.

Awad, E., Levine, S., Kleiman-Weiner, M., Dsouza, S., Tenenbaum, J. B., Shariff, A., Bonnefon, J. F., & Rahwan, I. (2020). Drivers are blamed more than their automated cars when both make mistakes. *Nature Human Behaviour, 4*(2), 134–143. https://doi.org/10.1038/s41562-019-0762-8

Belanche, D., Casaló, L. V., Flavián, C., & Schepers, J. (2020). Robots or frontline employees? Exploring customers' attributions of responsibility and stability after service failure or success. *Journal of Service Management, 31*(2), 267–289. https://doi.org/10.1108/JOSM-05-2019-0156

Bradley, G. W. (1978). Self-serving biases in the attribution process: A reexamination of the fact or fiction question. *Journal of Personality and Social Psychology, 36*(1), 56–71. https://doi.org/10.1037/0022-3514.36.1.56

Brown, M., Houghton, R., Sharples, S., & Morley, J. (2015). The attribution of success when using navigation aids. *Ergonomics, 58*(3), 426–433. https://doi.org/10.1080/00140139.2014.977827

Campbell, W. K., & Sedikides, C. (1999). Self-threat magnifies the self-serving bias: A meta-analytic integration. *Review of General Psychology*, *3*(1), 23–43. https://doi.org/10.1037/1089-2680.3.1.23.

Chandler, T. A., Shama, D. D., Wolf, F. M., & Planchard, S. K. (1981). Multiattributional causality: A five cross-national samples study. *Journal of Cross-Cultural Psychology*, *12*(2), 207–221. https://doi.org/10.1177/0022022181122006.

Crittenden, K. S., & Wiley, M. G. (1980). Causal attribution and behavioral response to failure. *Social Psychology Quarterly, 43*(3), 353–358. https://doi.org/10.2307/3033739

Dawes, J. (2017). The case for and against autonomous weapon systems. *Nature Human Behaviour, 1*(9), 613–614. https://doi.org/10.1038/s41562-017-0182-6

De Michele, P. E., Gansneder, B., & Solomon, G. B. (1998). Success and failure attributions of wrestlers: Further evidence of the self-serving bias. *Journal of Sport Behavior, 21*(3), 242.

Dijkstra, J. J., Liebrand, W. B., & Timminga, E. (1998). Persuasiveness of expert systems. *Behaviour & Information Technology, 17*(3), 155–163. https://doi.org/10.1080/014492998119526

Dunning, D. (1995). Trait importance and modifiability as factors influencing self-assessment and self-enhancement motives. *Personality and Social Psychology Bulletin*, *21*(12), 1297–1306. https://doi.org/10.1177/01461672952112007.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors*, *44*(1), 79–94. 10.1518/0018720024494856.

Ford, J. D. (1985). The effects of causal attributions on decision makers' responses to performance downturns. *Academy of Management Review, 10*(4), 770–786. https://doi.org/10.5465/amr.1985.4279100

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change, 114*, 254–280. https://doi.org/10.1016/j.techfore.2016.08.019

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. https://doi.org/10.1177/2515245919847202.

Grove, J. R., Hanrahan, S. J., & McInman, A. (1991). Success/failure bias in attributions across involvement categories in sport. *Personality and Social Psychology Bulletin*, *17*(1), 93–97. https://doi.org/10.1177/0146167291171014.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, *53*(5), 517–527. https://doi.org/10.1177/0018720811417254.

Heider, F. (1978). *The psychology of interpersonal relations* (M. Ohashi, Trans.). Seishinshobo. (Original work published 1958).

Hinds, P. J., Roberts, T. L., & Jones, H. (2004). Whose job is it anyway? A study of human-robot interaction in a collaborative task. *Human–Computer Interaction, 19*(1–2), 151–181. https://doi.org/10.1080/07370024.2004.9667343

Hong, J. W. (2020). Why is artificial intelligence blamed more? Analysis of faulting artificial intelligence for self-driving car accidents in experimental settings. *International Journal of Human–Computer Interaction, 36*(18), 1768–1774. https://doi.org/10.1080/10447318.2020.1785693

Huang, M. H., & Rust, R. T. (2018). Artificial intelligence in service. *Journal of Service Research*, *21*(2), 155–172. https://doi.org/10.1177/1094670517752459.

Hymes, R. W., & Akiyama, M. M. (1991). Depression and self-enhancement among Japanese and American students. *The Journal of Social Psychology, 131*(3), 321–334. https://doi.org/10.1080/00224545.1991.9713859

Kashima, Y., & Triandis, H. C. (1986). The self-serving bias in attributions as a coping strategy: A cross-cultural study. *Journal of Cross-Cultural Psychology*, *17*(1), 83–97. 0.1177/0022002186017001006.

Kelley, H. H., & Michela, J. L. (1980). Attribution theory and research. *Annual Review of Psychology, 31*(1), 457–501. https://doi.org/10.1146/annurev.ps.31.020180.002325

Lee, B., & Cranage, D. A. (2018). Causal attributions and overall blame of self-service technology (SST) failure: Different from service failures by employee and policy. *Journal of Hospitality Marketing & Management, 27*(1), 61–84. https://doi.org/10.1080/19368623.2017.1337539

Leo, X., & Huh, Y. E. (2020). Who gets the blame for service failures? Attribution of responsibility toward robot versus human service providers and service firms. *Computers in Human Behavior, 113*, 106520. https://doi.org/10.1016/j.chb.2020.106520

Lyons, J. B., & Stokes, C. K. (2012). Human–human reliance in the context of automation. *Human Factors*, *54*(1), 112–121. 10.1177/0018720811427034.

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: An integrative review. *Theoretical Issues in Ergonomics Science, 8*(4), 277–301. https://doi.org/10.1080/14639220500337708

Mamonov, S., & Koufaris, M. (2018). The effects of IT-related attributional style in voluntary technology training. *Information Systems Management, 35*(3), 220–233. https://doi.org/10.1080/10580530.2018.1477302

McMahan, I. D. (1973). Relationships between causal attributions and expectancy of success. *Journal of Personality and Social Psychology, 28*(1), 108–114. https://doi.org/10.1037/h0035474

Mednick, S. (1962). The associative basis of the creative process. *Psychological Review, 69*(3), 220–232. https://doi.org/10.1037/h0048850

Mezulis, A. H., Abramson, L. Y., Hyde, J. S., & Hankin, B. L. (2004). Is there a universal positivity bias in attributions? A meta-analytic review of individual, developmental, and cultural differences in the self-serving attributional bias. *Psychological Bulletin, 130*(5), 711–747. https://doi.org/10.1037/0033-2909.130.5.711

Microsoft. (2020). *Visual studio 2017* (Version 15.9.30) [Computer software]. Microsoft. Retrieved July 27, 2021, from https://docs.microsoft.com/ja-jp/visualstudio/releasenotes/vs2017-relnotes#15.9.30

Miller, D. T., & Ross, M. (1975). Self-serving biases in the attribution of causality: Fact or fiction? *Psychological Bulletin, 82*(2), 213–225. https://doi.org/10.1037/h0076486

Moon, Y., & Nass, C. (1998). Are computers scapegoats? Attributions of responsibility in human–computer interaction. *International Journal of Human-Computer Studies, 49*(1), 79–94. https://doi.org/10.1006/ijhc.1998.0199

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*(1), 81–103. https://doi.org/10.1111/0022-4537.00153

Peterson, D. K., Kim, C., Kim, J. H., & Tamura, T. (2002). The perceptions of information systems designers from the United States, Japan, and Korea on success and failure factors. *International Journal of Information Management, 22*(6), 421–439. https://doi.org/10.1016/S0268-4012(02)00033-6

Polman, R., Rowcliffe, N., Borkoles, E., & Levy, A. (2007). Precompetitive state anxiety, objective and subjective performance, and causal attributions in competitive swimmers. *Pediatric Exercise Science, 19*(1), 39–50. https://doi.org/10.1123/pes.19.1.39

Rascle, O., Le Foll, D., Charrier, M., Higgins, N. C., Rees, T., & Coffee, P. (2015). Durability and generalization of attribution-based feedback following failure: Effects on expectations and behavioral persistence. *Psychology of Sport and Exercise*, *18*, 68–74. 10.1016/j.psychsport.2015.01.003.

Reeves, B., & Nass, C. I. (2001). *The media equation: How people treat computers, television, and new media like real people and places* (H. Hosoma, Trans.). Shoeisha (Original work published 1996).

Sakamoto, S., & Kambara, M. (1998). A longitudinal study of the relationship between attributional style, life events, and depression in Japanese undergraduates. *The Journal of Social Psychology, 138*(2), 229–240. https://doi.org/10.1080/00224549809600374

Salem, M., Eyssel, F., Rohlfing, K., Kopp, S., & Joublin, F. (2013). To err is human (−like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics, 5*(3), 313–323. https://doi.org/10.1007/s12369-013-0196-9

Serenko, A. (2007). Are interface agents scapegoats? Attributions of responsibility in human–agent interaction. *Interacting with Computers, 19*(2), 293–303. https://doi.org/10.1016/j.intcom.2006.07.005

Sundar, S. S., & Nass, C. (2000). Source orientation in human-computer interaction: Programmer, networker, or independent social actor. *Communication Research, 27*(6), 683–703. https://doi.org/10.1177/009365000027006001

Thurman, N., Moeller, J., Helberger, N., & Trilling, D. (2019). My friends, editors, algorithms, and I: Examining audience attitudes to news selection. *Digital Journalism, 7*(4), 447–469. https://doi.org/10.1080/21670811.2018.1493936

Titova, L., & Sheldon, K. M. (2019). Why do I feel this way? Attributional assessment of happiness and unhappiness. *The Journal of Positive Psychology, 14*(5), 549–562. https://doi.org/10.1080/17439760.2018.1519081

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine, 25*(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics, 106*(4), 1039–1061. https://doi.org/10.2307/2937956

van der Woerdt, S., & Haselager, P. (2019). When robots appear to have a mind: The human perception of machine agency and responsibility. *New Ideas in Psychology, 54*, 93–100. https://doi.org/10.1016/j.newideapsych.2017.11.001

Waldrop, M. M. (2015). Autonomous vehicles: No drivers required. *Nature, 518*(7537), 20. https://doi.org/10.1038/518020a

Wang, H., & Hall, N. C. (2018). A systematic review of teachers' causal attributions: Prevalence, correlates, and consequences. *Frontiers in Psychology, 9*, 2305. https://doi.org/10.3389/fpsyg.2018.02305

Weiner, B. (1979). A theory of motivation for some classroom experiences. *Journal of Educational Psychology, 71*(1), 3–25. https://doi.org/10.1037/0022-0663.71.1.3

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review, 92*(4), 548–573. https://doi.org/10.1037/0033-295X.92.4.548

Weiner, B. (2010). The development of an attribution-based theory of motivation: A history of ideas. *Educational Psychologist, 45*(1), 28–36. https://doi.org/10.1080/00461520903433596

Weiner, B., Nierenberg, R., & Goldstein, M. (1976). Social learning (locus of control) versus attributional (causal stability) interpretations of expectancy of success. *Journal of Personality, 44*(1), 52–68 https://psycnet.apa.org/doi/10.1111/j.1467-6494.1976.tb00583.x

Wu, J., & Liu, D. (2007). The effects of trust and enjoyment on intention to play online games. *Journal of Electronic Commerce Research, 8*(2), 128–140.

Yan, W., & Gaier, E. L. (1994). Causal attributions for college success and failure: An Asian-American comparison. *Journal of Cross-Cultural Psychology*, *25*(1), 146–158. 10.1177/0022022194251009.

Yoshida, T., Kojo, K., & Kaku, H. (1982). A study on the development of self-presentation in children. *Japanese Journal of Educational Psychology, 30*(2), 30–37. https://doi.org/10.5926/jjep1953.30.2_120

Zuckerman, M. (1979). Attribution of success and failure revisited, or: The motivational bias is alive and well in attribution theory. *Journal of Personality, 47*(2), 245–287. https://doi.org/10.1111/j.1467-6494.1979.tb00202.x