Racism, Responsibility and Autonomy in HCI: Testing Perceptions of an AI Agent

Joo-Wha Hong

Dmitri Williams

University of Southern California Annenberg School for Communication and Journalism

3502 Watt Way, Los Angeles, CA 90089

joowhaho@usc.edu

Racism Blame and Autonomy in HCI: Testing Perceptions of an AI Agent

Abstract

This study employs an experiment to test subjects' perceptions of an artificial intelligence (AI) crime-predicting agent that produces clearly racist predictions. It used a 2 (human crime predictor/AI crime predictor) x 2 (high/low seriousness of crime) design to test the relationship between the level of autonomy and responsibility for the unjust results. The seriousness of crime was manipulated to examine the relationship between the perceived threat and trust in the authority's decisions. Participants (N=334) responded to an online questionnaire after reading one of four scenarios with the same story depicting a crime predictor unjustly reporting a higher likelihood of subsequent crimes for a black defendant than for a white defendant for similar crimes. The results indicate that people think that an AI crime predictor has significantly less autonomy than a human crime predictor. However, both the identity of the crime predictor and the seriousness of the crime showed insignificant results on the level of responsibility assigned to the predictor. Also, a clear positive relationship between autonomy and responsibility was found in both human and AI crime predictor scenarios. The implications of the findings for applications and theory are discussed.

Keywords: Attribution Theory, CASA, Predictive Policing, Racism, Artificial Intelligence

The movie *Minority Report* (2002) famously depicted a crime-free society that used a predictive policing system, identifying some members of society as likely to commit a crime and arresting them pre-emptively. The movie raised serious ethical concerns about surveillance, technology and fairness (Gad & Hansen, 2013). How can someone be guilty of a crime they haven't yet committed? As with most science fiction, the moral dilemma seems farfetched, given that we don't predict and sentence individuals before they act. And yet the fictional society depicted in the movie is already possible on one level, given the rise of artificial intelligence (AI) and big data to predict crime. Although we do not sentence the not-yet criminal, there have been attempts to build programs that predict future crimes by area or individual. An example is Predpol, which identifies potentially high crime regions using the theory that criminals repeat crimes in previous crime locations (Aradau, Blanke, Kaufmann, & Jeandesboz, 2017). Another is the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), a recidivism risk assessment instrument designed by Northpointe Institute for Public Management (Eaglin, 2017). COMPAS uses a micro-level approach by assessing each individual who has a criminal record and predicting whether the person is likely to commit a future crime. However, there are ethical concerns regarding the implication of a program that labels individuals as a potential criminal.

Behavior prediction has become central in AI development and applied to many fields, including crime prevention (Abbasi, Lau, & Brown, 2015). However, when predicting crime, there is no evidence that an AI predictor will make higher or lower accuracy predictions than human predictors (Dressel & Farid, 2018). And, of course, there is a crucial ethical implication to the technology. Even if a system is highly accurate, should it be used? The social costs of false positives and negatives are substantial, and conflict with the basic tenets of most Western judicial systems, which favor the presumption of innocence, let alone

the presumption of no future criminal action (Ashworth, 2011; Hardyns & Rummens, 2018). Any prediction—whether from a human or an AI—creates a large cost for those who are labeled as guilty, or likely to be guilty, in the future. This assume the prediction is accurate. If it is not, the person is doubly victimized—once for an action they have not taken, and once for an action they are unlikely to take. While the ethical implications of the technology are critical, this paper is focused on the empirical, specifically on our attitudes towards this rapidly developing technology, and how they may be impacted by our cultural context.

As a socially constructed technology, it is crucial that we examine predictive AI as an artifact of a particular time and place (Pinch & Bijker, 1999). Crime has long been a policy area that reflects our hopes and fears, sometimes based on our culture rather than on an objective assessment of risks (Glassner, 1999). Prejudices aside, lower crime rates have universal appeal. Thus, it is unsurprising to learn that there are promising views of the efficiency of the program and its application, and a general hope that its use will results in reduced crime rates (Duwe & Rocque, 2017; Land, 2017). However, others argue the predictive policing program is inevitably biased because it uses crime data that are reported from heavily policed regions, which leads to an overrepresentation of the social minorities who live in such areas (Kirkpatrick, 2017). Therefore, it is valuable to focus not just on the accuracy of the algorithm, but the public reception and understanding of the emerging technology. Knowing this will help guide policymakers to develop systems that are not merely good, but also fit within a context of values as acceptable and socially just (Weizenbaum, 1976; Fox, Yamagata, Najaka, & Soulé, 2018).

Drawing on Expectation-Disconfirmation Theory (EDT), Venkatesh and Goyal (2010) argue that expectations and perceived performance are crucial when deciding whether to adopt or accept a technology. In the case of an AI crime predictor, people are likely to

expect a level of fairness and justice, as they would from any aspect of their criminal justice system. Regularly biased outcomes would confound that basic expectation. However, in practice, an AI is merely a tool and an algorithm, and as such it is unlikely to be perfect. Much like humans in the criminal justice system, an AI will be fallible. The central question here is how people will react when an AI crime predictor fails to fulfill their expectations—and, whether that reaction would be different from the case of a human crime predictor. In other words, this study investigates the level of mistrust toward an AI crime predictor and compares it to the level of mistrust toward a human counterpart with the same misconduct.

AI and racism

Just as Weizenbaum (1976) anticipated biased algorithms, implementations of AI have shown group-based implications including many cases where race and gender have been a factor. The issue started to gain public attention with the advent of the Microsoft twitter chat bot, Tay, which notably used offensive language in 2016 (Beran, 2018). Recently, Buolamwini and Gebru noted that the efficiency of facial recognition programs varies based on race and gender (2018), finding that commercial face recognition was the most efficient in identifying lighter-skinned males and the least efficient when detecting darker-skinned females (Lohr, 2018). Buolamwini argues that this is a direct result of the engineers who code the program being mostly white, with the result that the training datasets mainly consist of white faces (Breland, 2017).

Another case of group-based differences in AI is when a search engine provides ethnically or sexually discriminating results. Noble (2018) argues that Google's search algorithm delivers systematically more negative results for black females, such as providing more obscene results when searching "black girls" compared to "white girls." Noble finds the

reason for such results stems from the bias of the people who built the algorithm. Because of the profit motive, the coders are incentivized to ensure that some results are privileged over others. In other words, Noble claims that Google programmers manipulated and altered an unbiased algorithm to maximize profits. Thus, the code itself is embedded in an economic system, not one built for social justice. The result may be that the consumption of those results is profitable, thus justifying and strengthening both the business practice, as well as the resulting existing sexism and racism caused by the results. One common argument from both Noble and Buolamwini is that AI programs themselves are neutral, but the people who built them are biased by either their own identity-based perspectives or by a profit motive.

The last example is directly related to the study at hand. The COMPAS recidivism prediction AI, which is used in this study below, has been shown to make racist decisions with regard to people with criminal records. An investigation from ProPublica revealed that COMPAS was biased against black prisoners and assigned them a higher probability of committing future crimes compared to white prisoners with a similar crime record. Also, after looking into the risk scores of about 7,000 arrested people in Florida and their crime relapses over the next two years, the investigation found that only 20% of COMPAS predictions were correct (Angwin, Larson, Kirchner, & Mattu, 2016). However, ProPublica reported only the results COMPAS produced and did not explain how the program generated these outcomes, since Northpointe, the company that built COMPAS, did not disclose their calculations in the program just as most private companies don’t share the details of their algorithms (Perel & Elkin-Koren, 2017). While some people are skeptical about this technology, there are others who support COMPAS and think it is unfairly framed as a biased tool (Zhang, Roberts, & Farabee, 2014; Flores, Bechtel, & Lowenkamp, 2016). Thus, whether COMPAS makes biased or neutral decisions that are merely viewed as biased, there are people who are

predisposed to have a strong feeling about each. Therefore, instead of investigating whether COMPAS is biased, this study will focus on how people would react to an AI crime-predicting system making unjust outcomes, particularly with regard to the violation of racial equality.

The autonomy of artificial intelligence

The purpose of creating AI is to produce computer programs that function autonomously to find the best possible answers to questions (Russell & Norvig, 2010). Studies investigating reactions to the autonomy of machines have found that perception comes from two dissimilar feelings: trustworthiness and threat. On the one hand, a study that found the autonomy of AI agent influences the perception of the agent's trustworthiness (Lee, Kim, Lee, & Shin, 2015). On the other hand, Złotowski, Yogeeswaran, & Bartneck (2017) found that people see autonomous machines as more of a threat. However, regardless of how the autonomy of artificial intelligence is perceived, there is a view that AI still possesses limitations to be deemed as having its own free will (Krausová & Hazan, 2013; Cevik, 2017). Thus, it is expected that people believe AI crime predictors have less autonomy compared to human crime predictors.

H1. A lower rating on autonomy will be given to a crime predictor in the AI crime predictor scenario compared to the human crime predictor scenario.

Attribution theory

Despite the reason why AI and computers make such unethical decisions, this study is

to see how people would react to the information that is the technologies are being unfair. Particularly, this study mainly investigates the case of artificial intelligence, which is often perceived as an autonomous technology (Weng et al., 2001; Zgrzebnicki, 2017), is culpable by comparing it with the same case with a human counterpart. To see the relationship between the level of blame and the identity of a crime predictor, attribution theory is chosen as another theoretical framework of this study. Attribution theory explains how people find a causal relationship to make judgments about an event by focusing on how "outcome dependent affect," "causal antecedents," "causal ascriptions," and the "causal dimensions" of an event are processed in order, which lead to psychological and behavioral consequences (Weiner, 2010; Fiske & Taylor, 1991). Therefore, this theory is often used to inquire how blame is attributed when events happen, such as accidents or crises, and what factors influence the process of blame (Jeong, 2009; Richard, 2014). As now AI performs human-like actions and behaviors, including crime predictions, there have been studies inquiring about how people blame AI when this technology causes blameworthy outcomes. For instance, Shank and DeSanti (2018) conducted a study based on attribution theory using actual events to look at how people blame AI when it commits moral violations. The study found that people tend to blame artificial intelligence more and external factors less, at marginal significance, when there is more specific information about the algorithm of the AI. Even though this study shares a similarity with Shank and DeSanti's study, as both are about blame toward artificial intelligence, this study focuses on how the level of blame varies by directly comparing reactions to the same action done by different "identities" of wrongdoers, human and artificial intelligence.

Furthermore, AI is expected to be blamed more due to an emotional distance coming from a different identity, which can be explained by the attribution theory. Defensive attribution explains the inclination of attributing more responsibility when there are fewer

similarities found between the blamer and the target of the blame, both personally and situationally (Burger, 1981; Shaver, 1970). From this perspective, more blame to AI for its racist decisions is expected if people identify less with AI than they do with a human. Thus, this study assumes that people will blame a crime predictor more when it is artificial intelligence.

H2. Participants will attribute more responsibility to a crime predictor in the AI crime predictor scenario than the human crime predictor scenario.

Seriousness of crime and acceptance of authority

Finally, it is expected that crime predictors will receive less blame if a defendant in given scenarios committed a more serious crime. Because the severity of the crime is found to have a positive relationship with perceived dangerousness (Sanderson, Zanna, & Darley, 2000), crimes with more serious outcomes may induce more trust in decisions by an authority. Authoritarianism derives precisely from the relationship between the perception of threat and the acceptance of authority (Feldman & Stenner, 1997). Sales (1973) found in archival data that threat is a cause of a positive attitude toward authority and acquiescence to an authority figure. Similarly, people support authoritarian crime controls when exposed to the news with serious crime (Krause, 2014). Yet, there have been few studies that examined the interaction between perceived threat and the acceptance of authority in the context of human-computer interactions, especially seeing computers as an authority. It is expected that, if the past crime that a defendant committed is serious, people will trust and support a crime predictor more and accept the risk scores rather than perceiving the scores as racist and problematic. Therefore, it is presumed that a crime predictor will be blamed less if people are

exposed to crime predictor scenarios with more serious crimes.

H3. Participants will blame a crime predictor less when the crime in the scenario is more serious.

Computers Are Social Actors (CASA)

Computers Are Social Actors (CASA) is a theoretical framework for examining how people perceive computers. According to Nass and Moon (2000), people tend to perform regular social behaviors in their human-computer interactions, as if the computer was another person. In CASA, these social norms are applied mindlessly as a heuristic shortcut, but have the effect of impacting our opinions of computers. Importantly, people tend to see computers as relatively autonomous, and do not focus on their nature as coded, artificial constructs, with parameters and algorithms decided upon and created by some other very separate person or persons. In a human-computer interacting setting, people tend to see machines as independent entities with their own independent sources of information (Sunder & Nass, 2000). As a result, CASA suggests that people consider a computer as an autonomous social entity. Hence, the framework is often employed when explaining why behaviors in human-machine interactions are similar to practices in interpersonal interaction.

In regular human-human relationships, we also consider the autonomy of others. In cases where we are assessing another's wrongdoing, we factor this in. As a general rule, when someone is perceived to be autonomous, we are more likely to assign them a higher level of blame for their actions, and vice versa (Nahmias, Shepard, & Reuter, 2014; Sankowski, 1992; Woolfolk, Doris, & Dailey, 2006). Russell, McAuley, & Tarico (1987) found that autonomy was a key predictor for responsibility in both successes and failures. Similarly, another study

found that people were more willing to give severe punishments to more autonomous people (Graham, Weiner, & Zucker, 1997). Based on CASA, it is reasonable to expect that the same patterns found in human-human interactions will be found in human-computer ones. A study using a delivery robot assisting people performing their task found that people attribute both blame and credit more to robots with higher autonomy (Kim & Hinds, 2006). Therefore, a positive relationship between autonomy and the level of responsibility is expected in both human and AI crime predictor scenarios, i.e. the more autonomous either the human or the AI are seen, the more they will be blamed for an unjust result.

H4. The amount of responsibility assigned to the predictor's unjust decision positively related to its perceived autonomy.

H5. The relationship between the autonomy of the crime predictor and the responsibility assigned to the crime predictor for both human and AI crime predictor scenarios will be similar.

Methods

In order to test the hypotheses, a 2x2 experiment was designed and conducted, varying both the kind of predictor (human or AI) as well as the seriousness of the crime (high or low). The dependent variables for this study are the perception of the autonomy of the crime predictor and the responsibility assigned to the predictor.

Participants

Amazon Mechanical Turk (MTurk) was used to recruit participants and incentive of $1 was awarded to each participant upon survey completion. Excluding participants who omitted any question in the survey was followed, leaving 334 participants from 353 initially recruited people. The youngest participant was 19 years old, while the oldest participant was 87 years

old (M=35.04, SD=12.64). Additionally, 149 participants were male, and 185 participants were female.

Procedures

Participants who agreed to participate in the study were given reading material based on an actual story retrieved from a news article in ProPublica, saying that black defendants receive a higher risk rating from a crime predictor, which indicates that the person is more likely to commit subsequent crimes compared to white defendants with more serious crime records who conducted a similar crime. Also, histograms that show black defendants have received higher risk scores than white defendants were provided (Angwin, Larson, Kirchner, & Mattu, 2016). This article is overtly framed around the racial injustice of the prediction. Versions of the story was altered into four different scenarios that fit a 2 (human or AI predictor) x 2 (high seriousness or low seriousness) experimental design. For human predictor scenarios, the language stated, “a person who is specialized to predict the possibility of subsequent offenses or crimes” and “the person who rated the scores.” This section was bolded in order to highlight the identity of the crime predictor as human. For the for AI predictor scenarios, the language stated, “a computer program with an algorithm that predicts the possibility of subsequent offenses or crimes” and “the computer crime predictor”. These again were bolded to stress that the identity of the crime predictor was artificial intelligence.

For the severity of the crime, the low seriousness conditions depicted criminals who conducted petty theft and shoplifting, while the high seriousness crime depicted criminals who drove under the influence (DUI). These were the same crimes dealt with in the original article. Other than these manipulations, the stimuli were equivalent in all four conditions.

Using Qualtrics©, an online survey tool, the four scenarios were randomly and evenly distributed to the participants.

After reading a given scenario, the participants were asked to answer two sets of questions (three items each), which used a seven-point Likert scale (from Strongly Disagree to Strongly Agree) and were edited from the causal dimension scale (CDS) (Russell, 1982). The original CDS scale consists of three parts; (a) controllability, which is similar to the measurement of autonomy in this study, (b) locus of control, which is similar to the responsibility assigned to the crime predictor in this study, and (c) stability, which refers to whether the event would stably and repeatedly occur. Because provided reading materials imply that the crime predictor giving higher risk scores to black defendants than white defendants has been stably occurred over times, the stability of the event was not used for this study since it is expected not to produce significant outcomes.

Measures

Autonomy of the crime predictor. Because the information given about the crime predictor explains the identity (human x artificial intelligence), the scores of this measure show how people differentiate the autonomy of human and artificial intelligence. This dependent variable was measured with a three item scale consisting of these questions: (a) Was the crime predictor fully self-controllable when making the decision?; (b) Was the decision intended by the crime predictor?; (c) Is the crime predictor responsible for making the decision? This three-question measure reached statistical significance (α=0.70). Higher scores indicate that the crime predictor made the decision independently from the influence of others

Responsibility of the crime predictor. Since this measure is also based on the identity of the crime predictor, the scores reflect how people differentiate the characteristics of human and artificial intelligence in terms of assessing the responsibility. This is the dependent variable for the rest of the hypotheses. A three-item scale was used, consisting of these questions: (a) Is the decision something that reflects a characteristic of the crime predictor?; (b) Was the decision made due to an intrinsic aspect of the crime predictor?; (c) Is the decision something about the crime predictor itself? This three-question measure showed good reliability (α=0.78). Higher scores indicate that the crime predictor is responsible for the decision.

In order to evaluate the effect of the identity of the crime predictor on the perception of autonomy, a t-test was conducted. Also, a two-way analysis of variance (ANOVA) was carried out on the scores of the responsibility assigned to the crime predictor to evaluate the influence of identity and the seriousness of crimes. Finally, two sets of simple linear regressions were conducted to compare the relationship of autonomy of the crime predictor and its assigned responsibility between the human crime predictor scenario and the AI crime predictor scenario.

Results

To verify the efficacy of the manipulations, the responses from “Do you think the type of crime mentioned in the reading is a serious crime?” and “To which extent do you think the crime predictor has human characteristics (not humanistic)?” were analyzed and compared between different scenarios using a t-test. The efficacy of the “seriousness of crime” manipulation showed a significant outcome between low seriousness (M=5.03,

SD=1.33) and high seriousness (*M*=5.54, *SD*=1.28) crime scenarios; $t(332) = 3.54$, $p < .001$, and the efficacy of the "human - AI distinction" manipulation also showed a significant outcome between a human crime predictor (*M*=4.62, *SD*=1.98) and an AI crime predictor (*M*=4.11, *SD*=1.92); $t(332) = -2.47$, $p = .014$. These results indicate that participants distinguished different seriousness of crimes, and the identity between human and AI crime predictors.

A t-test was conducted For H_1, to test whether people think human crime predictors are more autonomous than an AI counterpart. The dependent variable for the t-test was the autonomy of the crime predictor. The data analyzed with the t-test demonstrated the influence of the crime detector's identity at the level of perceived autonomy of the crime detector. A higher number indicates a crime predictor possessing more self-control when making the decision. From the analyzed data, hypothesis 1 was supported, which argues participants think an AI crime predictor has less autonomy compared to a human crime predictor when making a racist decision. Based on the results of the t-test, there was a small but statistically significant effect of the identity of the crime predictor on its autonomy, $p<0.05$ [$t(332)=-2.95$, $p=0.003$, $d=0.322$]. The autonomy of the predictor was rated lower in the AI scenario (*M*=4.61, *SD*=1.40) than the human one (*M*=5.04, *SD*=1.27).

The two-way ANOVA was conducted to see the level of blame based on the identity of the crime predictor and the seriousness of the crime. Levene's test was conducted to assess the equality of variances, and the result of the test rejects the homogeneity of variances; $F(3, 330)=0.86$, $p=0.46$. The data analyzed with this two-way ANOVA showed the influence of the identity of the crime detector and the seriousness of crime on the level of the blame against the crime detector for its decision. A higher number indicates more responsibility was attributed to internal factors, which means more blame to the crime predictor, and a smaller

number indicates the attribution of responsibility to external factors, which means less blame to the crime predictor.

Two hypotheses in regard to attributed responsibility were rejected. The second hypothesis was that participants attribute more blame of a racism decision to a human crime predictor than to an AI crime predictor, and the third hypothesis presumed that participants direct less blame toward the crime predictor when the crime in a scenario is more serious. The results from two-way ANOVA showed an insignificant outcome for the effects of the different identity of a crime predictor [$F(1, 330) = 0.92$, $p = .34$] and the seriousness of the crime on the attribution of responsibility for the decision [$F(1, 330) = 0.17$, $p = .68$]. These results reject the hypotheses that the identity of crime predictors and the seriousness of crimes will influence the level of blame. Additionally, an insignificant result was found from the two-way interaction of identity of crime predictor x seriousness of crime on the level of attributed responsibility [$F(1, 330) = .462$, $p = .50$]. Table 1 shows the descriptive statistics for analyzed data on the level of blame.

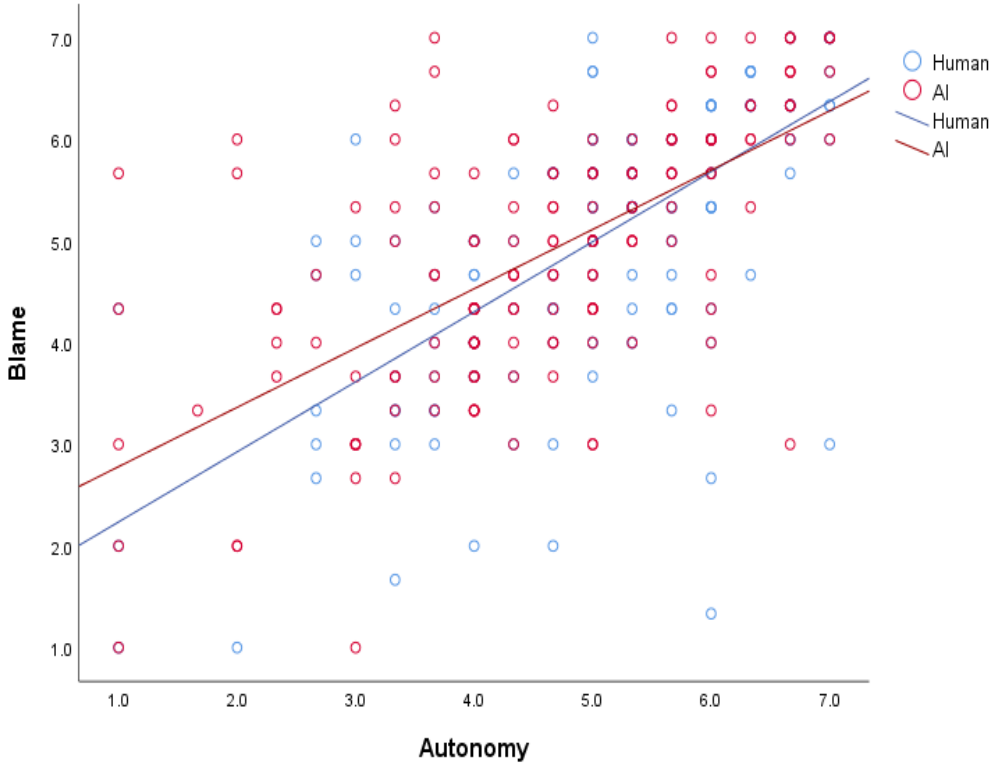
Table 1. Descriptive statistics for the level of blame based on crime predictor identity and seriousness of the crime

Seriousness of	**Human Crime Predictor**			**AI Crime Predictor**		
Crime	**M**	**SD**	**N**	**M**	**SD**	**N**
High	5.00	1.38	88	4.96	1.35	85
Low	5.04	1.24	79	4.81	1.22	82

Based on results from the t-test and ANOVA, it was found that the influence of the identity of a crime predictor, whether human or artificial intelligence, was significant on the perceived autonomy of the crime predictor in the decision-making but not significant on the attributed responsibility. In order to have an in-depth understanding of the relationship

between the autonomy of the crime predictor and the level of blame, two sets of regression analyses were conducted, one set using data only from the AI crime predictor scenario and the other set using data only from the human crime predictor scenario. Simple linear regressions were conducted to see the influence of the level of autonomy on the level of blame in the AI crime predictor scenario and the human crime predictor scenario, respectively. For the AI crime predictor scenario, a significant effect was found [$F(1, 165)=112.08$, $p<0.001$], with $r^2=40.5$ ($\beta=0.64$). For the human crime predictor scenario, a significant effect was also found [$F(1, 165)=134.14$, $p<0.001$], with $r^2=44.8$ ($\beta=0.67$). The results from two regression analyses show positive relationships between the level of autonomy and blame in both human-crime predictor cases and AI-crime predictor cases, which support the fourth hypothesis. The results also suggest that the relationship between the level of autonomy and blame toward the crime predictor is similar across the two scenarios, supporting the fifth hypothesis. Figure 1 depicts the relationship between the autonomy of the crime predictor and the level of blame to the crime predictor for its decision, based on the identity of the crime predictor.

Figure 1. The regression analysis of the relationship between the autonomy of the crime predictor and the level of blame



Discussion

The results of this experimental study show that there are strong correlations between autonomy and blame, which are found in both human and AI crime predictor cases. These outcomes support CASA theory because they illustrate that people blame AI at similar rates as they blame humans for a racist decision. Because these outcomes are correlations, further studies using an experiment to inquire about causal relationships between autonomy and the level of blame in human-computer interacting settings are necessary. Also, even though they acknowledge that AI is less autonomous when making decisions, the study shows that the level of blame on artificial intelligence is similar to the level of blame on human actors. Based on the EDT perspective, this indicates that people had a similar level of expectation of

fair judgments from a crime predictor regardless of its identity (human or AI). CASA theory was first derived based on studies that measured the attitudes of people who interacted with a computer directly (Nass, Moon, & Green, 1997; Moon, & Nass, 1996). This study extends CASA by finding that it is also applicable for indirect interactions with machines, such as assessing their performance from more of a distance. Moreover, recent CASA-based studies have been conducted by focusing solely on participants' attitudes to various types of machine without the consideration how the participants would react to human agents, which led to an experiment design that compared reactions to computers with and without human characteristics (Carolus, Muench, Schmidt, & Schneider, 2019; Edwards, Edwards, Stoll, Lin, & Massey, 2019). To overcome the shortcoming of the previous studies, this study tested CASA further by finding a similarity between human-human interaction and human-computer interaction by comparing them directly. This suggests that the method used here is a fruitful one for future research.

On the other hand, the result shows that the attribution theory may not be suitable to explain public blame against AI. There are two approaches to explain the discordance between the level of autonomy and blame: there is no relationship between autonomy and blame, or a ceiling effect. Because the relationship between autonomy and blame was found in both AI and human crime predictor scenarios using linear regression analyses, it is more likely that the result is due to a ceiling effect, which is a measurement limitation due to an extreme skewness (Ho & Yu, 2015; Salkind, 2010). In other words, a crime predictor predicting higher plausibility of subsequent crimes to a black defendant than a white defendant without explainable reason is exceptionally unacceptable to participants that the identity of the crime predictor was not carefully concerned. The reason for the discordance can be confirmed through similar future studies with slightly more publicly acceptable cases.

Other than the ceiling effect, this study has a few limitations. First, the attitude toward artificial intelligence was not measured before the experiment. Hence, it is not clear whether preexisting attitudes toward AI has any influence on the blame toward an AI crime predictor. Similarly, the understanding of artificial intelligence technology was also not measured, and information about AI was not provided, though we can assume that there was even understanding through random assignment.

One aspect of the results of this study, which AI industries should be aware of, is that people's expectations of AI labor are not any lower than they are for human labor, particularly in terms of racial discrimination. In other words, even though people are aware that AI may have fewer intentions and lower autonomy on any outcome it produces, they have the same expectations of fairness. Moreover, because people acknowledge that the AI program is less autonomous than humans, blame will likely shift more to programmers and the company, and less to the program itself. In other words, even if people understand computers or that AI has less intentionality, this does not mean the customers are naïve. Thus, companies in the AI industry should treat unethical outputs from their AI products with an equal level of seriousness compared to unethical behaviors done by their human representative.

On the other hand, there should also be a consideration of these results from an academic perspective. Unexpected adverse outcomes may interrupt, or even cease, technology developments and related studies, just like shutting down Tay, a chatbot developed by Microsoft, due to its inappropriate and unethical comments (Wakefield, 2016). Putting efforts into AI's making ethical choices is inevitable since artificial intelligence is expected to both provide benefits and constrain our lives by making decisions that have been made by humans (Casacuberta & Guersenzvaig, 2018; Diakopoulos, 2014). Also, more studies and attention on how people would perceive and react to unexpected consequences

from AI are necessary because the technology is still unfolding and evolving.

Finally, there should be more research on ethnicity issues and racism in terms of recent technologies, particularly robots and artificial intelligence. The reason more studies are required in this field is that such technologies with biases may emphasize and reinforce the wrong idea if they are commercialized and permeate into our lives (Howard & Borenstein, 2017). The AI industry is showing fast growth, with 70% growth in business value over the past year, which indicates there will be more AI products that people will use and interact with (Coleman, L. D., 2018). This fast commercialization of AI technologies means it would be detrimental if the algorithm in AI products were placed into the wrong hands. Thus, the efforts to inquire about racism is needed not only in interpersonal interaction settings but also in the field of human-computer interaction.

Reference

Abbasi, A., Lau, R., & Brown, D. (2015). Predicting Behavior. *Intelligent Systems, IEEE*, 30(3), 35–43. https://doi.org/10.1109/MIS.2015.19

Angwin, J., Larson, J., Kirchner, L., & Mattu, S. (2016). Machine Bias. *ProPublica*. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Aradau, C., Blanke, T., Kaufmann, M., & Jeandesboz, J. (2017). Politics of prediction: Security and the time/space of governmentality in the age of big data. *European Journal of Social Theory*, 20(3), 373-391.

Ashworth, A. (2011). The Unfairness of Risk-Based Possession Offences. *Criminal Law and Philosophy*, 5(3), 237–257. https://doi.org/10.1007/s11572-011-9112-2

Beran, O. (2018). An Attitude Towards an Artificial Soul? Responses to the "Nazi Chatbot." *Philosophical Investigations*, 41(1), 42–69. https://doi.org/10.1111/phin.12173

Breland, A. (2017, December 4). How white engineers built racist code – and why it's dangerous for black people. *The Guardian*. Retrieved from: https://www.theguardian.com/technology/2017/dec/04/racist-facial-recognition-white-coders-black-people-police

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on Fairness, *Accountability and Transparency* (pp. 77-91).

Burger, J. (1981). Motivational Biases in the Attribution of Responsibility for an Accident: A Meta-Analysis of the Defensive-Attribution Hypothesis. *Psychological Bulletin*, 90(3), 496–512. https://doi.org/10.1037/0033-2909.90.3.496

Carolus, A., Muench, R., Schmidt, C., & Schneider, F. (2019). Impertinent mobiles - Effects

of politeness and impoliteness in human-smartphone interaction. *Computers in Human Behavior*, 93, 290–300. https://doi.org/10.1016/j.chb.2018.12.030

Casacuberta, D., & Guersenzvaig, A. (2018). Using Dreyfus' legacy to understand justice in algorithm-based processes. *AI & Society*, 1-7.

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, 29(6), 2156-2160 .

Cevik, M. (2017). Will It Be Possible for Artificial Intelligence Robots to Acquire Free Will and Believe in God?. *Beytulhikme: An International Journal of Philosophy*, 7(2).

Coleman, L. D. (2018, May 31). Inside Trends And Forecast For The $3.9T AI Industry. *Forbes*. Retrieved from https://www.forbes.com/sites/laurencoleman/2018/05/31/inside-trends-and-forecast-for-the-3-9t-ai-industry/#125bd7a42c86

Diakopoulos, N. (2014). Algorithmic Accountability. *Digital Journalism*, 3(3), 1-18.Dressel, J. & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1), Eaao5580-eaao5580.

Duwe, G. & Rocque, M. (2017). Effects of Automating Recidivism Risk Assessment on Reliability, Predictive Validity, and Return on Investment. *Criminology & Public Policy*, 16(1), 235-269.

Eaglin, J. (2017). Constructing Recidivism Risk. *Emory Law Journal*, 67(1), 59-122.

Edwards, C., Edwards, A., Stoll, B., Lin, X., & Massey, N. (2019). Evaluations of an artificial intelligence instructor's voice: Social Identity Theory in human-robot interactions. *Computers in Human Behavior*, 90, 357–362. https://doi.org/10.1016/j.chb.2018.08.027

Feldman, S., & Stenner, K. (1997). Perceived Threat and Authoritarianism. *Political Psychology*, 18(4), 741–770. https://doi.org/10.1111/0162-895X.00077

Flores, A. W., Bechtel, K., & Lowenkamp, C. T. (2016). False Positives, False Negatives, and False Analyses: A Rejoinder to Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks. *Fed. Probation*, 80, 38.

Fiske, S.T. & Taylor, S.E. (1991) Social cognition (2nd ed.). New York: McGraw-Hill

Fox, D., Yamagata, H., Najaka, S., & Soulé, D. (2018). Improving Judicial Administration Through Implementation of an Automated Sentencing Guidelines System. Criminal Justice Policy Review, 29(5), 489–504. https://doi.org/10.1177/0887403416628603

Gad, C. & Hansen, L. (2013). A Closed Circuit Technological Vision: On Minority Report, event detection and enabling technologies. *Surveillance & Society*, 11(1/2), 148-162.

Glassner, B. (1999). *The culture of fear: Why Americans are afraid of the wrong things*. New York, Basic Books.

Graham, S., Weiner, B., & Zucker, G. (1997). An Attributional Analysis of Punishment Goals and Public Reactions to O. J. Simpson. *Personality and Social Psychology Bulletin*, 23(4), 331-346.

Hardyns, W., & Rummens, A. (2018). Predictive Policing as a New Tool for Law Enforcement? Recent Developments and Challenges. *European Journal on Criminal Policy and Research*, 24(3), 201–218. https://doi.org/10.1007/s10610-017-9361-2

Ho, A., & Yu, C. (2015). Descriptive Statistics for Modern Test Score Distributions: Skewness, Kurtosis, Discreteness, and Ceiling Effects. *Educational and Psychological Measurement*, 75(3), 365–388. https://doi.org/10.1177/0013164414548576

Howard, A., & Borenstein, J. (2017). The Ugly Truth About Ourselves and Our Robot Creations: The Problem of Bias and Social Inequity. *Science and Engineering Ethics*, 1–16. https://doi.org/10.1007/s11948-017-9975-2

Jeong, S. (2009). Public's Responses to an oil spill accident: A test of the attribution theory

and situational crisis communication theory. *Public Relations Review*, 35(3), 307–309. https://doi.org/10.1016/j.pubrev.2009.03.010

Karsh, B. T. (2004). Beyond usability: designing effective technology implementation systems to promote patient safety. *BMJ Quality & Safety*, 13(5), 388-394.

Kim, T., & Hinds, P. (2006). Who Should I Blame? Effects of Autonomy and Transparency on Attributions in Human-Robot Interaction. In Robot and Human Interactive Communication, 2006. ROMAN 2006. *The 15th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 80–85). IEEE. https://doi.org/10.1109/ROMAN.2006.314398

Kirkpatrick, K. (2017). It's not the algorithm, it's the data. *Communications of the ACM*, 60(2), 21–23. https://doi.org/10.1145/3022181

Krause, K. (2014). Supporting the Iron Fist: Crime News, Public Opinion, and Authoritarian Crime Control in Guatemala. *Latin American Politics and Society*, 56(1), 98–119. https://doi.org/10.1111/j.1548-2456.2014.00224.x

Krausová, A., & Hazan, H. (2013). Creating Free Will in Artificial Intelligence. *Beyond AI: Artificial Golem Intelligence*, 96.

Land, K. (2017). Automating Recidivism Risk Assessment. *Criminology & Public Policy*, 16(1), 231-233.

Lee, J., Kim, K., Lee, S., & Shin, D. (2015). Can Autonomous Vehicles Be Safe and Trustworthy? Effects of Appearance and Autonomy of Unmanned Driving Systems. *International Journal of Human-Computer Interaction*, 31(10), 682–691. https://doi.org/10.1080/10447318.2015.1070547

Lohr, S. (2018, February 12). Facial Recognition Works Best If You're a White Guy. *New York Times*, p. B.1.

Moon, Y., & Nass, C. (1996). How "Real" Are Computer Personalities?: Psychological Responses to Personality Types in Human-Computer Interaction. *Communication Research*, 23(6), 651–674. https://doi.org/10.1177/009365096023006002

Nahmias, E., Shepard, J., & Reuter, S. (2014). It's OK if "my brain made me do it": People's intuitions about free will and neuroscientific prediction. *Cognition*, 133(2), 502–516. https://doi.org/10.1016/j.cognition.2014.07.009

Nass, C. & Moon, Y. (2000). Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues*, 56(1), 81-103.

Nass, C., Moon, Y., & Green, N. (1997). Are Machines gender neutral? Gender-stereotypic responses to computers with voices. *Journal of Applied Social Psychology*, 27(10), 864–876. https://doi.org/10.1111/j.1559-1816.1997.tb00275.x

Noble, S. (2018). *Algorithms of Oppression*. NYU Press.

Oudeyer, P. (2017). Autonomous development and learning in artificial intelligence and robotics: Scaling up deep learning to human-like learning, *40*, e275. https://doi.org/10.1017/S0140525X17000243

Perel, M., & Elkin-Koren, N. (2017). Black Box Tinkering: Beyond Disclosure In Algorithmic Enforcement. *Florida Law Review*, 69(1), 181–221.

Pinch, T. and T. Bijker (1999). The Social Construction of Facts and Artifacts: Or How the Sociology of Science and the Sociology of Technology Might Benefit Each Other. In *The Social Construction of Technological Systems* (pp. 17-50). W. E. Bijker, Hughes, P. & Pinch, T. Cambridge, Massachusetts, The MIT Press.

Rickard, L. (2014). Perception of Risk and the Attribution of Responsibility for Accidents. *Risk Analysis*, 34(3), 514–528. https://doi.org/10.1111/risa.12118

Russell, D. (1982). The Causal Dimension Scale: A Measure of How Individuals Perceive

Causes. *Journal of Personality and Social Psychology*, 42(6), 1137-1145.

Russell, D., McAuley, E., & Tarico, V. (1987). Measuring Causal Attributions for Success and Failure: A Comparison of Methodologies for Assessing Causal Dimensions. *Journal of Personality and Social Psychology*, 52(6), 1248.

Russell, S., & Norvig, P. (2010). *Artificial intelligence: a modern approach* (3rd ed.). Upper Saddle River, N.J.: Prentice Hall.

Sales, S. M. (1973). Threat as a factor in authoritarianism. *Journal of Personality and Social Psychology*, 28, 44–57

Salkind, N. J. (2010). *Encyclopedia of research design Thousand Oaks*, CA: SAGE Publications, Inc. doi: 10.4135/9781412961288

Sanderson, C., Zanna, A., & Darley, J. (2000). Making the punishment fit the crime and the criminal: Attributions of dangerousness as a mediator of liability. *Journal Of Applied Social Psychology*, 30(6), 1137–1159. https://doi.org/10.1111/j.1559-1816.2000.tb02514.x

Sankowski, E. (1992). Blame and Autonomy. *American Philosophical Quarterly*, 29(3), 291–299.

Shaver, K. (1970). Defensive Attribution: Effects of Severity and Relevance on the Responsibility Assigned for An Accident. *Journal of Personality and Social Psychology*, 14(2), 101-113.

Shank, D. & DeSanti, A. (2018). Attributions of Morality and Mind to Artificial Intelligence after Real-World Moral Violations. *Computers in Human Behavior*. Advance online publication. https://doi.org/10.1016/j.chb.2018.05.014

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:10.1177/00936500002700600

Venkatesh, V., & Goyal, S. (2010). Expectation Disconfirmation and Technology Adoption: Polynomial Modeling and Response Surface Analysis. *MIS Quarterly*, 34(2), 281–303. https://doi.org/10.2307/20721428

Wakefield, J. (2016, March 25). Microsoft chatbot is taught to swear on Twitter. *BBC News*. Retrieved from: https://www.bbc.com/news/technology-35890188.

Weiner, B. (2010). The Development of an Attribution-Based Theory of Motivation: A History of Ideas. *Educational Psychologist*, 45, 28-36.

Weizenbaum, J. (1976). *Computer power and human reason : from judgment to calculation*. San Francisco: W. H. Freeman.

Weng, J., Mcclelland, J., Pentland, A., Sporns, O., Stockman, I., Sur, M., & Thelen, E. (2001). Artificial intelligence - Autonomous mental development by robots and animals. *Science*, 291(5504), 599–600.

Woolfolk, R., Doris, J., & Dailey, J. (2006). Identification, Situational Constraint, and Social Cognition: Studies in the Attribution of Moral Responsibility. *Cognition: International Journal of Cognitive Science*, 100(2), 283–301. https://doi.org/10.1016/j.cognition.2005.05.002

Zgrzebnicki, P. (2017). Selected Ethical Issues in Artificial Intelligence, Autonomous SystemDevelopment and Large Data Set Processing. *Studia Humana*, 6(3), 24–33. https://doi.org/10.1515/sh-2017-0020

Złotowski, J., Yogeeswaran, K., & Bartneck, C. (2017). Can we control it? Autonomous robots threaten human identity, uniqueness, safety, and resources. *International Journal of Human - Computer Studies*, 100, 48–54. https://doi.org/10.1016/j.ijhcs.2016.12.008

Zhang, S., Roberts, R., & Farabee, D. (2014). An Analysis of Prisoner Reentry and Parole Risk Using COMPAS and Traditional Criminal History Measures. *Crime & Delinquency*,

60(2), 167–192. https://doi.org/10.1177/0011128711426544