Computers in Human Behavior 125 (2021) 106944

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

AI, you can drive my car: How we evaluate human drivers vs. self-driving cars

Joo-Wha Hong [a,*], Ignacio Cruz [b], Dmitri Williams [a]

[a] University of Southern California, Annenberg School for Communication and Journalism, 3502 Watt Way, Los Angeles, CA, 90089, USA
[b] Northwestern University, School of Communication, 70 Arts Cir Dr, Evanston, IL, 60208, USA

ARTICLE INFO

Keywords:
Self-driving cars
Schema theory
Computers-are-social-actors
Attribution theory
Human-agent communication
Human-computer interaction

ABSTRACT

This study tests how individuals attribute responsibility to an artificial intelligent (AI) agent or a human agent based on their involvement in a negative or positive event. In an online, vignette experimental between-subjects design, participants (n = 230) responded to a questionnaire measuring their opinions about the level of responsibility and involvement attributed to an AI agent or human agent across rescue (i.e., positive) or accident (i.e., negative) driving scenarios. Results show that individuals are more likely to attribute responsibility to an AI agent during rescues, or positive events. Also, we find that individuals perceive the actions of AI agents similarly to human agents, which supports CASA framework's claims that technologies can have agentic qualities. In order to explain why individuals do not always attribute full responsibility for an outcome to an AI agent, we use Expectancy Violation Theory to understand why people credit or blame artificial intelligence during unexpected events. Implications of findings for practical applications and theory are discussed.

In simple terms, a self-driving car is a vehicle that uses an artificial intelligent (AI) system to evaluate information from sensors to make decisions while on the road (Kim, Na, & Kim, 2012; Vellinga, 2017). The deployment of these vehicles has been touted as a catalyst to revolutionize the future of transportation by decreasing accidents, reducing traffic, and improving the environment by lowering emissions (Rahman, Hamid, & Chin, 2017). The belief that self-driving cars will decrease accident rates is attributed to their limited reliance on human input, which implies that AI programming has the ability to outperform human judgment (Teoh & Kidd, 2017). The progress of automation technologies in cars within recent years has been significant, and these technologies are expected to be deployed on a mass scale within the next decade (Borraz, Navarro, Fernández & Alcover, 2018; Lee, Jung, Jung, & Shim, 2018).

Despite this hype of potential, a series of negative events involving self-driving cars have called their reliability into question and prompted legislative calls for regulation (Wakabayashi, 2018). For example, in March 2018, an Uber Technologies Inc.-owned self-driving car struck and killed a pedestrian in Phoenix, Arizona. The fatality garnered widespread public attention and created controversy about the safety and legality of this form of AI-enabled technology. Even before the tragic Uber fatality, self-driving cars had not gained significant recognition, much less trust from the public. Kohl, Knigge, Baader, Böhm, and Krcmar (2018) examined general attitudes toward self-driving cars (using Twitter data before the Uber accident) and found that individuals expressed more tweets about risks about employing a self-driving car than they do for their benefits. Another study about the perception of self-driving cars found that the acceptance of self-driving cars increases when the AI driver is expected to show better performance than human drivers (Gambino & Sundar, 2019).

Unlike previous empirical work in this area of research, this current study seeks to understand perceptions of AI agents in self-driving cars using a human-computer interaction (HCI) approach, which focus on the interaction between users and computing devices to help create more user-friendly environments (Ebert, Gershon, & Veer, 2012). We draw on this line of thinking since scholars have generally theorized about AI as more than a simple tool, but as an agent—an entity that interacts with human beings (Guzman & Lewis, 2019). The central issue this study takes up is perception—how do we think about AI-enabled drivers compared to human drivers on the road? Answering this question allows us to inform better policy, improve marketing and understand the diffusion and adoption of what is anticipated to be a fundamental part of modern life. As AI technology diffuses throughout society, we inevitably confront its costs and benefits against its practicality and the perceptions

* Corresponding author.
E-mail address: joowhaho@usc.edu (J.-W. Hong).

https://doi.org/10.1016/j.chb.2021.106944
Received 27 November 2020; Received in revised form 29 June 2021; Accepted 4 July 2021
Available online 7 July 2021

we have about them. In turn, we may form differences in perceptions from our relationships with AI technologies compared with other forms of technological tools we currently use. Therefore, studies about how responsibility is attributed to AI are crucial and urgently needed (Coeckelbergh, 2019). The following section outlines how we set out to understand the ways that people perceive AI in both positive and negative situations, and we apply the lens of two theoretical frameworks: schema theory and attribution theory.

1. The distinction between the perception of agents and their actions

Research examining how responsibility and blame are attributed to agents contends that the evaluation of an event and the evaluation of an agent should be considered separately when forming moral judgements about others (Malle, Guglielmo, & Monroe, 2014). Therefore, in this paper, we argue that the agents (human vs AI) and their actions should be evaluated separately. Since artificial intelligence has often been deemed as a less autonomous being (Cevik, 2017), it is possible that reactions to AI's actions may not be the same as reactions toward the AI itself (as an entity). For instance, people might think that Apple's Siri conversation style is kind, but that Siri herself is not kind because her responses are programmed. Thus, this study will test both approaches, examining self-driving car scenarios by looking at the agent, as well as their actions. The perception of an AI agent, or an AI-enabled self--driving car, is theorized using schema theory while perceptions about actions are theorized with attribution theory.

2. Schema theory and Computers-Are-Social-Actors

Recent studies characterize self-driving cars not only as vehicles that do not rely on human control and input, but rather as autonomous beings that possess their own agency (Ratan, 2019). This form of agency is primarily derived by the way these technologies are anthropomorphized. By having more humanness-associated characteristics, machines can be perceived differently based on how we expect them to interact (Go & Sundar, 2019). It has been found that self-driving cars with high levels of anthropomorphism lead people to trust their actions (Waytz, Heafner, & Epley, 2014). Therefore, it is expected that people may perceive a self-driving car similarly as an "agent" that can embody human-like qualities rather than a typical vehicle that is not autonomous and requires direct action from a human. To determine the extent to which people can perceive this degree of similarity toward a self-driving car, this study employs schema theory to help explain the cognitive processes of information perception.

Schema theory is an approach that can explain how people may perceive an artificial intelligent agent, or what we refer to as an AI driver, to be more or less similar to a human driver. A schema is a cognitive framework that is built upon an individual's past experiences in order to perceive, comprehend, and recall new information (Bartlett, 1932; Brewer & Treyens, 1981). Schemata function to decrease the complexity of situations in order to facilitate receiving new information so that typical situations can be processed without the effort of acknowledging and interpreting familiar objects and ideas (Harris & Sanborn, 2014; Kleider, Pezdek, Goldinger, & Kirk, 2008).

While schema theory has often been used in media effects research (Dixon, 2006; Meirick, 2006; Scheufele, 1999; Scheufele & Tewksbury, 2006), it has not yet been applied to perceptions of performance for emerging forms of technology like AI. One recent study found that schema, mental shortcuts for efficient information processes, induced from human-human interaction were applicable to both human and AI agents, while schema triggered from human-machine interaction can only be applied to AI agents (Velez et al., 2019). When behavior performed by AI appears similar to human behavior, it is likely that people apply their schemata and evaluate AI similarly to humans. For instance, people attribute responsibility to AI agents when they engage in ethical violations as they would to human violators of the same behavior (Hong & Williams, 2019; Shank & Desanti, 2018).

A similar argument was made by Computers-Are-Social-Actors (CASA). CASA argues that people mindlessly apply the same social heuristics used for human interactions to computers (Nass & Moon, 2000). CASA is from the concept of mindlessness of Langer (1992), who claims that the unconscious awareness state of an individual depends on the context of interactions and relies on past experiences and knowledge. Gambino, Fox, and Ratan (2020) suggested the extended application of CASA through the development of social scripts for human-machine interactions. Previous CASA studies examining human-computer interaction find that humans attribute similar levels of social behavior biases such as gender stereotyping, personality, and expertise recognition toward machines (Nass, Moon, & Green, 1997; Nass & Lee, 2001). As a theoretical framework that focuses on how humans perceive technology as social actors, CASA can be applied to types of human-machine communication that do not necessitate verbal communicative interactions (e.g., operating a self-driving car as a driver). We arrive at this claim by upholding original CASA experimental conditions, which were experiments which found that humans could perceive computers, who engaged in non-reciprocal interactions, as social actors (Nass & Moon, 2000). Recently, many CASA studies utilize experimental vignettes to examine how people evaluate machine agents (Dang & Liu, 2021; Höddinghaus, Sondern, & Hertel, 2021; Pelau, Dabija, & Ene, 2021).

Both schema theory and CASA assume people's intuitive perceptions are based on preexisting knowledge and experiences, which leads to indistinctive behaviors toward machines and humans. Based schema theory and CASA, we argue that the schema people have about human drivers will be triggered when seeing AI-enabled drivers' humanlike performances, which can lead to an unconscious attribution of social norms. Therefore, people will evaluate AI drivers as they assess human drivers. In other words, it is assumed that AI and human drivers will be treated similarly. The context of this study tests how different actors are perceived based on their involvement in a driving scenario that results in either a negative and positive outcome. Taking these factors together, following hypothesis is proposed. Therefore, the following hypothesis is proposed:

H_1: For both an AI-enabled driver and a human driver, the actions of a driver in a positive event will be more positively assessed than the actions of a driver in a negative event.

3. Attribution theory—defensive attribution

After a condemnable offence occurs, people often search for information to determine who or what should accept fault. Attribution theory explains how people identify causal links in order to form judgments about an event (Fiske & Taylor, 1991). Based on the theory, this study examines how an individual recognizing that a driver is a form of AI instead of a human may influence the level of perceived attribution a person places on it. Since computers now function as social actors, some studies have investigated how liability should be allocated to non-human agents when they cause blameworthy outcomes. This theory has been often used in research examining blame attribution in accidents (e.g., Hill, 1975; Jeong, 2009; Rickard, 2014). Thus, we expect this theory can extend our thinking to include a self-driving car as an active social actor in a driving scenario with an accident.

Defensive attribution, which is a component of attribution theory, argues that people attribute more responsibility to harm-doers with less personal or situational similarity (Shaver, 1970; Burger, 1981). This is basic ego protection just in case the same event happens to them. Adding to this, defensive attribution is often theorized to explain how people perceive accidents caused by non-human agents. In this process, humans refer to machine heuristics when employing defensive attribution to other entities that are not humans. Machine heuristics is a concept that helps explain how humans and machines are distinguished and

perceived differently from one another (Sundar & Kim, 2019). Although CASA claims that individuals treat technology as social actors, previous research examining machine heuristics find that people assume humans have qualities that are distinctive from technology. Machine heuristics argues that people distinguish themselves from machines; they think machines are colder, but also more just and trustworthy than humans. For instance, there are cases in which people evaluated AI-written news articles more credibly than articles written by humans (Liu & Wei, 2018; Tandoc, Yao, & Wu, 2020). On the other hand, people expect less interactivity from machines than humans (Go & Sundar, 2019). These studies illustrate that people have different assumptions toward AI agents, because they distinguish machines from other humans. According to machine heuristics, it is expected that people will differentiate humans and machines, which leads to them to feel more attached to human drivers than AI-enabled drivers. In the context of self-driving cars and applying attribution theory, we can expect that people would generally identify with an AI driver less than they do with a human driver because of its machine-distinctive traits. In this case, participants are hypothesized to identify more with a human driver, compared to an AI driver, in an accident based on inherent human-like similarities. Therefore, because attribution theory posits that we are more likely to blame others unlike us, it suggests that people would be more likely to blame an AI-enabled driver than a human driver for the same error:

H_2. People will attribute more responsibility to an AI-enabled driver than a human driver when involved in a negative event.

The usage of attribution theory is not only limited toward negative situations; it has also often been employed to explain the attribution of positive events (Sirin & Villalobos, 2011). In a landmark study by Medway and Lowe (1975), the valence of an event's outcome and the level of attributed responsibility for an event was dependent on the severity of the overall outcome. However, the relationship between the perceived similarity between agents and the level of attribution in positive settings (which this paper seeks to examine) has yet to be investigated. Based on the same logic of how responsibility is attributed in negative events, it can be expected that AI-drivers will be attributed with less credit and praise than human drivers when they both achieve the same positive task. A recent study found that people often experience anxiety and concern when giving out compliments for fear of misestimating the impact their praise might have on others, thus people reduce the amount of compliments they give (Boothby & Bohns, 2020). In other words, to the extent that people prefer to distinguish themselves from machines (Cha et al., 2020; Fox et al., 2015), they would be hesitant to compliment AI agents that they are less attached to than humans. This lessened degree of attribution of responsibility may be attributed to how humans perceive themselves dissimilarly to technology. Taken together, these suggest the following two research questions:

RQ_1. Is there be an interaction between the valence of an event's outcome (i.e., positive or negative) and the type of driver (human vs AI)?

RQ_2: Will people attribute less responsibility to an AI-enabled driver than a human driver in a positive outcome?

4. Method

A 2 × 2 experiment, between-subjects design was conducted where both the identity of agents in a situation (e.g., human vs. artificial intelligence) and the varying valence of an event's outcomes (e.g., positive vs. negative) were considered. Employing a vignette design, a fictitious news article with a positive scenario was presented as a rescue, in which the agent (AI vs human) saved a driver and took them to a hospital. The other version of a fictitious news article with the negative scenario was presented as a crash. The manipulation used fictitious news articles for all four combinations (AI rescue, AI crash, human rescue, and human crash, see below). The dependent variables for the study were the subject's perception of the agents and their level of responsibility attributed to the driver.

4.1. Participants

Amazon Mechanical Turk (MTurk) was used to recruit participants. Participants who participated in the survey more than once were excluded, leaving 230 participants from the 273 who were initially recruited. Power analyses using G-Power (Faul, Erdfelder, Lang, & Buchner, 2007) suggested the potential to detect medium-sized effects. The youngest participant was 19 years old, while the oldest participant was 70 years old (M = 33.21, SD = 12.64). In terms of gender, 62% of participants identified as male and 38% participants identified as female.

4.2. Procedure

Participants who agreed to participate in the study were given one of two types of reading stimuli in the form of a fictitious news article. Using articles as vignettes is a method that has been often used to test people's attitudes and reactions (Billard, 2018; Hong, 2020). One article was based on an actual news report detailing how an AI driver saved its passenger from a potentially fatal acute pulmonary embolism (Ferris, 2018). The following paragraphs of the story were from the rescuing condition with an AI driver:

The artificial intelligent (AI) driving system in a self-driving car is being credited with having helped save a man's life after its AI driver mode was enabled and drove him to a hospital when he suffered a pulmonary embolism.

Yesterday, 29-year-old David McGill was driving to work when he felt an excruciating pain in his abdomen and chest. McGill recounted how he set the autonomous driving function on his self-driving car. "I thought it was easier to have the car drive me to the hospital rather than calling an ambulance," McGill said. After being enabled, the self-driving car's AI driver drove McGill to a nearby hospital.

In the human driver condition, this article replaced the AI by introducing the driver as a "rideshare driver" in a Greenlight rideshare, a fictitious company that does not exist.

The negative outcome/crash news articles were based on a hypothetical car accident situation that led to the death of a passenger. A story about an accident with the death of a passenger was used because it demonstrated a contrast to the article that was about saving a person from immediate death. The following paragraphs are a sample reflecting the accident condition with the human driver:

Yesterday, a GreenLight rideshare car was involved in an accident after it suddenly lost control due to a slippery road condition because of heavy rain. A passenger in the car died in the accident.

Local police said the car was driven by a 41-year-old Michael Smith. The car suddenly lost control and collided with a tree. There was one passenger, 29-year-old David McGill, inside the car at the time of the crash. He was pronounced dead when first responders arrived at the scene of the accident. The car did not hit any other pedestrians and the passenger was the only casualty.

In the AI condition, the human driver was replaced with a self-driving AI agent. The full articles are attached in an appendix A.

After reading a given news article, all participants were asked to report their perceptions of the driver and their thoughts on how responsible the driver was about the depicted incident. Also, their attitudes toward AI, their competency in having knowledge about AI, and their demographic information were asked. Participants were debriefed after they finished their survey.

5. Measures

Because this study examines different attitudes towards the agent and the outcome, different measurements were used to measure each. Also, scales measuring attitudes toward AI, and their competency in AI knowledge were created and measured (see appendix B). The order of both scales and the questions within each were randomized.

Perception of drivers. The impression of the driver was measured using a scale for opinions about the agent (Brave, Nass, & Hutchison, 2005). The measurement included questions asking how caring the driver is and how trustworthy the driver seems, such as choosing between sincere vs. insincere or friendly vs. unfriendly. An exploratory factor analysis (EFA) of 16 items using principal axis factoring with orthogonal (varimax) rotation yielded one factor with eigenvalues >1 (KMO = 0.97), accounting for 69.32% of the total variance with significant Bartlett's test of sphericity, $\chi^2 = 3550.34$, $p < .001$. This seven-point bipolar scale reached high reliability (α = 0.97). Higher average ratings indicate more positive perceptions toward the driver in the given article.

Evaluation of attributed responsibility. To measure how much responsibility is attributed to the driver, this study used the revised causal dimension scale (CDS-II) with four subscales: locus of causality (α = 0.84), external control (α = 0.85), personal control (α = 0.86), and stability (α = 0.83) (McAuley, Duncan, & Russell, 1992). Since there are four subscales that are suggested, a confirmatory factor analysis (CFA) was conducted. The results (NFI = 0.92, CFI = 0.96, GFI = 0.93, TLI = 0.94, SRMR = 0.05, RMSEA = 0.06) showed a good model fit of this measurement. Participants were asked to report how much they agree with given statements, such as 1) The cause of the event is something that was manageable by the driver vs. was not manageable by the driver and 2) The cause of the event is something that the driver will do again vs. the driver will not do again. This 12-item seven-point bipolar scale overall reached high reliability (α = 0.88). Higher average ratings indicate that more responsibility is attributed to the driver.

6. Results

We begin by testing the experiment's manipulation on the valence of a driving scenario and participant's perceived similarity to a driver. Responses were analyzed between different scenarios with an independent samples *t*-test. The manipulation effects of the valence of scenario (e.g., Is the event described in the article positive or negative?) revealed a significant difference between positive ($M = 5.71$, $SD = 1.27$) and negative ($M = 3.36$, $SD = 2.20$) events; $t(228) = 9.97$, $p < .001$, and the manipulation effects of "perceived similarities with drivers" (e.g., How much do you find yourself similar to the driver?) also showed a significant difference between human and AI drivers ($M = 5.07$, $SD = 1.51$) and AI ($M = 4.55$, $SD = 1.87$); $t(228) = -2.33$, $p = .021$. The results from the manipulation check questions confirmed that the conditions were distinct, which allowed the analysis of main effects.

Two sets of independent t-tests were conducted for H_1, which presumed more positive assessments of drivers in the rescuing scenario compared to the accident scenario, regardless of the driver types. One set was analyzed using data only from the AI driver scenario and the other set using data only from the human driver scenario. There was a statistically significant difference between the rescuing scenario ($M = 4.22$, $SD = 1.53$) versus the accident scenario ($M = 3.63$, $SD = 1.35$) when the driver was artificial intelligence [$t(113) = 2.20$, $p = .030$, $d = 0.41$]. Similarly, there was a significant difference between the rescuing scenario ($M = 5.00$, $SD = 1.67$) versus the accident scenario ($M = 4.00$, $SD = 1.18$) when the driver was a human [$t(113) = 3.70$, $p < .001$, $d = 0.69$]. The *t*-test results confirmed that H_1 was supported.

To test H_2 claiming more responsibility attribution toward an AI driver than a human driver in the accident scenario, the level of blame and praise were respectively analyzed using analysis of variance (ANOVA). Levene's tests were conducted and showed the followings: blame [F (1, 115) = 0.45, $p = .51$] and praise [F (1, 111) = 0.21, $p = .65$]. A one-way ANOVA with the same dependent variable using data only from the accident scenario was conducted for H_2, which asked whether people blame AI more than human drivers. There was no statistically significant effect from the identity of drivers on the level of blame [F (1,115) = 0.05, $p = .83$]. The level of blame toward the AI driver ($M = 4.46$, $SD = 1.09$) was similar to the human driver ($M = 4.41$, $SD = 1.15$). H_2 was not supported.

A two-way ANOVA was conducted to test RQ_1, which argues that the identity of drivers and the valence of an event's outcome have an interaction effect on the level of attributed responsibility to drivers. The dependent variable for this analysis was the evaluation of attributed responsibility. Levene's test was conducted to assess the equality of variances, and rejected the homogeneity of variances [F (3, 226) = 0.86, $p = .46$], meaning that the ANOVA was appropriate to conduct. RQ_1 considered an interaction effect between the valence of the incident and the identity of drivers in terms of attributing responsibility. The overall CDS-II outcomes showed an insignificant result for the valence of event [F (1, 226) = 0.20, $p = .66$], the identity of drivers [F (1, 226) = 3.06, $p = .08$], and the interaction [F (1, 226) = 2.12, $p = .15$].

However, there was an interaction effect found in one of the subscales of CDS-II. While the external control [F (1, 224) = 0.938, $p = .334$], the personal control [F (1, 224) = 2.26, $p = .134$], and the locus of control [F (1, 224) = 0.170, $p = .680$] did not show any interaction effect, there was an interaction between the valence of events and the identity of drivers in terms of the stability [F (1, 224) = 5.44, $p = .021$, $\eta p^2 = 0.02$]. Table 1 shows the descriptive statistics for ANOVA. Additionally, the identity of drivers showed significant outcomes regarding stability [F (1, 224) = 6.40, $p = .012$, $\eta p^2 = 0.03$] and personal control [F (1, 224) = 5.40, $p = .021$, $\eta p^2 = 0.02$]. Participants thought AI drivers ($M = 4.78$, $SD = 1.19$) were more likely to repeat the action when placed in similar circumstances in the future compared to human drivers ($M = 4.37$, $SD = 1.24$). Also, they reported that the given situations were more manageable by AI drivers ($M = 4.71$, $SD = 1.52$) than human drivers ($M = 4.23$, $SD = 1.64$).

A one-way ANOVA was conducted using data only from the rescuing scenario to test RQ_2, which assumed less responsibility attribution toward an AI driver than a human driver in the rescuing scenario. There was a statistically significant effect of the identity of drivers on the level of praise [F (1,109) = 4.57, $p = .04$, $\eta p^2 = 0.04$]. The level of praise was higher for the AI driver ($M = 4.76$, $SD = 1.24$) than the human driver ($M = 4.25$, $SD = 1.25$). The outcomes suggest that being AI or a human can influence the level of praise. Also, subscales revealed relevant findings. While the external control [F (1, 111) = 2.54, $p = .114$] and the locus of control [F (1, 111) = 0.01, $p = .93$] did not show a significant outcome, there was a significant difference human and AI drivers in terms of the stability [F (1, 111) = 10.83, $p = .001$, $\eta p^2 = 0.09$] and the personal control [F (1, 111) = 6.34, $p = .013$, $\eta p^2 = 0.05$]. The AI driver was deemed to have more control ($M = 4.97$, $SD = 1.58$) than the human driver ($M = 4.17$, $SD = 1.79$) at the time of the rescuing and more likely to repeat it ($M = 4.95$, $SD = 1.28$) than the human driver ($M = 4.18$, $SD = 1.20$).

7. Discussion

This study aims to understand how individual's attribute

Table 1
Descriptive statistics for ANOVA regarding identity of drivers and the outcome valence.

Valence of Incidents	AI driver			Human Driver		
	M	SD	N	M	SD	N
Rescuing (CDS-II)	4.76	1.24	57	4.25	1.25	56
Accident (CDS-II)	4.46	1.09	58	4.41	1.15	59
Rescuing (Stability)	4.95	1.28	57	4.18	1.20	56
Accident (Stability)	4.61	1.08	58	4.55	1.25	59
Rescuing (Locus of Control)	4.49	1.55	57	4.46	1.56	56
Accident (Locus of Control)	4.31	1.50	58	4.44	1.48	59
Rescuing (Personal Control)	4.97	1.58	57	4.17	1.79	56
Accident (Personal Control)	4.45	1.42	58	4.28	1.51	59
Rescuing (External Control)	4.61	1.44	57	4.20	1.30	56
Accident (External Control)	4.46	1.52	58	4.38	1.22	59

Note. The scale ranges from 1 (*strongly negative*) to 7 (*strongly positive*).

responsibility differently between two types of agents—an AI technology and a human—when they perform the same action. One major finding of this study from both the overall measurement of responsibility attribution and its subscales is that people praise an AI-enabled technology significantly more compared to humans when the event results in a positive outcome. On the other hand, there was no significant difference in attribution in a negative event. This result explains how the outcome valence of an event may influence how people attribute responsibility toward an AI agent. It suggests that we are more likely to praise AI than a human given the same results. However, it contradicts the hypothesis based on attribution theory which suggested that we would insulate human agents from negative outcomes. In cases of attribution that results from a negative-valence event, an AI agent would be deemed more responsible than a human and vice-versa. Given the conditions and findings of the experimental design, attribution theory based on interpersonal communication may not be as applicable to situations involving human-machine interaction. In order to better explain why this contradiction emerges, we turn to Expectancy Violation Theory (EVT). This theory has been applied to human-computer interaction (Spence, Westerman, Edwards, & Edwards, 2014) and can offer a clearer understanding of how individuals may praise AI agents more strongly than human agents in a given situation.

7.1. Using EVT to explain actions by AI

EVT is a communication theory that examines how individuals react and respond to unexpected interpersonal events. The theory states that when individuals face positive violations of expectations in an interaction, they perceive the outcome as more favorable. On the flip side, a negative violation within an interaction can cause them to perceive the outcome as less favorable (Burgoon & Hale, 1988; Burgoon & Jones, 1976). While EVT was originally based on proxemics, the theory has been applied to verbal, computer-mediated communication, and now has been used in human-computer interaction environments (Bonito, Burgoon, & Bengtsson, 1999; Edwards, Edwards, Spence, & Westerman, 2016). For example, Burgoon et al. (2016) and Westerman, Cross, and Lindmark (2019) used EVT to investigate how people see embodied agents or chatbots that deviate from their social expectations of an interaction both positively and negatively. Similarly, a recent finding showed that the evaluation of AI-composed music relied on expectancy violations (Hong, Peng, & Williams, 2020).

We assume that this theory can better explain the differences in attribution between positive and negative outcomes coming from AI and human agents. In other words, people may experience different expectancy violations with human-machine interactions than human-human ones. For instance, if people have low expectations of self-driving cars compared to humans (much less the expectation that an AI driver could save a passenger compared to the expectation that a human driver could), then reading the stimulus in this experiment would not change their attitudes. While this study did not measure their expectations of a self-driving car, it is possible that people have a low expectation of a self-driving car's ability. It is likely the case that they would not expect a self-driving car to proactively transport a rider to a hospital because as Sundar and Kim (2019, p. 538) noted, people consider machines to be mechanical. Therefore, when AI does save lives, which defies the assumption that machines can only perform a given assignment, it elicits a positive violation of their expectations. This could lead to more positive responsibility attribution. The unexpected performance of AI saving a person may have affected the understanding of self-driving car producing consistent results. We expect EVT can explain this better.

There may be different expectations deriving from the anthropomorphic aspect of self-driving cars. Ratan (2019) suggests the concept of "Avacars" to explain that cars may increasingly be seen as an autonomous being having their own identity, rather than being a mere vehicle. Nass argues that CASA is a concept that explains people's anthropomorphic behaviors, which refers to treating non-human entities like humans, such as attributing social rules to machines (1994), but it does not mean that people think machines are human (2000). However, recent CASA research found that an anthropomorphic factor (i.e., a cartoon character) moderates the level of social attributes applied to machines (Lee, 2010). Therefore, people may have different expectations about self-driving cars based on how much they think the vehicles are autonomous and anthropomorphic. This possible expectancy violation may have been relevant to the evaluation of the locus of control.

This study cannot confirm whether people had expectations about self-driving cars and whether there was any expectancy violation because it did not measure the AI heuristics a priori. However, the results suggest that people think AI has more control of itself and is more likely to repeat the same behaviors than human beings. Future research should specifically investigate these expectations to test whether EVT is a viable theoretical foundation going forward.

7.2. Theoretical implications

While the participants in this study reported dissimilar reactions based on the identity of agents in terms of attributing responsibility, their perceptions of these agents were based on the valence of the events. Participants perceived drivers in the rescuing scenarios more positively compared to those in accident scenarios regardless of their identity. It can be inferred from these findings that people evaluate AI agents just as they assess human agents. This finding lends support to explaining how schema theory can be extended from human-computer interactions to human-AI interactions. Also, this study suggests that perceiving AI agents similarly to human agents without any interaction can be explained by schema theory. While the CASA paradigm is often used to explain how people interact with machines (see Guzman, 2018), these studies only examine whether social behaviors are attributed to machines only in direct interaction settings. Previous work does not compare whether the level of attribution of a machine is regarded the same as individuals attributing responsibility to humans. By comparing perceptions of human and AI performances directly with schema theory, it can be possible to explain how people evaluate AI just as they assess human beings. While the applicability of CASA is only limited to direct interacting settings between humans and machines, schema about machines can further explain behaviors toward machines that do not involve face-to-face interactions. This approach can be valuable in guiding future research in other human-computer contexts.

Also, this study partially supported the defensive attribution argument that personal or situational similarity influences the level of responsibility attribution (Burger, 1981; Shaver, 1970). While the theory was first devised based on interpersonal interactions, the results of this study suggest that the attribution of responsibility toward machines should consider particular aspects of human-machine communication. It is assumed that various biases and understandings of machines play a crucial role when blaming and praising AI performances.

7.3. Limitation and future directions

Aside from not considering EVT in our initial theorizing, another limitation of this study is that people's experiences of driving were not taken into account, such as having a driver's license or experiences with car accidents. Not only their understanding of AI but also attitudes of driving would have influenced the level of attributing responsibilities. For instance, people who have experienced a car accident as a driver would perceive the given car accident article about differently compared to the ones without any experience. Future studies about self-driving cars should consider this factor. Also, this study used only one positive and one negative event. Using multiple different scenarios for each positive and negative case would have decreased potential biases coming from the article. Therefore, using different stories is advised for future studies. The lack of comparability between the negative and positive scenarios should have been more delicately controlled because

the positive scenario was an unusual case, compared to the accident scenario, which was more mundane. There was no attention check question in this study. Excluding participants who did not participate in this study mindfully would have shown clearer findings.

This project was primarily concerned with examining how human perceptions of responsibility of an AI-enabled driver were impacted by the outcome of a driving scenario. The results of the study indicate the complexity in the sociotechnical relationship humans share with technology. This study adds to extant literature on the relationship humans are developing with AI-technologies. In a recent article, Guzman and Lewis (2019) urge communication scholars to rethink how the traditional boundaries of communication may not fit with how humans interact with emerging forms of artificial intelligent agents. The authors lay out a research agenda for scholars by highlighting three areas of inquiry that include the functional, relational, and metaphysical elements AI-technologies may have with humans. This study attempts to answer their call to extend our understanding of how we perceive our relationship toward these emerging technologies as they become part of our day-to-day lives. With the inevitable development and integration of self-driving cars in the near future, research in this area is imperative to provide designers and consumers of these technologies with knowledge about the use and effects of these agents.

Author contributions

Joo-Wha Hong: Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, and Project Administration Ignacio Cruz: Conceptualization, Methodology, Investigation, and Writing - Review & Editing. Dmitri Williams: Conceptualization, Methodology, Writing – review & editing, and Supervision

Acknowledgement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

No potential conflict of interest was reported by the authors.

Appendix A

I. AI driver and Accident scenario

Passenger died in self-driving car accident on a slippery road.

By Alex Robbins.

May 13, 2019.

SAN FRANCISCO — Yesterday, a self-driving car was involved in an accident after it suddenly lost control due to an unidentified cause by the artificial intelligent (AI) driver that is currently being investigated. A passenger in the self-driving car died in the accident.

Fatal accident caused by AI driver.

Local police said the self-driving car was set in AI autonomous driving, where the AI driver had a full control of the car at the time of the accident. The AI driver system, which had been driving for two years, suddenly lost control of the car and collided with a tree. There was a one passenger, 29-year-old David McGill, inside the car at the time of the crash. He was pronounced dead when first responders arrived to the scene of the accident. The car did not hit any other pedestrians and the passenger was the only casualty.

Nexus, the self-driving car's manufacturer, pointed out the severity of this accident. The company released a statement earlier today that prioritized the safety of its passengers.

II. AI driver and Rescue scenario

AI driver saves passenger's life by steering him to hospital.

By Alex Robbins.

May 13, 2019.

SAN FRANCISCO — The artificial intelligent (AI) driving system in a self-driving car is being credited with having helped save a man's life after its AI driver mode was enabled and drove him to a hospital when he suffered a pulmonary embolism.

Passenger rescued by AI driver.

Yesterday, 29-year-old David McGill on his way work using the AI driver system in his car when he felt an excruciating pain in his abdomen and chest. McGill recounted how he was glad he set the autonomous driving function on his self-driving car. "It was easier to have the car drive me to the hospital rather than calling an ambulance," McGill said.

The self-driving car's AI driver, which McGill had been using for two years, drove him to a nearby hospital.

Nexus, the self-driving car's manufacturer, pointed out the severity of this incident. The company released a statement earlier today that prioritized the safety of its passengers.

III. Human driver and Accident scenario

Passenger died in self-driving car accident on a slippery road.

By Alex Robbins.

May 13, 2019.

SAN FRANCISCO — Yesterday, a GreenLight rideshare car was involved in an accident after it suddenly lost control due an unidentified cause by the driver that is currently being investigated. A passenger in the car died in the accident.

Fatal accident caused by driver.

Local police said the car was driven by a 41-year-old Michael Smith. Smith, who had been driving for GreenLight for two years, suddenly lost control of the car and collided with a tree. There was one passenger, 29-year-old David McGill, inside the car at the time of the crash. He was pronounced dead when first responders arrived to the scene of the accident. The car did not hit any other pedestrians and the passenger was the only casualty.

GreenLight, the rideshare company, pointed out the severity of this accident. The company released a statement earlier today that prioritized the safety of its passengers.

IV. Human driver and Rescue scenario

Rideshare driver saves passenger's life by steering him to hospital.

By Alex Robbins.

May 13, 2019.

SAN FRANCISCO — A GreenLight rideshare driver is being credited with having helped save a man's life after the rideshare driver drove his passenger to a hospital when he suffered a pulmonary embolism during a ride.

Passenger rescued by driver.

Yesterday, 29-year-old David McGill rode in a Greenlight rideshare to work when he felt an excruciating pain in his abdomen and chest. He could not speak because of the pain at that time, but his GreenLight driver, 41-year-old Michael Smith who had been driving for GreenLight for two years, recognized that his passenger was in pain and drove him to the closest hospital. "I thought it was easier to drive the passenger to the hospital rather than calling an ambulance," Smith said.

GreenLight, the rideshare company, pointed out the severity of this incident. The company released a statement earlier today that prioritized the safety of its passengers.

Appendix B

Perception of drivers

Indicate how well the adjective represents the driver in the article

you just read.

Compassionate - - - - - - - Not Compassionate.
Unselfish - - - - - - - Selfish.
Friendly - - - - - - - Unfriendly.
Cooperative - - - - - - - Competitive.
Likable - - - - - - - Unlikable.
Pleasant - - - - - - - Unpleasant.
Appealing - - - - - - - Unappealing.
Not irritating - - - - - - - Irritating.
Trustworthy - - - - - - - Untrustworthy.
Honest - - - - - - - Dishonest.
Reliable - - - - - - - Unreliable.
Sincere - - - - - - - Insincere.
Intelligent - - - - - - - Unintelligent.
Smart - - - - - - - Dumb.
Capable - - - - - - - Incapable.
Warm - - - - - - - Cold.

Evaluation of attributed responsibility.

To what extent do you agree with the following statements?

The cause of the event is something that …

Reflects more of the driver - - - - - - - reflects more of the situation
Was manageable by the driver - - - - - - - was not manageable by the driver
The driver will do again - - - - - - - the driver will not do again
The driver could regulate - - - - - - - the driver could not regulate
Others have control over - - - - - - - others have no control over
Pertains to the driver - - - - - - - does not pertain to the driver
Is stable over time - - - - - - - is variable over time
Is under the influence of other factors - - - - - - - is not under the influence of other factors
Is something about the driver - - - - - - - is something about other factors
The driver had influence over - - - - - - - the driver had no influence over
Is unchangeable - - - - - - - is changeable
Other factors can regulate - - - - - - - other factors cannot regulate.

Attitudes toward AI.

(From “Strongly Disagree” to “Strongly Agree”) Please rate the extent to which you agree with the following statements:

AI is a positive force in the world.
AI research should be funded more.
AI is generally helpful.
There is a need to use AI.

Competency in AI knowledge.

(From “Extremely Unconfident” to “Extremely Confident”) How would you rate your confidence in the following:

Explaining what artificial intelligence is.
Having a conversation about artificial intelligence.
My knowledge about artificial intelligence.

References

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge University Press.

Billard, T. J. (2018). Attitudes toward transgender men and women: Development and validation of a new measure. *Frontiers in Psychology, 9*, 387. https://doi.org/10.3389/fpsyg.2018.00387

Bonito, J. A., Burgoon, J. K., & Bengtsson, B. (1999). The role of expectations in human-computer interaction. In *Proceedings of the international ACM SIGGROUP conference on Supporting group work* (pp. 229–238). https://doi.org/10.1145/320297.320324. ACM.

Boothby, E. J., & Bohns, V. K. (2020). Why a simple act of kindness is not as simple as it seems: Underestimating the positive impact of our compliments on others. *Personality and social psychology bulletin*, Article 0146167220949003. https://doi.org/10.1177/0146167220949003

Borraz, R., Navarro, P. J., Fernández, C., et al. (2018). Cloud incubator car: A reliable platform for autonomous driving. *Applied Sciences, 8*(2), 303. https://doi.org/10.3390/app8020303

Brave, S., Nass, C., & Hutchinson, K. (2005). Computers that care: Investigating the effects of orientation of emotion exhibited by an embodied computer agent. *International Journal Of Human-Computer Studies, 62*(2), 161–178. https://doi.org/10.1016/j.ijhcs.2004.11.002

Brewer, W., & Treyens, J. (1981). Role of schemata in memory for places. *Cognitive Psychology, 13*(2), 207–230. https://doi.org/10.1016/0010-0285(81)90008-6

Burger, J. M. (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

Burgoon, J. K., Bonito, J., Lowry, P., Humpherys, S., Moody, G., Gaskin, J., et al. (2016). Application of Expectancy Violations Theory to communication with and judgments about embodied agents during a decision-making task. *International Journal of Human-Computer Studies, 91*, 24–36. https://doi.org/10.1016/j.ijhcs.2016.02.002

Burgoon, J. K., & Hale, J. (1988). Nonverbal expectancy violations: Model elaboration and application to immediacy behaviors. *Communication Monographs, 55*(1), 58–79. https://doi.org/10.1080/03637758809376158

Burgoon, J. K., & Jones, S. B. (1976). Toward a theory of personal space expectations and their violations. *Human Communication Research, 2*(2), 131–146. https://doi.org/10.1111/j.1468-2958.1976.tb00706.x

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), 75–87.

Cha, Y. J., Baek, S., Ahn, G., Lee, H., Lee, B., Shin, J. E., et al. (2020). Compensating for the loss of human distinctiveness: The use of social creativity under Human–Machine comparisons. *Computers in Human Behavior, 103*, 80–90. https://doi.org/10.1016/j.chb.2019.08.027

Coeckelbergh, M. (2019). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and Engineering Ethics*, 1–18. https://doi.org/10.1007/s11948-019-00146-8

Dang, J., & Liu, L. (2021). Robots are friends as well as foes: Ambivalent attitudes toward mindful and mindless AI robots in the United States and China. *Computers in Human Behavior, 115*, 106612. https://doi.org/10.1016/j.chb.2020.106612

Dixon, T. L. (2006). Schemas as average conceptions: Skin tone, television news exposure, and culpability judgement. *Journalism & Mass Communication Quarterly, 83*(1), 131–149. https://doi.org/10.1177/107769900608300109

Ebert, A., Gershon, N., & Veer, G. (2012). Human-computer interaction. *KI - Künstliche Intelligenz, 26*(2), 121–126. https://doi.org/10.1007/s13218-012-0174-7

Edwards, C., Edwards, A., Spence, P., & Westerman, D. (2016). Initial interaction expectations with robots: Testing the human-to-human interaction script. *Communication Studies, 67*(2), 227–238. https://doi.org/10.1080/10510974.2015.1121899

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/BF03193146

Ferris, R. (2018). *May 22) Self-driving cars are scaring more people. CNBC. Available at:*. https://www.cnbc.com/2018/05/22/self-driving-cars-are-scaring-more-people.html.

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

Fox, J., Ahn, S. J., Janssen, J. H., Yeykelis, L., Segovia, K. Y., & Bailenson, J. N. (2015). Avatars versus agents: A meta-analysis quantifying the effect of agency on social influence. *Human-Computer Interaction, 30*, 401–432. https://doi.org/10.1080/07370024.2014.921 494

Gambino, A., Fox, J., & Ratan, R. A. (2020). Building a stronger CASA: Extending the computers Are social actors paradigm. *Human-Machine Communication, 1*, 71–86. https://doi.org/10.30658/hmc.1.5

Gambino, A., & Sundar, S. S. (2019). April). Acceptance of self-driving cars: Does their posthuman ability make them more eerie or more desirable?. In *Extended abstracts of the 2019 CHI conference on human factors in computing systems* (p. LBW2513). ACM.

Go, E., & Sundar, S. (2019). Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior, 97*, 304–316. https://doi.org/10.1016/j.chb.2019.01.020

Guzman, A. (2018). What is human-machine communication anyways? In A. Guzman (Ed.), *Human-machine communication: Rethinking communication technology and ourselves* (pp. 1–28). New York, NY: Peter Lang.

Guzman, A. L., & Lewis, S. C. (2019). *Artificial intelligence and communication: A human–machine communication research agenda*. New Media & Society. https://doi.org/10.1177/1461444819858691

Harris, R. J., & Sanborn, F. W. (2014). *A cognitive psychology of mass communication* (6th ed.). New York, NY: Routeledge.

Hill, F. A. (1975). Attribution of responsibility in a campus stabbing incident. *Social Behavior and Personality, 3*(2), 127–131. http://search.proquest.com/docview/60862664/.

Höddinghaus, M., Sondern, D., & Hertel, G. (2021). The automation of leadership functions: Would people trust decision algorithms? *Computers in Human Behavior, 116*, 106635. https://doi.org/10.1016/j.chb.2020.106635

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

Hong, J. W., Peng, Q., & Williams, D. (2020). *Are you ready for artificial mozart and skrillex? An experiment testing expectancy violation theory and AI music*. New Media & Society. https://doi.org/10.1177/1461444820925798

Hong, J. W., & Williams, D. (2019). Racism, responsibility and autonomy in HCI: Testing perceptions of an AI agent. *Computers in Human Behavior, 100*, 79–84. https://doi.org/10.1016/j.chb.2019.06.012

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

Kim, T. S., Na, J. C., & Kim, K. J. (2012). Optimization of an autonomous car controller using a self-adaptive evolutionary strategy. *International Journal of Advanced Robotic Systems, 9*(3), 73. https://doi.org/10.5772/50848

Kleider, H., Pezdek, K., Goldinger, S., & Kirk, A. (2008). Schema-driven source misattribution errors: Remembering the expected from a witnessed event. *Applied Cognitive Psychology, 22*(1), 1–20. https://doi.org/10.1002/acp.1361

Kohl, C., Knigge, M., Baader, G., Böhm, M., & Krcmar, H. (2018). Anticipating acceptance of emerging technologies using twitter: The case of self-driving cars. *Journal Of Business Economics, 88*(5), 617–642. https://doi.org/10.1007/s11573-018-0897-5

Langer, E. (1992). Matters of mind: Mindfulness/mindlessness in perspective. *Consciousness and Cognition, 1*(3), 289–305. https://doi.org/10.1016/1053-8100(92)90066-J

Lee, E. (2010). What triggers social responses to flattering computers? Experimental tests of anthropomorphism and mindlessness explanations. *Communication Research, 37*(2), 191–214. https://doi.org/10.1177/0093650209356389

Lee, U., Jung, J., Jung, S., & Shim, D. (2018). Development of a self-driving car that can handle the adverse weather. *International Journal of Automotive Technology, 19*(1), 191–197. https://doi.org/10.1007/s12239-018-0018-z

Liu, B., & Wei, L. (2019). Machine Authorship in Situ: Effect of news organization and news genre on news credibility. *Digital Journalism, 7*(5), 635–657. https://doi.org/10.1080/21670811.2018.1510740

Malle, B., Guglielmo, S., & Monroe, A. (2014). A theory of blame. *Psychological Inquiry, 25*(2), 147–186. https://doi.org/10.1080/1047840X.2014.877340

McAuley, E., Duncan, T., & Russell, D. (1992). Measuring causal attributions: The revised causal dimension scale (CDSII). *Personality and Social Psychology Bulletin, 18*, 566–573. https://doi.org/10.1177/0146167292185006

Medway, F., & Lowe, C. (1975). Effects of outcome valence and severity on attribution of responsibility. *Psychological Reports, 36*(1), 239–246. https://doi.org/10.2466/pr0.1975.36.1.239

Meirick, P. C. (2006). Media schemas, perceived effects, and person perceptions. *Journalism & Mass Communication Quarterly, 83*(3), 632–649. https://doi.org/10.1177/107769900608300310

Nass, C., Lee, K., & Nass, C. (2001). Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied, 7*(3), 171–181. https://doi.org/10.1037/1076-898X.7.3.171

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

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

Pelau, C., Dabija, D. C., & Ene, I. (2021). What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. *Computers In Human Behavior, 122*. https://doi.org/10.1016/j.chb.2021.106855, 106855.

Rahman, A. A., Hamid, U., & Chin, T. A. (2017). Emerging technologies with disruptive effects: A review. *Perintis eJournal, 7*(2), 111–128.

Ratan, R. (2019). When automobiles are avacars: A self-other-utility approach to cars and avatars. *International Journal of Communication, 13*, 1–19.

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

Scheufele, D. (1999). Framing as a theory of media effects. *Journal of Communication, 49*(1), 103–122. https://doi.org/10.1111/j.1460-2466.1999.tb02784.x

Scheufele, D. A., & Tewksbury, D. (2006). Framing, agenda setting, and priming: The evolution of three media effects models. *Journal of Communication, 57*(1), 9–20. https://doi.org/10.1111/j.1460-2466.2006.00326_5.x

Shank, D., & Desanti, A. (2018). Attributions of morality and mind to artificial intelligence after real-world moral violations. *Computers in Human Behavior, 86*, 401–411. https://doi.org/10.1016/j.chb.2018.05.014

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. https://doi.org/10.1037/h0028777

Sirin, C., & Villalobos, J. (2011). Where does the buck stop? Applying attribution theory to examine public appraisals of the president. *Presidential Studies Quarterly, 41*(2), 334–357. https://doi.org/10.1111/j.1741-5705.2011.03857.x

Spence, P., Westerman, D., Edwards, C., & Edwards, A. (2014). Welcoming our robot overlords: Initial expectations about interaction with a robot. *Communication Research Reports, 31*(3), 272–280. https://doi.org/10.1080/08824096.2014.924337

Sundar, S. S., & Kim, J. (2019). Machine heuristic: When we trust computers more than humans with our personal information. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. ACM. https://doi.org/10.1145/3290605.3300768.

Tandoc, E. C., Jr., Yao, L. J., & Wu, S. (2020). Man vs. machine? The impact of algorithm authorship on news credibility. *Digital Journalism, 8*(4), 548–562. https://doi.org/10.1080/21670811.2020.1762102

Teoh, E., & Kidd, D. (2017). Rage against the machine? Google's self-driving cars versus human drivers. *Journal of Safety Research, 63*, 57–60. https://doi.org/10.1016/j.jsr.2017.08.008

Velez, J. A., Loof, T., Smith, C. A., Jordan, J. M., Villarreal, J. A., & Ewoldsen, D. R. (2019). Switching schemas: Do effects of mindless interactions with agents carry over to humans and vice versa? Journal of Computer-Mediated Communication, 24, 335-352. https://doi.org/10.1093/jcmc/zmz016.

Vellinga, N. E. (2017). From the testing to the deployment of self-driving cars: Legal challenges to policymakers on the road ahead. *Computer Law & Security Report: The International Journal of Technology Law and Practice, 33*(6), 847–863. https://doi.org/10.1016/j.clsr.2017.05.006

Wakabayashi, D. (2018, March 19). Self-driving Uber car kills pedestrian in Arizona, where robots roam. *New York Times*. https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html.

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology, 52*, 113–117.

Westerman, D., Cross, A., & Lindmark, P. (2019). I believe in a thing called bot: Perceptions of the humanness of "chatbots. *Communication Studies, 70*(3), 295–312. https://doi.org/10.1080/10510974.2018.1557233