From Cute to Incompetent: The Impact of Anthropomorphic Design on Responsibility Attribution in Autonomous Driving

Uwe Messer
Universität der Bundeswehr München
uwe.messer@unibw.de

Denise Pape
University of Goettingen
denise.pape@uni-goettingen.de

Nadine Lukas
Universität der Bundeswehr München
nadine.lukas@unibw.de

Leonore Peters
University of Bamberg
leonore.peters@uni-bamberg.de

Abstract

In the era of artificial intelligence (AI) and automation, the shift from human labor to machines is evident. This study focuses on autonomous vehicles (AVs) and explores the attribution of responsibility in the case of accidents, considering anthropomorphic design elements in the vehicle front. Prior research emphasizes the positive effects of anthropomorphizing technology but has overlooked potential drawbacks. By examining specific facial schemas, we aim to understand how design elements influence responsibility attribution in AVs. Our findings suggest that a baby-faced design reduces responsibility attribution in non-autonomous vehicles but increases it in fully autonomous vehicles.

Keywords: anthropomorphism, autonomous vehicles, responsibility attribution, babyfacedness

1. Introduction

In light of the rapid advancement and adoption of artificial intelligence and automation technology, a consequential shift is observed as human labor is being substituted by machines (Waytz et al., 2014). These machines have emerged as viable alternatives to human effort for a range of tasks that traditionally required a human mind (Dwivedi et al., 2021). Examples include robots and drones, medical diagnostics, and the advent of self-driving, autonomous vehicles (AVs). Autonomous driving has the potential to transform transportation by improving safety, comfort, and carbon efficiency in road traffic. The adoption of autonomous vehicles depends on two interrelated issues: 1) user acceptance and 2) regulations for the distribution of responsibility among road users in the event of an accident. Despite numerous trials providing compelling evidence of the superior safety performance of AVs in comparison to human drivers, research indicates that acceptance of these vehicles decreases as the level of autonomy increases (Schoettle & Sivak, 2016). Ironically, there is a widespread belief that autonomous vehicles do not perform as well as human drivers (Schoettle & Sivak, 2016). Given that skepticism remains one of the most significant barriers to AV adoption (Choi & Ji, 2015), the cultivation of favorable perceptions of autonomous vehicles is a critical factor in achieving widespread acceptance and a prerequisite for the promises the technology holds for society. Further, it is important to understand how users allocate responsibility to AVs to fully unleash their potential and effectively contribute to a positive transformation of road traffic.

To tackle the acceptance problem engineers and psychologists have attempted to increase trust in technology by endowing it with human-like features (e.g., Waytz et al., 2014; Song & Luximon, 2021) – a strategy that is known as ‘anthropomorphism’ (Epley et al., 2007). The merit of anthropomorphic design resides in the innate tendency of humans to perceive human-like characteristics, such as faces, in inanimate objects (Landwehr et al., 2011). This propensity to recognize facial features in nonhuman entities is attributed to the fact that the human face is processed as a biologically significant stimulus receiving heightened attention, compared to other environmental stimuli (Mondloch et al., 1999). The car front is the most important visual stimulus that shapes the first impression of a vehicle and users tend to process it similarly to how they process human faces (Maeng & Aggarwal, 2018). In automotive design, it is well known that design elements can resemble specific facial expressions, that can elicit emotional responses from viewers (Landwehr et al., 2011; Aggarwal & McGill, 2007; Windhager et al., 2008). For instance, a car grille with a downturned and slanted headlight is perceived as more aggressive, while a grille with an upturned shape and arched headlights resembles a smiling expression and is being interpreted

URI: https://hdl.handle.net/10125/106589
978-0-9981331-7-1

as friendly (Aggarwal & McGill, 2007). The practice of imbuing cars with anthropomorphic qualities can lead to a heightened level of trust in the car's abilities (Waytz et al., 2014) – because anthropomorphic features increase the perception of fundamental mental capacities and render the car more proficient in carrying out its tasks (Pierce et al., 2013).

However, a review of the extant literature reveals a one-sided focus on the positive aspects of anthropomorphizing vehicles in the context of low automation levels (e.g., Waytz et al., 2014; Niu et al., 2018). Research on negative and unintended downstream consequences of anthropomorphism for autonomous vehicles, in particular, is virtually non-existent. Further, we find that research takes a very broad perspective on anthropomorphism which does not consider how the particular arrangement of design features and their resemblance to human cognitive schema might affect users' perception of AVs. Upon examination of the literature on facial physiology, it is evident that specific facial schemas are linked to perceptions of capabilities (Gorn et al., 2008; Carré et al., 2009) and even guilt ascription (Wilson & Rule, 2015). Notably, a wider face shape is typically associated with a dominant and powerful disposition, while a babyface characterized by a high forehead, small nose, chin, and large eyes – is commonly associated with traits such as naiveté (Gorn et al., 2008). Therefore, it is plausible that the direction of the effects of endowing vehicles with human-like features depends on the composition of design elements and their resemblance to human schema. Amidst the backdrop of car manufacturers incorporating aggressive or angry faces (e.g., Lamborghini Huracan, Ford Raptor) or adopting more friendly, baby-faced features (e.g., BMW Mini, Fiat 500), it is also of paramount practical importance to understand how these design choices might shape user perceptions in case of accidents when such vehicles become autonomous. Hence, we aim to answer the following research question:

RQ1: How do different anthropomorphic design elements influence the attribution of responsibility in autonomous vehicles in the case of an accident and what are the underlying mechanisms?

Our research encompassed three comprehensive studies, supplemented by a pretest. In Study 1, we manipulated a series of anthropomorphic car design features to examine their impact on responsibility attribution. Our findings reveal that while a baby-faced design reduces the attribution of responsibility to the car when it lacks autonomous driving capabilities, a notable shift occurs when the car is fully autonomous. Study 2 replicates the effects observed in Study 1 and examines two crucial control variables: the general tendency to anthropomorphize and technology optimism. Building upon the findings from Studies 1 and 2, which demonstrate that the babyface design for autonomous vehicles backfires as it increases responsibility for accidents compared to a control design, Study 3 aims to explore the underlying mechanisms behind this effect. Our findings indicate that a baby-faced vehicle demonstrates a heightened sense of benevolence (while not exhibiting a corresponding increase in ability) when operating at a low level of automation. On the other hand, when the vehicle functions autonomously, the baby-faced design is perceived as less benevolent and less capable, resulting in a higher assignment of responsibility in case of an accident.

We integrate the anthropomorphism literature on AVs with the psychological face perception literature (e.g., Oosterhof & Todorov, 2008; Peterson et al., 2022; Carré et al., 2009). By doing so, we elucidate that anthropomorphism can generate an unintended backlash effect, in contrast to the prevailing discourse, which predominantly emphasizes positive effects. Furthermore, our study contributes to the understanding that the assignment of responsibility for autonomous technologies is influenced by superficial visual features, similar to the attribution of responsibility to human wrongdoers depending on their appearance (see e.g., Willis & Todorov, 2006). Given the potentially detrimental outcomes associated with anthropomorphism in the context of autonomous vehicles, it becomes imperative to acquire a comprehensive understanding of whether humanizing these entities can result in counterproductive outcomes when product failures occur. The implications of our research extend to designers and manufacturers of autonomous vehicles, shedding light on the critical importance of incorporating appropriate design considerations for autonomous systems.

The article is structured as follows: We begin with a comprehensive overview of research on responsibility attribution and anthropomorphic design elements in autonomous driving. Next, we delve into the theoretical foundations of anthropomorphism, vehicle design, and responsibility attribution, leading to the formulation of our hypotheses. We then conduct an empirical analysis using regression analysis to examine these relationships. Finally, we discuss the implications of our findings for theory and managerial practice.

2. Related work on responsibility attribution and anthropomorphic design in autonomous driving

Research on AVs has gained traction in recent years. This study adopts the terminology established by the Society of Automotive Engineers (SAE) J3016 information report, which categorizes automated vehicles into different levels ranging from 0 (no automation) to 5 (full automation, no driver input required) (SAE International, 2014). The level of vehicle automation is contingent upon the complexity of the autonomous technology utilized and the degree of involvement of the human driver in driving decisions, constituting a vital determinant of AV safety (Wang et al., 2020). Previous studies have thoroughly investigated various aspects of autonomous vehicles, including factors such as general attitudes (e.g., Charness et al., 2018), acceptance and adoption (e.g., Hein et al., 2018), and willingness to pay for self-driving cars (e.g., Bansal & Kockelmann, 2018). Furthermore, trust has received a considerable amount of attention in the literature (e.g., Choi & Ji, 2015), being an essential prerequisite in the adoption of the technology. Insufficient attention has been given to the issue of responsibility attribution in the context of accidents, particularly from a psychological standpoint rather than a legal one. To date, there has only been preliminary research exploring the attribution of moral responsibility in this domain. Shariff et al. (2017) acknowledge that autonomous vehicles can adopt utilitarian perspectives to minimize risk or prioritize self-protection, particularly for the occupant. This ethical dilemma gives rise to a social dilemma, wherein citizens may perceive the utilitarian approach as more ethically justifiable, while occupants may prioritize self-protective strategies. Furthermore, McManus and Rutchick (2019) have demonstrated that the attribution of responsibility is influenced by the driver's capacity to influence the outcome and sequence of their decisions. Notably, reducing the driver's agency, whether directly or indirectly, tends to diminish blame for negative outcomes while eliciting greater praise for positive outcomes. Further results are provided by Copp et al. (2021) who examine the attribution of responsibility in a fatal autonomous vehicle accident. In this study, participants assign responsibility to various entities involved in autonomous vehicle operation, including human drivers, AV manufacturers, pedestrians, and external factors like acts of God. The outcomes consistently reveal that human drivers were predominantly held accountable, irrespective of the monitoring conditions in place, whereas AV manufacturers remained accountable regardless of the requirement for human driver oversight. Another study by Beckers et al. (2022) identifies that people tend to hold human drivers primarily responsible for crashes involving partially automated vehicles, despite recognizing the challenges drivers face when taking over control from automation. This finding suggests that attributing sole responsibility to drivers may be unreasonable, emphasizing the importance of raising public awareness regarding the impact of automation on driver capabilities in such scenarios. However, a crucial aspect that has received scarce attention is whether the attribution of responsibility can be influenced by the appearance of autonomous vehicles. Despite the extensive research conducted on responsibility attribution and anthropomorphism, no previous studies have integrated these research streams to examine whether specific design features can alter the assignment of responsibility.

However, there are a few studies on anthropomorphic elements in autonomous vehicles incorporating features such as a designated name, gender, and voice for the vehicle (e.g., Waytz et al., 2019), the anthropomorphic embodiment of information (e.g., Niu et al., 2018; Kraus et al., 2009) or employment of on-board voice-based agent interfaces (e.g., Forster et al., 2017). In previous research on anthropomorphism in conventional (non-autonomous) cars, prominent facial physiology frameworks such as baby-faced and aggressive features (e.g., wide, facial width-to-height ratio (fWHR)) have been employed. In our study, we aim to bridge the gap between the literature on anthropomorphism in autonomous cars and the findings from facial physiology research, investigating the extent to which changes in the appearance of AVs influence the attribution of moral responsibility and exploring the underlying mechanisms that may explain this interaction.

3. Theoretical background

3.1. Anthropomorphism

Anthropomorphism describes the phenomenon of assigning human-like characteristics to nonhuman agents. Simply put, anthropomorphism is the "attribution of human characteristics or behavior to a god, animal, or object" (Oxford Dictionary, Soanes & Stevenson, 2005). The attributes allocated to nonhuman agents include motivations, goals, or emotions (Epley et al., 2007). Moreover, it also includes conscious experience, metacognition, and intentions (Gray et al., 2007). Treating nonhuman agents like humans has a significant impact on recognizing them as moral agents, allocating certain behavioral expectations and interpretations thereof (Epley et al., 2007). Gazzola et al. (2007) provide neuroscientific evidence demonstrating that making judgments about anthropomorphized agents activates the same neural systems as when making judgments about humans. The extent of ascribing human-like characteristics depends on the degree of the human-like appearance of the nonhuman agents (e.g., Kiesler et al., 2008; Waytz et al., 2010), including responsiveness, voice, and facial characteristics. This fact is widely used in branding strategies and product design. Since human-like facial characteristics in particular trigger our neural circuits, those facial design elements are very popular in the automotive sector. Examples are vehicles with human-like facial features such as the BMW Mini, the Fiat 500, or the Lamborghini Aventador.

3.2. Vehicle design and responsibility attribution

Previous research indicates that emotional expressions in vehicles, resembling human faces, trigger perceptual mechanisms similar to those activated by human faces, recognizing and categorizing a dominant or submissive face. For example, research on decoding emotional expressions shows that the perception of aggressiveness is cued by the shape of both the mouth and the eyes (Landwehr & Herrmann, 2011). In the context of vehicle design, the grille can be interpreted as a human mouth, while the headlights resemble human eyes (Aggarwal & McGill, 2007). Additionally, the ratio of facial width to upper facial height (fWHR) is another important cue used by humans to form impressions about others (Carré et al., 2009). High fWHR is often associated with attributes such as aggression and dominance. In contrast, baby-faced features, upturned grille, and arched headlights are more likely to be perceived as submissive and innocent (Gorn et al., 2008). Studies have shown that baby-faced individuals are perceived as less likely to deceive, are characterized as warm and submissive (Zebrowitz, 1997), and are more moral in general (McArthur, 1985). Gorn et al. (2008) find that people conclude that baby-faced decision-makers have a lower deception intent than their mature-faced counterparts. Drawing from these findings, we hypothesize that the use of design elements that resemble a babyface will have a negative effect on the attribution of responsibility to the vehicle.

H1: Individuals attribute less responsibility to the vehicle when it resembles a baby-faced design.

As AVs are perceived as sentient entities, embodying intelligence, competence, and independent decision-making (Waytz & Epley, 2014), the attribution of responsibility undergoes a notable shift, with accountability placed on the autonomous vehicle itself. A potential discrepancy arises between the baby-faced design and the autonomy of the vehicle. Drawing on Mandler's (1982) schema congruence theory, this disparity may result in an altered overall evaluation of responsibility attribution, as there is a mismatch between the car's apparent capabilities and the simplistic cognitive frameworks associated with a baby-faced design. While a baby-faced design may elicit positive associations of honesty and good intentions in a non-autonomous vehicle, it becomes incongruous in the realm of autonomous driving, where the vehicle's competence and expertise hold paramount importance (Gorn et al., 2008). Thus, for AVs, we assume an effect of baby-faced design on attribution of responsibility that is opposite to H1:

H2: Individuals attribute more responsibility to the vehicle resembling a baby-faced design when it is autonomous (level 5) vs. partially automated (level 3).

For the special case of babyfacedness, we want to dive deeper into the underlying mechanisms of responsibility attribution. Two major aspects are crucial in responsibility attribution: capabilities and intentions (Hartman et al., 2022). The capability to perform successfully seems to play a pivotal role in responsibility attribution. It captures the knowledge and skills needed to do a specific job (Gabarro, 1978), and incorporates the "can-do" component, meaning the expertise and competence needed to act in an appropriate fashion (Colquitt et al., 2007). Apparent intentions can amplify responsibility attribution. Benevolence – defined as the extent to which someone is believed to want to do good, apart from any profit motives (Mayer et al., 1995) – can create an emotional attachment, fostered by warmth and supportiveness. Thus, we include these two constructs – ability and benevolence – as mechanisms in our theoretical framework and resume:

H3: Benevolence and ability explain the effects of design and automation level on responsibility attribution.

4. Empirical examination

We conducted three studies (with one pre-study, total N = 779) to test our hypotheses. Study 1 investigates anthropomorphic car design features (baseline, aggressive, baby-faced) and their influence on responsibility attribution. Based on these initial findings, we replicate the study (baseline vs. baby-faced) to assess the robustness of the results (Study 2). In Study 3, we further explore the connection between babyfacedness and responsibility attribution by examining the mediating effects of benevolence and ability as explanatory variables.

4.1. Pretest: Anthropomorphic design features

Our experimental manipulation of design is applied to the front of the car (the “face”). Grille and headlines were interpreted as mouth and eyes (see Landwehr et al., 2011; Windhager et al., 2010; Aggarwal & McGill, 2007). To manipulate aggressiveness, we build on research by Třebický et al. (2013), endowing the vehicle with slanted headlights (“eyes”) and a larger grille (“mouth”). The manipulation of babyfacedness involved enlargement of the headlights (“eyes”) and a smaller grille (“mouth”, see Miesler et al., 2011), see Figure 1.

Prior to our main studies, we conducted a pre-study to test the efficacy of our manipulations of car designs (baseline vs. aggressive vs. babyface). We recruited participants through Prolific. Participants received monetary compensation and did not have to meet any specific requirements to participate. We tested the manipulations using a between-subject design (N = 113, 63% female, M_{age}= 36.8 years SD = 11.95, Min = 18, Max = 71) in which participants were randomly assigned to one of three groups (n_{base}= 38, $n_{aggressive}$= 36, $n_{babyfaced}$= 39; see Figure 1) and requested to answer our manipulation checks. We measured participants’ perception of a regular car design (“I think the face of the car is regular/typical”, Cronbach’s α = 0.93, AVE = 0.87, CR = 0.93), aggressiveness (“I think the face of the car is aggressive/threatening”; Cronbach’s α = 0.92, AVE = 0.90, CR = 0.95; adapted from Landwehr et al., 2011) and babyfacedness (“I think the face of the car is baby-faced/childlike”; Cronbach’s α = 0.95, AVE = 0.90, CR = 0.95; adapted from Poutvaara et al., 2009) with two items each, all items anchored by strongly disagree (1) and strongly agree (7).

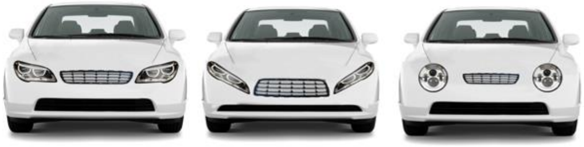

Figure 1. Manipulations for baseline, aggressive, and babyface car design.

Tukey’s multiple comparison tests show that our manipulation checks were successful. Participants perceive the baseline manipulation as more regular than the babyface (mean difference of 2.39, $p < 0.001$) and the aggressive manipulation (mean difference of 1.04, $p < 0.001$). The aggressive car design is perceived as more aggressive than the baseline (mean difference of 1.10, $p < 0.001$) or babyface (mean difference of 1.54, $p < 0.001$), and the babyface design is perceived as more childlike than the baseline (mean difference of 2.46, $p < 0.001$) or the aggressive car design (mean difference of 2.81, $p < 0.001$).

4.2. Study 1: Responsibility attribution

4.2.1. Design and sample. We recruited 262 participants from a European university via social media (M_{age}= 37.02 years, SD = 13.65; 51% female) and assigned them randomly to one of six conditions in a 3 (anthropomorphic car design: baseline vs. aggressive vs. babyface) × 2 (car type: level 3 vs. level 5) between-subjects design. Following a brief survey introduction, we provided participants with contextual information concerning various levels of automated vehicles, which span from level 0, indicating no automation, to level 5, signifying full automation. For our experiment, we specifically utilized stimuli from level 3 and level 5. For level 3 participants read that the vehicle can perform some tasks autonomously but the driver needs to be ready to take over at any time. For level 5, participants read that the car is fully autonomous and that the vehicle performs all tasks without human oversight. The manipulation check for perceived autonomy (“The decisions of this car are (1) completely in the hands of the driver/occupant (7) completely in the hands of the car) is significant at $p < 0.001$, with respondents perceiving the level 5 vehicle as significantly more autonomous than the level 3 vehicle (M_{level3} = 7.62, SD = 2.20; M_{level5} = 8.93, SD = 2.29; F = 22.18, $p < 0.001$). Subsequently, participants were presented with a newspaper article reporting an incident between a bicyclist and the vehicle, resulting in minor injuries to the bicyclist. Additionally, the newspaper article featured a picture of the car showing one of the three anthropomorphic design conditions (baseline vs. aggressive vs. baby-faced). Participants were then asked to indicate the extent to which they attributed responsibility for the collision to the vehicle, the occupant, and the manufacturer using a scale anchored at 1 (= not at all responsible) and 7 (= completely responsible).

4.2.2. Results. To examine the effect hypothesized in H1, two separate regression analyses were run with car design and automation level and their interaction as the independent variables and responsibility attribution as the dependent variable, where the baseline car design was used as the reference category. First, we examined the main effect of car design on responsibility attribution to the vehicle, occupant, and manufacturer. Results show no significant effect of the aggressive design on car responsibility, but a significant negative effect for babyface design (β = -0.75, CI95% = [-1.1, -0.37], $p < 0.001$), implying that a vehicle with a more childlike appearance is attributed with less responsibility. Further, we did not find any main effects of design on responsibility attribution to the occupant or the manufacturer.

In the next step, we examined the main effect of automation level on responsibility. We again conducted regression analysis with responsibility attributions as the dependent variables and car level as the independent variable, using level 3 design as the reference category. Our findings show no main effect on car responsibility. However, there is a negative significant effect for passenger responsibility attribution when the vehicle is autonomous compared to being operated by a driver (β = -1.0, CI95% = [-1.2, -0.61], $p < 0.001$), and, as expected, a significant increase in responsibility attribution to the manufacturer (β = 0.43, CI95% = [0.03, 0.83], $p < 0.05$)

Finally, we investigate the interaction effect between car design and automation level. Results reveal a significant interaction effect between babyface design and automation level on car responsibility, such that highly autonomous cars featuring a babyface design were ascribed a significantly higher level of responsibility attributed to the car compared to the traditional design (β = 0.95, CI95% = [0.38, 1.5], $p < 0.001$), supporting H2. Further, we found a significant interaction effect of baby-faced design and automation for the responsibility of the occupant (β = -0.58, CI95% = [-1.1, -0.07], $p < 0.01$). When the vehicle was baby-faced and autonomous, the occupant received significantly lower levels of responsibility for the accident, while this was reversed for the aggressive design (β = 0.54, CI95% = [0.04, 1.0], $p < 0.05$), where the occupant was seen as more responsible for the accident than in the baseline condition (we did not make any predictions). There was no interaction effect between design and automation level on responsibility attribution to the manufacturer.

4.2.3. Discussion. Non-autonomous vehicles with a babyface design receive lower levels of responsibility, whereas fully autonomous vehicles with a babyface receive higher levels of responsibility (while occupants receive lower levels of responsibility). This shift can be explained by schema congruence theory (Mandler, 1982), which suggests that product evaluation is influenced by the alignment between features and activated human schemas. The competence and capabilities of the vehicle contradict the naive schemas associated with babyface design. While babyface design may connote good intentions in less automated cars, it becomes contradictory when considering the expertise required for autonomous driving. To ensure the robustness of our findings and address potential alternative explanations, we replicate the study controlling for anthropomorphism and technology optimism.

4.3. Study 2: Replication

4.3.1. Design and sample. We conducted a between-subject experiment: 2 (anthropomorphic car design: baseline vs. babyface) × 2 (automation: level 3 vs. level 5), focusing on the effects of a baby-faced schema. Respondents were recruited via social media (as in Study 1) and randomly assigned to one of the four conditions. We realized a sample of N = 146 (M_{age} = 38.08 years, SD = 12.15; Min = 18, Max = 62, 44% female). For scenario design, we relied on the same materials as in Study 1. The manipulation check (identical to Study 1) for babyfacedness is significant at $p < 0.001$, with the baby-faced car design again being perceived as more childlike than the baseline design (M_{base} = 1.72, SE = 1.53, M_{baby} = 6.50, SE = 1.26, F(1,144) = 417.1, $p < 0.001$); while the level 5 car again being perceived as significantly more autonomous than the level 3 car (M_{level3} = 7.95, SE = 2.00, M_{level5} = 9.84, SE = 2.21, F(1,144) = 29.5, $p < 0.001$). Participants were presented with a newspaper article reporting an incident between a bicyclist and the vehicle, resulting in minor injuries to the bicyclist.

4.3.2. Results. We conducted an ANOVA with responsibility attribution towards the car as the dependent variable and anthropomorphic car design (control vs. babyface), automation level (level 3 vs. level 5), and their interaction as independent variables. Consistent with the results of Study 1, our results show a significant negative main effect of car design on responsibility attribution confirming that a childlike-looking car (M = 4.74, SE = 1.76) is attributed less responsibility in an accident situation than the control car (M = 6.03, SE = 1.14, F(1,142) = 51.75, $p < 0.001$). There was a significant main effect of automation level (level3: M = 4.55, SE = 1.50; level 5: M = 6.37, SE = 1.07, F(1,142) = 108.97, $p < 0.001$). As hypothesized, we find a significant interaction between car design and automation level (F(1,142) = 11.83, $p < 0.001$). A babyface car design is significantly more responsible in the fully autonomous condition (M = 6.00, SE = 1.30) than in the lower autonomy condition (M = 3.47, SE = 1.13).

4.3.3. Robustness checks. In addition, participants answered eleven items measuring a trait-like tendency to anthropomorphize in general (Waytz et al., 2014 Neave et al., 2015; sample item: "I sometimes wonder if my computer deliberately runs more slowly after I have shouted at it.", Cronbach's α = 0.93, AVE = 0.55, CR = 0.93) and a five-item measure of technology optimism (Parasuraman, 2000, sample item: "Technology allows people to have more control over their daily lives."; Cronbach's α = 0.88, AVE = 0.64, CR = 0.89). An ANOVA with these two constructs as controls reveals that the main and interaction effects did not change in direction or significance. We find a significant direct effect of anthropomorphism on responsibility attribution (F(1, 140) = 5.79, p < 0.05).

4.3.4. Discussion. Study 2 successfully replicates the effects found in Study 1. Next, we aim to shed light on the mechanisms that transmit the interaction between design and automation level on responsibility attribution.

4.4. Study 3: The role of benevolence and ability

4.4.1. Design and sample. The outcomes obtained from Studies 1 and 2 have established that a babyface design for AVs backfires because it receives more blame for an accident than a control design. We intend to explore the underlying mechanisms that account for this effect. Participants who were recruited on Prolific received monetary compensation and did not have to meet any specific requirements to participate. We used a 2 (anthropomorphic car design: baseline vs. babyface) × 2 (autonomy level: level 3 vs. level 5) between-subjects design in which participants were randomly assigned to one of the four conditions (N = 258, M_{age}= 37.47 years, SD = 12.43, Min = 18, Max = 77, 40% female, 7% students). All manipulation checks were again successful at p < 0.001, with the level 5 car being perceived as more autonomous than the level 3 car (M_{level3} = 3.79; M_{level5}= 5.21, F (1,256) = 58.63, p < 0.001) and the babyface car design being perceived as more baby-faced than the baseline design ($M_{control}$= 2.56, SD = 1.43; M_{baby}= 4.94, SD = 1.42; F(1,256) = 179.4, p < 0.001). Next, participants read the newspaper article as in Study 1. In addition to the constructs used in the earlier studies, we included assessments of car's benevolence (sample item: "I feel very confident about the cars' skills."; Song & Luximon, 2021, Cronbach's α = 0.91, AVE = 0.66, CR = 0.90) and ability (sample item: "The car seems concerned about others' welfare."; Song & Luximon, 2021, Cronbach's α= 0.93, AVE = 0.77, CR = 0.93).

4.4.2. Results. We ran regressions to test the postulated relationships and found that a babyface car design exerts a substantial positive impact on the perception of benevolence (β = 1.11, t = 5.07, CI95% = [0.68, 1.54], p < 0.001). However, we failed to discern a significant influence of car level (β = 0.14, t = 0.68, CI95% = [-0.28, 0.58], p > 0.1). Notably, an examination of the interaction effect of car design and car level revealed a significant negative interaction effect of a baby-faced autonomous car (β = −0.82, t = −2,67, CI95% = [-1.43, -0.21], p < 0.001). The conditional effects additionally unveiled that the introduction of a baby-faced design yields a significant positive effect on the perception of benevolence when autonomy is low (level 3) (β = 1.11, t = 5.07, CI95% = [0.68, 1.54]); conversely, no significant effect is observed when autonomy is high (level 5) (β = 0.28, t = 1.32, CI95% = [-0.14, 0.71]). Furthermore, our analysis reveals a lack of significant effect concerning either car design (β = 0.27, t = 1.14, CI95% = [-0.19, 0.73], p > 0.1) or car level (β = 0.01, t = 0.04, CI95% = [-0.45, 0.47], p > 0.1) on ability perception. However, a significant negative interaction effect was identified towards ability and a babyface autonomous car, signifying that an autonomous vehicle with a childlike appearance is perceived as less competent (β = -0.74, t = -2.24, CI95% = [-1.40, -0.09], p < 0.05). Upon scrutinizing the conditional effects, our results lend credence to the notion that a car featuring a baby-faced design does not exert a significant impact on benevolence perception when automation is low (level 3) (β = 0.27, t = 1.14, CI95% = [-0.19, 0.73]). In contrast, a significant negative effect is observed when automation is high (level 5) (β = -0.47, t = -2.03, CI95% = [-0.93, -0.02]).

In the analysis of the relationship between benevolence and responsibility attribution, no discernible significant effect was observed (β = -0.02, t = 0.10, CI95% = [-0.23, 0.19], p > 0.1). Conversely, the impact of ability on responsibility attribution was found to be negative, thereby affirming the hypothesis (β= -0.44, t = -4.35, CI95% = [-0.64, -0.24], p < 0.001). Consequently, it can be inferred that as the perceived ability of a car increases, the corresponding allocation of responsibility decreases.

Finally, we conducted a moderated mediation analysis (Model 7 in PROCESS; Hayes, 2017) using 10,000 resamples with design as the independent variable, autonomy level as the moderator, ability and benevolence as the mediators, and responsibility attribution to the vehicle as the dependent variable. Ability mediated the effect of design on responsibility in the level 5 condition (ß = 0.22, CI95% = [0.001,0.49]) but not in the level 3 condition (b = -0.11, CI95% = [-0.32,0.10]). Index of moderated mediation: b = 0.33, CI95% = [0.03, 0.68]; indirect-only. There was a main effect of design on benevolence but no effect of the mediator on responsibility attribution.

4.4.3. Discussion. The results provide important insights into the mechanisms explaining the effect of design on responsibility attribution. A babyface vehicle appears to be more benevolent (but not more able) when it exhibits a low automation level. Yet, an autonomous baby-faced vehicle appears to be both less benevolent and less able and is ascribed more responsibility in case of an accident. These findings align with the existing body of research on babyfacedness, which attributes benevolent intentions to entities falling within the babyface-schema, but also characterizes them as exhibiting more naive behavior and diminished capabilities in case of negative behavior (Gorn et al., 2018).

5. Theoretical and practical implications

While existing research on the anthropomorphic design of vehicles generally highlights their positive effects, our research reveals that in the context of autonomous vehicles, these elements can have unintended negative consequences. Specifically, the presence of immature-looking features in the car front can lead to adverse effects. We find that cute features in autonomous cars reduce inferred ability, amplifying the attribution of responsibility in the event of an accident. A potential explanation is the mismatch between the expected capabilities of fully autonomous vehicles and the naive schemas associated with a babyface design. When an entity that is perceived as cute behaves in a harmful way (e.g., injures a cyclist), there is a cognitive dissonance between the expected endearing and innocent qualities associated with its design and the actual behavior of the entity (Festinger, 1957; Harmon-Jones & Mills, 1999), which we refer to as the *cuteness paradox*. Consequently, this dissonance reinforces the attribution of responsibility to the entity in question. Considering the popularity of the cuteness strategy in conventional vehicle design this finding is of great practical relevance to manufacturers and designers.

From a theoretical perspective, the current research contributes to three interrelated research streams: anthropomorphism, baby schema, and social perception of technology. The literature has mostly focused on positive outcomes of anthropomorphism and research on detrimental outcomes is rather scarce (for an exception see Kim et al., 2016). Our findings emphasize its multifaceted nature and the imperative to rigorously explore the spectrum of anthropomorphic effects – also the potentially detrimental ones (see also Kim et al., 2016). By linking the literature on anthropomorphism with the psychological face shape literature (Oosterhof & Todorov, 2008), we show that triggering human schemas through design elements can change the perception of technology by affecting inferred ability. Inferences from faces are well documented in traditional research on human subjects (e.g., Oosterhof & Todorov, 2008) but not when it comes to technology. Our study elucidates that even nuanced design elements can substantially influence the impressions of a car and how responsibility is attributed. This sheds light on the necessity to re-examine and adapt existing theories of responsibility attribution, especially in light of the pervasive evolution of technology.

We also add to the baby schema literature and extend it to the context of non-human entities. While the effects of the baby schema are well understood in the context of human faces (e.g., Gorn et al., 2008), research in the context of technology, and in particular vehicles, is scarce. We show that there are important similarities and differences compared to the human context, and we examine an important moderator of its effects: autonomy. Technology with higher levels of autonomy suffers from a baby-faced design, while technology with low levels of autonomy is seen as more benevolent and receives lower levels of responsibility in the case of "wrongdoing". These findings open up possibilities for further research, as robots, avatars, and other entities with human-like "faces" may suffer from immature design elements depending on their level of autonomy.

Once viewed primarily as a tool, technology's anthropomorphic attributes are changing its relationship with users. There is a growing literature on social perception of technology, with previous work focusing on intent detection in AI systems (McKee et al., 2021). As discussed above, inferring intent and ability becomes more important for vehicles as they become more autonomous. By showing that perceptions of ability and benevolence – which are closely related to social impressions, as outlined in the Stereotype Content Model (Fiske et al., 2002) – are systematically dependent on design features (i.e., cues in the front of the car), we demonstrate that social categorization can be easily triggered and provide a systematic approach to design options for autonomous cars. By demonstrating that deep-seated cognitive biases shape users' perceptions of technology and its capabilities, we show that design can create a mismatch between the technology's objective and inferred capabilities. This gap suggests the existence of underlying psychological processes that may shape our interactions with advanced technologies and warrants further attention from human-computer interaction researchers.

From a practical perspective, this research shows that anthropomorphism needs to be implemented in the right way – because appearance biases impressions not only of other people but also of technology. *Manufacturers* should be aware of the unintended adverse consequences of anthropomorphizing vehicles. In particular, design elements that suggest naiveté or cuteness can lead to a higher attribution of responsibility in the case of accidents. This is even more relevant considering how popular the cuteness strategy is for conventional vehicles (see e.g. Fiat 500, Volkswagen Beetle; Mini). The findings also have *legal and regulatory implications*, as current regulatory frameworks often have difficulty assigning responsibility in accidents involving AVs. By understanding how design features affect the allocation of responsibility, we can promote the development of more sophisticated and effective regulations. The results may also be relevant to other areas of *human-AI collaboration*. In contexts where artificial intelligence is used to help humans overcome performance-related challenges, understanding the role of anthropomorphism in assigning responsibility is critical. This knowledge can guide the development of AI systems that are perceived as more capable and trustworthy, ultimately leading to better user acceptance and more effective collaboration. Our findings transcend the realm of car designers and manufacturers and have broader implications for the design of anthropomorphized technology in various domains, including humanoid robots and avatars.

6. Limitations and future research

Our findings contribute to a better understanding of anthropomorphized technology and the intricate mechanisms involved in the attribution of responsibility. However, it is important to note that our research focuses only on anthropomorphic design elements in the vehicle front and explores only one type of accident, without further differentiating between different levels of severity. Future research endeavors could therefore investigate how responsibility attribution varies across different levels of severity. By doing so, these studies can provide valuable insights that can inform the development of an appropriate design framework for human-like devices.

7. References

Aggarwal, P., & McGill, A. L. (2007). Is that car smiling at me? Schema congruity as a basis for evaluating anthropomorphized products. *Journal of Consumer research, 34*(4), 468-479.

Bansal, P., & Kockelman, K. M. (2018). Are we ready to embrace connected and self-driving vehicles? A case study of Texans. *Transportation, 45*, 641-675.

Beckers, N., Siebert, L. C., Bruijnes, M., Jonker, C., & Abbink, D. (2022). Drivers of partially automated vehicles are blamed for crashes that they cannot reasonably avoid. *Scientific Reports, 12*(1), 16193.

Carré, J. M., McCormick, C. M., & Mondloch, C. J. (2009). Facial structure is a reliable cue of aggressive behavior. *Psychological Science, 20*(10), 1194-1198.

Charness, N., Yoon, J. S., Souders, D., Stothart, C., & Yehnert, C. (2018). Predictors of attitudes toward autonomous vehicles: The roles of age, gender, prior knowledge, and personality. *Frontiers in psychology, 9*, 2589.

Choi, J. K., & Ji, Y. G. (2015). Investigating the importance of trust in adopting an autonomous vehicle. International *Journal of Human-Computer Interaction, 31*(10), 692–702.

Colquitt, J. A., Scott, B. A., & LePine, J. A. (2007). Trust, trustworthiness, and trust propensity: A meta-analytic test of their unique relationships with risk taking and job performance. *Journal of Applied Psychology, 92*(4), 909–927.

Copp, C. J., Cabell, J. J., & Kemmelmeier, M. (2023). Plenty of blame to go around: Attributions of responsibility in a fatal autonomous vehicle accident. *Current Psychology, 42*(8), 6752-6767.

Fiske, S. T., Cuddy, A. J., Glick, P., & Xu, J. (2002). A Model of (Often Mixed) Stereotype Content: Competence and Warmth Respectively Follow From Perceived Status and Competition. *Journal of Personality and Social Psychology*, 82(6), 878-902.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition, 108*, 353–380.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., ... & Williams, M. D. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management, 57*, 101994.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: a three-factor theory of anthropomorphism. *Psychological review, 114*(4), 864.

Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Stanford University Press.

Forster, Y., Naujoks, F., & Neukum, A. (2017, June). Increasing anthropomorphism and trust in automated driving functions by adding speech output. In *2017 IEEE intelligent vehicles symposium* (IV), 365-372.

Gabarro, J. J. (1978). The development of trust, influence and expectations. *Interpersonal behavior: Communication and understanding in relationships*, 290-303.

Gazzola, V., Rizzolatti, G., Wicker, B., & Keysers, C. (2007). The anthropomorphic brain: the mirror neuron system responds to human and robotic actions. *Neuroimage, 35*(4), 1674-1684.

Gorn, G. J., Jiang, Y., & Johar, G. V. (2008). Babyfaces, trait inferences, and company evaluations in a public relations crisis. *Journal of Consumer Research, 35*(1), 36-49.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science, 315*, 619.

Harmon-Jones, E., & Mills, J. (1999). *Cognitive Dissonance: Progress on a Pivotal Theory in Social Psychology*. American Psychological Association.

Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford publications.

Hein, D., Rauschnabel, P., He, J., Richter, L. and Ivens, B., (2018). What drives the adoption of autonomous cars?, *ICIS 2018 Proceedings*.

Kiesler, S., Powers, A., Fussell, S. R. and C. Torrey (2008). Anthropomorphic interactions with a robot and robot–like agent. *Social Cognition 26*(2), 169−181.

Kim, S., Chen, R. P., & Zhang, K. (2016). Anthropomorphized helpers undermine autonomy and enjoyment in computer games. *Journal of Consumer Research*, 43(2), 282-302.

Kraus, S., Althoff, M., Heißing, B., & Buss, M. (2009). Cognition and emotion in autonomous cars. In *2009 IEEE intelligent vehicles symposium*, 635-640.

Landwehr, J. R., McGill, A. L., & Herrmann, A. (2011). It's got the look: The effect of friendly and aggressive "facial" expressions on product liking and sales. *Journal of Marketing, 75*(3), 132-146.

Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics, 35*(10), 1243-1270.

Lee, J. G., Kim, K. J., Lee, S., & Shin, D. H. (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.

Mandler, G. (1982). The structure of value: Accounting for taste. In M. S. Clark & S. T. Fiske (Eds.), *Affect and cognition: The 17th annual Carnegie symposium* (pp. 3-36). Hillsdale, NJ: Erlbaum.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*, 709 –734.

McKee, K. R., Bai, X., & Fiske, S. T. (2023). Humans perceive warmth and competence in artificial intelligence. *Iscience*, 26(8).

McManus, R. M., & Rutchick, A. M. (2019). Autonomous vehicles and the attribution of moral responsibility. *Social psychological and personality science, 10*(3), 345-352.

Miesler, L., Leder, H., & Herrmann, A. (2011). Isn't it cute: An evolutionary perspective of baby-schema effects in visual product designs. *International Journal of Design, 5*(3), 17-30.

Mondloch, C. J., Lewis, T. L., Budreau, D. R., Maurer, D., Dannemiller, J. L., Stephens, B. R., & Kleiner-Gathercoal, K. A. (1999). Face perception during early infancy. *Psychological science, 10*(5), 419-422.

Neave, N., Jackson, R., Saxton, T., & Hönekopp, J. (2015). The influence of anthropomorphic tendencies on human hoarding behaviours. *Personality and Individual Differences, 72*, 214-219.

Niu, D., Terken, J., & Eggen, B. (2018). Anthropomorphizing information to enhance trust in autonomous vehicles. *Human Factors and Ergonomics in Manufacturing & Service Industries, 28*(6), 352-359.

Nittono, H., Fukushima, M., Yano, A., & Moriya, H. (2012). The power of kawaii: Viewing cute images promotes a careful behavior and narrows attentional focus. *PloS one, 7*(9), e46362.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences, 105*(32), 11087-11092.

Parasuraman, A. (2000). Technology Readiness Index: A Multiple-Item Scale to Measure Readiness to Embrace New Technologies. *Journal of Service Research, 2*(4), 307-320

Peterson, J. C., Uddenberg, S., Griffiths, T. L., Todorov, A., & Suchow, J. W. (2022). Deep models of superficial face judgments. *Proceedings of the National Academy of Sciences, 119*(17), e2115228119.

Pierce, J. R., Kilduff, G. J., Galinsky, A. D., & Sivanathan, N. (2013). From glue to gasoline: How competition turns perspective takers unethical. *Psychological science, 24*(10), 1986-1994.

Poutvaara, P., Jordahl, H., & Berggren, N. (2009). Faces of politicians: Babyfacedness predicts inferred competence but not electoral success. *Journal of Experimental Social Psychology, 45*(5), 1132-1135.

SAE International. (2014). Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems (J3016) (pp. 1–12).

Schoettle, B., & Sivak, M. (2016). *Motorists' preferences for different levels of vehicle automation (Report No. SWT-2016-8).* Ann Arbor: University of Michigan Transportation Research Institute, Sustainable Worldwide Transportation.

Shariff, A., Bonnefon, J. F., & Rahwan, I. (2017). Psychological roadblocks to the adoption of self-driving vehicles. *Nature Human Behaviour, 1*(10), 694-696.

Soanes, C., & Stevenson, A. (Eds.). (2005). *Oxford dictionary of English* (2nd ed.). New York: Oxford University Press.

Song, Y., & Luximon, Y. (2021). The face of trust: The effect of robot face ratio on consumer preference. *Computers in Human Behavior, 116*, 106620.

Třebický, V., Havlíček, J., Roberts, S. C., Little, A. C., & Kleisner, K. (2013). Perceived aggressiveness predicts fighting performance in mixed-martial-arts fighters. *Psychological Science, 24*(9), 1664-1672.

Wang, J., Zhang, L., Huang, Y., Zhao, J., & Bella, F. (2020). Safety of autonomous vehicles. *Journal of advanced transportation, 2020*, 1-13.

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

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological science, 17*(7), 592-598.

Wilson, J. P., & Rule, N. O. (2015). Facial trustworthiness predicts extreme criminal-sentencing outcomes. *Psychological science, 26*(8), 1325-1331.

Windhager, S., Slice, D. E., Schaefer, K., Oberzaucher, E., Thorstensen, T., & Grammer, K. (2008). Face to face: The perception of automotive designs. *Human Nature, 19*, 331-346.

Zebrowitz, Leslie A. and Susan M. McDonald (1991). The Impact of Litigants' Baby-Facedness and Attractiveness on Adjudications in Small Claims Courts. *Law and Human Behavior, 15*(6), 603–23.