미래사회
Journal of Future Society
2025. Vol. 16, No. 3, 218~239.
https://doi.org/10.22987/jifso.2025.16.3.218

Human-Likeness and Moral Responsibility in AI: The Roles of Dyadic Morality and Emotional Unpleasantness[*]

Eunjin Yoon[**] Eunkyoung Chung[***]

Kangwon National University

This study examines how the perceived human-likeness of artificial intelligence (AI) influences judgments of AI's moral responsibility. Specifically, it investigates (1) the direct effect of human-likeness on moral responsibility, (2) the sequential mediating roles of perceived moral patiency and perceived moral agency, and (3) the moderating effect of emotional unpleasantness on the relationship between human-likeness and moral patiency. Data were collected from 270 adult participants in South Korea using an online survey. Results revealed a significant negative direct effect of human-likeness on moral responsibility, contrary to initial expectations. However, a significant positive indirect effect emerged through a sequential mediation pathway involving patiency and agency, suggesting a suppression effect. Emotional unpleasantness further moderated the association between human-likeness and moral patiency, such that the relationship was stronger when unpleasantness was high. These findings provide new insights into how the physical appearance of AI, moral cognition, and affective responses interact to influence moral evaluations in human-AI interactions.

Keywords: artificial intelligence, dyadic morality, emotional unpleasantness, human-likeness, moral responsibility

[*] This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2023S1A5A2A01078593)

[**] First author: Eunjin Yoon/Completed Ph.D. Coursework, Department of Psychology, Kangwon National University/
1, Kangwondaehak-gil, Chuncheon-si, Gangwon-do, Republic of Korea

[***] Corresponding author: Eunkyoung Chung/Professor, Department of Psychology, Kangwon National University/
1, Kangwondaehak-gil, Chuncheon-si, Gangwon-do, Republic of Korea/
Tel: 033-250-6854/E-mail: ekchung@kangwon.ac.kr

Ⅰ. Introduction

As artificial intelligence (AI) has become increasingly integrated into everyday life, the number of incidents involving the harm caused by AI systems is also increasing. A notable case occurred in Belgium in 2023, where a man died by suicide after interacting with the chatbot "Eliza" for six weeks. The chatbot reportedly encouraged suicidal ideation by affirming his belief that his death could save the world and suggesting that they could reunite in heaven (El Atillah, 2023). More recently, the parents of a 14-year-old boy filed a lawsuit against another chatbot company, claiming that their son's addiction to the chatbot contributed to his suicide (Dodd, 2024).

These incidents have intensified the debate on the moral responsibility of AI. Although these discussions have traditionally been rooted in moral philosophy, several scholars have challenged the exclusion of nonhuman agents from moral considerations. For instance, Gunkel (2012) criticized conventional ethical theories for unilaterally denying moral standing to machines and advocated for a relational, inclusive ethical framework that could encompass AI. Floridi and Sanders (2004) introduced the notion of "mind-less morality," which removes the requirement of consciousness or intentionality and redefines moral agency more broadly. In line with these inclusive approaches, Noh (2022) proposed a "literalism" grounded in posthumanist semantics, arguing that psychological and moral categories (e.g., decision, intention, and suffering) can be extended beyond humans when they are applied consistently within scientific explanatory models. This perspective may provide a conceptual foundation for considering AI not merely as a metaphorical subject but also as a legitimate expansion of moral categories.

However, some scholars argue that consciousness is a prerequisite for moral standing. For example, Mosakas (2021) claimed that phenomenal consciousness is essential for granting moral status to social robots. In addition, Himma (2009) contended that moral agency must involve rational attributes such as consciousness, deliberation, and free will—features that current AI lacks—making moral responsibility attribution inappropriate. Torrance (2008) further argued that moral agency presupposes moral patiency, the capacity to be a moral subject, and requires the ability to fulfill ethical obligations.

Apart from philosophical debates about whether AI *should* be granted a moral status, empirical studies suggest that people *do* attribute moral agency and patiency to it. As the incidents mentioned earlier illustrate, people may perceive AI not merely as a tool but as an entity capable of acting or suffering. Prior research has shown that anthropomorphism is significantly influenced by human-likeness, which is the degree to which an AI resembles a human in appearance or behavior (Epley et al., 2007; Duffy, 2003; Fink, 2012; Kahn et al., 2007; Fong et al., 2003).

This study examined how the human-likeness of AI influences the perception of moral responsibility. Drawing on the theory of dyadic morality (Gray & Wegner, 2012a), this study investigated the mediating roles of perceived moral agency (the capacity for intentional action) and perceived moral patiency (the capacity to experience suffering). Although previous research emphasized perceived agency as a key determinant of moral responsibility attribution (Monroe et al., 2014; Shank & DeSanti, 2018), the role of patiency has received comparatively less attention. Patiency is typically associated with being a victim, but few studies have empirically examined how it interacts with agency to shape moral judgment.

To fill this gap, this study analyzed a sequential mediation model in which AI's human-likeness affects perceptions of experience and agency, influencing moral responsibility judgments. This approach expands the theoretical understanding of how the two moral dimensions—of agency and patiency—jointly shape AI-related moral evaluations.

Additionally, this study explored the moderating role of emotional unpleasantness, a response commonly triggered by highly human-like AI and conceptualized as the uncanny valley phenomenon. Such discomfort may distort or intensify the perceptions of an AI's moral capacities. By incorporating emotional responses into the analysis, this study aims to provide a more comprehensive account of how human-likeness influences moral judgments.

Through these analyses, this study contributes to both the empirical and theoretical discussions on AI ethics by elucidating the psychological mechanisms through which people ascribe moral significance to human-like machines.

Ⅱ. Human—Likeness of AI and Moral Judgment

Human—likeness refers to the degree to which an AI or robot resembles a human in its external appearance, encompassing visual cues such as facial structure, skin texture, eye direction, and body shape (Kätsyri et al., 2015). Prior studies have demonstrated that as human—likeness increases, individuals are more likely to attribute psychological characteristics such as thoughts, emotions, intentions, and desires to AI or robots (Kühne & Peter, 2023). For example, human—like robots tend to elicit greater trust and likability from users and increase intentions for cooperation and interaction (You & Robert, 2018), and evoke perceptions of internal states such as emotions and intentions, particularly when equipped with features such as eyes (Linnunsalo et al., 2023).

This inclination to attribute human-like qualities to entities is commonly referred to as anthropomorphism. Anthropomorphism is a cognitive and psychological process through which individuals uniquely project human mental states, such as emotions, intentions, and subjective experiences, onto AI systems (Epley et al., 2007). According to Waytz et al. (2010), anthropomorphism can occur through two distinct routes: physical and psychological. Physical anthropomorphism involves perceiving human-like bodily features, such as facial structures or limbs, whereas psychological anthropomorphism involves attributing internal mental states. Physical anthropomorphism is typically observed in humanoid robots, whereas psychological anthropomorphism can arise even in nonembodied systems, such as chatbots or text-based digital assistants that display human-like interaction patterns.

Human−likeness is a key driver of anthro−pomorphism. Numerous studies have shown that the more a robot or AI resembles a human, the more likely people are to assign human−like qual−ities to it (Bartneck et al., 2009; Broadbent et al., 2013; Fong et al., 2003; Powers & Kiesler, 2006). One such quality is the capacity for moral judg−ment, a defining feature of humanity and a core dimension of anthropomorphism. Therefore, hu−man−likeness may influence how people evaluate AI morally.

Research has demonstrated that when robots or AI agents display human-like appearance or behavior, people tend to apply moral standards to judge their actions. For instance, in moral dilemma scenarios such as the trolley problem, participants evaluated highly human-like robots more strictly and held them to moral standards similar to those of humans (Malle et al., 2016). Moreover, Kahn et al. (2012) found that people attributed higher levels of moral responsibility for causing harm to humanoid robots with more human-like appearances. A large-scale study by Ladak et al. (2024) identified a human-like body and prosocial behavior as the two most significant predictors of moral considerations of AI. Notably, the study emphasized that physical similarity enhances the perception that an AI possesses a "mind." Based on this theoretical framework, the following hypothesis was proposed:

Hypothesis 1 (H1): The perceived human−like−ness of an AI is positively associated with its perceived moral responsibility.

Ⅲ. Dyadic Morality as Mediators

Attributing the mind to nonhuman entities is a foundational step in perceiving them as morally accountable. By assigning human-like mental capacities to AI, anthropomorphism allows people to view them as autonomous agents capable of intentional action, thereby rendering them potential targets of moral praise or blame (Waytz et al., 2010). Supporting this notion, Stuart and Kneer (2021) found that individuals attribute moral blame to AI systems when they perceive the AI to possess knowledge and desires. Similarly, entities perceived to have high levels of agency tend to be judged more harshly for transgressions (Gray et al., 2007), and AI perceived to be making purposeful decisions is more likely to be assigned moral responsibility (Monroe et al., 2014).

The mind perception theory (Gray et al., 2007) explains how individuals attribute the mind to entities along two dimensions: agency (capacity for intention, planning, and communication) and experience (capacity for sensations and emotions such as pain or joy). Adult humans are typically seen as high in both dimensions, whereas robots are often attributed with moderate agency but low experience.

Subsequently, Gray and Wegner (2009) extended this framework to the moral domain, clarifying that moral agency refers to "the capacity to do right or wrong," whereas moral patiency refers to "the capacity to be a target of right or wrong." These two dimensions are central to the theory of dyadic morality (Gray & Wegner, 2012a), which links agency to moral agency (i.e., the ability to act as a perpetrator) and experience to moral patiency (i.e., the ability to be harmed). Adult humans, who are perceived as both agents and patients, are held morally responsible. In contrast, entities such as children or animals, high in experience but low in agency, are held less accountable for harm.

When applied to AI, perceived increases in both agency and experience may enhance moral responsibility judgments. The theory of dyadic morality suggests that attributing either dimension to the mind can influence subsequent moral responses. However, previous findings have indirectly addressed these dimensions, focusing on their affective and behavioral consequences rather than on explicit moral evaluations. For instance, Yam et al. (2021) found that when robots were perceived as having feelings (i.e., sentience), users evaluated them more positively and were more forgiving of their mistakes. This result reflects how the perception of sentience—though conceptually distinct from moral patiency—can evoke empathic concern and leniency, processes that conceptually align with the protective side of moral patiency within a dyadic framework. Similarly, Yam et al. (2022) showed that anthropomorphized robot supervisors are perceived as more agentic, which increases user resistance to negative feedback. Although this resistance is not a direct moral judgment, it illustrates how heightened perceptions of agency can elicit social responses consistent with moral blame. In other words, perceiving agency tends to activate the "moral agent" side of the dyad, leading to reactions such as defensiveness or retributive behavior. Overall, these studies extend the dyadic morality perspective by demonstrating that even non-moral outcomes, such as forgiveness or resistance, arise from the same underlying mind-perception processes that lead to moral judgments.

Choi and Jang (2022) explored the mediating role of moral agency and patiency in the relationship between AI anthropomorphism and moral responsibility judgments. Perceived moral patiency significantly mediated this relationship, whereas perceived agency did not. This contrasts with earlier findings that emphasized the role of agency. A potential explanation lies in their parallel mediation model, which assumes that anthropomorphism independently increases perceived agency and patiency. Building on prior research in moral psychology and dyadic morality, we posit that perceived moral patiency precedes perceived moral agency. Emotional sensitivity and empathy, which are the core components of patiency, provide a foundation for moral cognition (Haidt, 2001; Moll et al., 2003). As Gray and Wegner (2012b) showed, human-like cues primarily elicit perceptions of vulnerability and sentience before the cognitive attributions of agency emerge. This sequential process is consistent with the neurosentimentalist view (Gerrans & Kennett, 2010) and the developmental findings that moral understanding begins with affective attunement before evaluative reasoning (Nucci & Gingo, 2010). Thus, perceived patiency can be conceptualized as a psychological gateway to moral agency. Therefore, we propose Hypothesis 2.

Hypothesis 2 (H2): Perceived moral patiency and perceived moral agency sequentially mediate the relationship between AI human-likeness and perceived moral responsibility, with higher human-likeness increasing perceived patiency, which in turn increases perceived agency, finally leading to greater moral responsibility judgments.

Ⅳ. Moderating Role of Emotional Unpleasantness

Discomfort elicited by highly human-like robots, known as the uncanny valley effect, was first introduced by Mori (1970). He proposed that as robots become more human-like, they are initially perceived more favorably—but when they become “almost” human, a sharp drop in likability occurs, evoking a sense of eeriness. This phenomenon has been widely replicated and extended in subsequent studies (Cheetham et al., 2011, 2014; Ferrey et al., 2015; Gray & Wegner, 2012b).

Several theories have been proposed to explain the emotional response. MacDorman and Entezari (2015) framed it as an evolutionarily adaptive disgust response rooted in pathogen avoidance: entities that appear nearly human but are not slightly off in appearance may signal disease risk. Cheetham et al. (2011, 2014) emphasized categorization ambiguity, suggesting that difficulty in classifying human-like machines creates perceptual conflict. Ferrey et al. (2015) further proposed the inhibitory-devaluation hypothesis, in which cognitive inhibition in response to ambiguous stimuli leads to a negative affect, similar to the mechanisms observed in cognitive dissonance. Diel et al. (2021) summarized the contributing factors, including perceptual mismatches, behavioral expectation violations, and category conflicts between human and robotic cues.

These emotional reactions significantly affect social and moral evaluations. For instance, Song and Shin (2024) found that increasing the human-likeness of chatbots in e-commerce contexts heightens eeriness, lowers trust, and reduces consumer intent. Laakasuo et al. (2021, 2023) demonstrated that robots with human-like faces that make moral decisions are judged as less moral, because of the affective discomfort they trigger. Shin et al. (2019) demonstrated that highly realistic robotic avatars reduce information processing and judgment accuracy.

Prior research suggests that emotional unpleasantness affects perceptions of moral patiency —the capacity to feel and suffer—more strongly than moral agency (Gray & Wegner, 2012b; Tsukimoto, 2017). When human-likeness induces eeriness, it may sharpen attention to the robot’s capacity to feel emotions, thereby intensifying the perceptions of vulnerability or sentience. In other words, the affective salience created by the uncanny valley may amplify the attribution of emotional capacity, and enhance perceived moral patiency. Based on this theoretical reasoning, we proposed the following hypothesis:

Hypothesis 3 (H3): Emotional unpleasantness moderates the relationship between the perceived human-likeness of AI and perceived moral patiency, rendering this relationship stronger when emotional unpleasantness is high.

Ⅴ. Method

1. Participants and Procedures

In total, 270 adults (135 men and 135 women) residing in South Korea participated in this online study. The age of the participants ranged from 20 to 50 years, and they were recruited through Embrain, a professional online survey panel provider. Stratified sampling was used to ensure balanced representation across gender and age groups. All participants provided informed consent before participation and received a monetary incentive of 3,000 KRW for completing the study. This study was approved by the Institutional Review Board of the Kangwon National University. The participants were randomly assigned to one of three conditions that varied in the degree of human–likeness of the AI agent. Each participant was first presented with an image of the AI agent along with a description of a tax–related task. After this initial exposure, they completed a set of questionnaires measuring perceived human–likeness, emotional unpleasantness, and dyadic morality (i.e., perceived moral patiency and agency). Next, the participants read a scenario in which the AI tax agent committed a service failure. They then evaluated the AI's responsibility for the failure and completed demographic questions.

2. Measurements

(1) Human–Likeness of AI

The perceived human–likeness of the AI agent was manipulated across the three conditions using visual stimuli. In the low human–likeness (HL) condition, an image of a generic laptop (nonanthropomorphic) was presented. In the middle HL and high HL conditions, images of humanoid robots with varying degrees of human–likeness were used. To select the appropriate stimuli, a pilot study was conducted to evaluate three robot images of perceived HL and personification, based on the criteria used by Choi and Jang (2022). The two images that scored the lowest and highest on both dimensions were selected to ensure a meaningful contrast between the conditions. Across all conditions, participants were introduced to the AI as a government–affiliated tax agent with the following standardized description.

“*The government has recently introduced TAX-76, an advanced AI developed by Company K for tax administration. TAX-76 is the first AI capable of making tax-related decisions by utilizing complex tax-related data and personal information. It is known to significantly improve the efficiency of complex tax-related tasks and decision-making processes.*”

The only variation across conditions was the name and visual appearance of the AI (Low HL: TAX–76, Middle HL: Coco and High HL: James). To assess the perceived HL of the AI, participants responded to two items adapted from Mathur and Reichling (2016): “How mechanical does this AI look?” (reverse–coded) and “How human does this AI look?” Responses were recorded on a 7–point Likert scale, with higher scores indicating greater perceived HL. The internal consistency of the two-item scale was acceptable (Cronbach's α = .61).

(2) Perceived Moral Agency and Moral Patiency

Participants' perceptions of the moral mind of AI were measured using items adapted from Yam et al. (2021). Perceived moral agency was assessed with four items that evaluated the extent to which the AI was considered capable of intentional and autonomous action (e.g., "This AI can plan its own actions" and "This AI can think"). Perceived moral patiency was measured with four items assessing the AI's capacity for emotional and sensory experience (e.g., "This AI can feel pain" and "This AI can feel happiness"). All items were rated on a 7-point Likert scale (1 = strongly disagree and 7 = strongly agree), with higher scores indicating stronger agreement with the respective dimensions. Internal consistency was high for both subscales (α = .84 for moral agency and α = .95 for moral patiency).

In addition, the "agency" items primarily reflect intentional stance adoption and index intentionality attribution, rather than full-fledged moral agency. Similarly, the "patiency" items capture sentience, which is conceptually related to—but not identical with—moral patiency. In line with the theory of dyadic morality (Gray & Wegner, 2009, 2012), we interpret these dimensions as proxies for broader constructs of moral agency (the capacity to do right or wrong) and moral patiency (the capacity to be a target of right or wrong). Thus, although our measures operationalized intentionality and sentience at the item level, we analyzed them as perceived moral agency and moral patiency, which is consistent with prior empirical research.

(3) Emotional Unpleasantness

Emotional unpleasantness toward the AI tax agent was measured using four adjective-based items that capture negative emotional responses: *uncanny, creepy, scary,* and *eerie.* These items were adapted from the emotional response dimensions proposed by Kim et al. (2022). Participants were asked to indicate, for example, "*How creepy is this?*", using a 7-point Likert scale ranging from 1 (not at all) to 7 (very much). Higher scores indicate greater emotional unpleasantness. The internal consistency of the scale was excellent (Cronbach's α = .92).

(4) Moral Responsibility of AI

To assess perceptions of the AI agent's moral responsibility in the context of service failure, participants were presented with a scenario adapted from a real case involving a tax AI system used by the Dutch government. In this case, an AI system designed to detect fraudulent child benefit claims mistakenly flagged approximately 26,000 individuals, resulting in severe personal consequences, including financial distress, family breakdown, and, in some cases, suicide. Based on this real-world incident, the following vignette was presented to participants:

"*Based on the decision made by the tax AI TAX-76, the government issued a notice to single mother Janet Lamesa stating, 'You have fraudu-*

lently received child benefits and must repay €40,000 (approximately 60 million KRW).' Janet later contacted the tax office to explain her situation, but her request was denied. As a result, she took on significant debt to repay the alleged overpayment, lost her job due to financial instability, and ultimately lost custody of her child to her ex-husband. However, an audit later revealed that TAX-76's decision was incorrect."

After reading the scenario, participants rated the extent to which they believed the AI was responsible for the outcome using a single item: "*To what extent do you think the AI is responsible for the outcome in this situation?*" Although the item measured general responsibility, we interpreted this attribution—based on previous studies (e.g., Alicke, 2000; Kahn et al., 2012; Shank & DeSanti, 2018)—within the framework of dyadic morality as reflecting judgments of moral responsibility. Responses were recorded on a 7-point Likert scale (1 = not at all responsible and 7 = completely responsible).

To provide a more balanced evaluation of attribution, the participants also rated the perceived responsibility of the government and human users in the same scenario using comparable items.

3. Statistical Analyses

A confirmatory factor analysis was performed to evaluate the adequacy of the measurement model. Factor loadings and model fit indices were examined to ensure the construct validity and measurement reliability.

Subsequently, the hypothesized structural model was compared with an alternative model to evaluate the overall model fit and significance of the path coefficients. the model selection was based on comparative fit indices and explanatory power of the respective path structures.

To test the moderating effect of emotional unpleasantness on perceived moral patiency, we employed Ping's (1996) two-step approach. In this procedure, an interaction term was incorporated into the measurement model in the first step to ensure model identification. Second, the estimated parameter values were fixed in the structural model, which were then re-estimated to assess the moderation effects within a covariance structure framework.

Ⅵ. Results

1. Preliminary Analysis

Table 1 presents the descriptive statistics and intercorrelations among the study variables. As expected, the perceived HL of the AI was positively correlated with perceived moral patiency (r = .41, p < .01), perceived moral agency (r = .23, p < .01), and emotional unpleasantness (r = .27, p < .01). Perceived moral patiency was positively associated with moral agency (r = .51, p < .01) and emotional unpleasantness (r = .39, p < .01). Additionally, moral agency was modestly but significantly correlated with emotional unpleasantness (r = .15, p < .05).

All skewness and kurtosis values were within the thresholds recommended by Kline (2005)—less than 3 for skewness and less than 10 for kurtosis—suggesting no significant violations of the nor mality assumption and supporting the suitability

<Table 1> Descriptive statistics and correlations among study variables (N = 270)

	1	2	3	4	5
1. Human-likeness of AI	–				
2. Perceived moral patiency	.41**	–			
3. Perceived moral agency	.23**	.51**	–		
4 Moral responsibility of AI	−.12	−.00	.11	–	
5. Emotional unpleasantness	.27**	.39**	.15**	−.01	–
Mean	2.97	2.34	3.98	5.23	3.35
Standard Deviation	1.40	2.34	1.22	1.73	1.43
Skewness	−.07	.92	−.52	−1.01	.19
Kurtosis	−1.05	−.02	.08	.28	−.51

* *p* < .05, ** *p* < .01

of the data for structural equation modeling.

2. Measurement Model

The measurement model was tested using maximum likelihood estimation and included four latent variables: the perceived HL of AI, moral patiency, moral agency, and moral responsibility. Model fit indices indicated an overall fit: $\chi^2(39) = 175.368$, $p < .001$, $\chi^2/df = 4.497$, CFI = .930, TLI = .901, RMSEA = .114 (90% CI [.097, .131]), SRMR = .075, AIC = 229.368, and BIC = 326.526. While the RMSEA slightly exceeded the recommended threshold of .08 (Hu & Bentler, 1999), other indices, including CFI and TLI values above .90 and an SRMR below .08, suggest that the model demonstrates acceptable fit when evaluated using multiple criteria (Kline, 2015). These results suggest that the model provides a reasonable approximation of the observed data.

3. Comparison of Structural Models for Mediation

The model fit indices for the hypothesized (sequential mediation) and alternative (parallel mediation) models are summarized in Table 2. The hypothesized model demonstrated a better fit, as indicated by a lower χ^2/df ratio (4.385 vs.

<Table 2> Comparison of fit indices between the research model and the alternative model

Model	χ²	df	χ²/df	CFI	TLI	RMSEA [90% CI]	SRMR	AIC	BIC
Research model	179.799***	41	4.385	.929	.904	.112 [.096, .129]	.076	229.799	319.759
Alternative model	211.375***	41	5.155	.912	.883	.124 [.108, .141]	.087	261.375	351.335

*** *p* < .001

<Table 3> Indirect effects in the research model

Pathway	B	β	S.E.	p
Human−likeness → Patiency → Agency → Responsibility	.14	.06* [.01, .11]	.08	.024

Note. Human−likeness=human−likeness of AI, Patiency=perceived moral patiency, Agency=perceived moral agency, Responsibility=moral responsibility of AI; values in brackets represent 95% confidence intervals(CI). * p < .05

<Table 4> Indirect effects in the alternative model

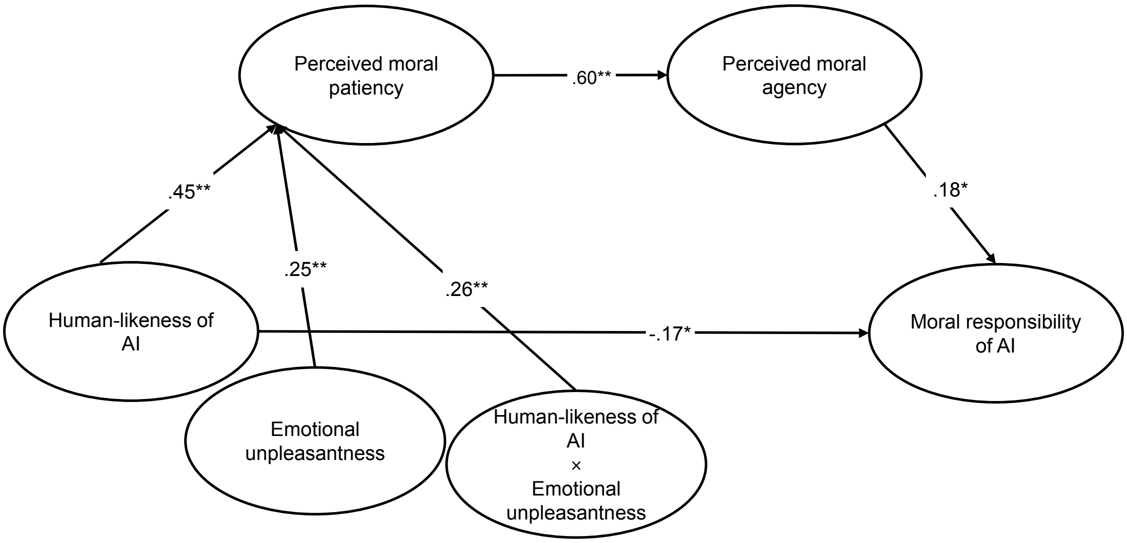
Pathway	B	β	S.E.	p
Human-likeness → Patiency → Responsibility	-.28 [-.84, .09]	.12	.25	.123
Human-likeness → Agency → Responsibility	.41 [.04, 1.13]	-.08*	.31	.024

Note. Values in brackets represent 95% confidence intervals (CI). * p < .05

5.155), higher CFI (.929 vs. .912), and higher TLI (.904 vs. .893) compared to the alternative model. Although both models exceeded the recommended threshold for RMSEA (.112 and .124), the sequential model yielded comparatively better fit indices. Additionally, the AIC and BIC values were lower for the sequential model (AIC = 229.799 and BIC = 319.759) than for the parallel model (AIC = 261.375 and BIC = 351.335), suggesting greater parsimony.

In the sequential mediation model, perceived HL of AI significantly predicted perceived moral patiency (β = .52, p < .01), which in turn predicted perceived moral agency (β = .59, p < .01). Perceived moral agency significantly predicted moral responsibility (β = .18, p < .05). Notably, HL of AI showed a significant direct negative effect on moral responsibility (β = -.17, p < .05). Furthermore, contrary to expectations, the total effect of HL on moral responsibility was negative and not statistically significant (β = -.12, n.s.). Indirect effect analysis using bootstrapping (Table 3) confirmed a significant sequential mediation effect through moral patiency and agency (β = .055, p < .05). In contrast, the parallel mediation model showed significant direct effects from HL to both moral agency (β = .69, p < .01) and moral patiency (β = .79, p < .01), and a significant effect from moral agency to moral responsibility (β = .17, p < .05). However, the path from moral patiency to moral responsibility was not significant (β = -.10, p > .05), and the indirect effect through moral patiency was also nonsignificant (Table 4).

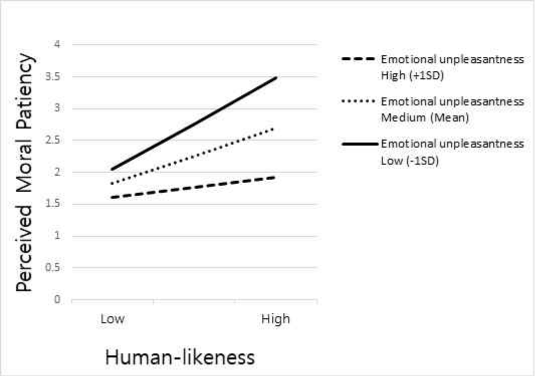
These findings support Hypothesis 2, indicating that perceived moral patiency and moral agency function as sequential mediators rather than as parallel processes in the relationship between AI HL and perceptions of moral responsibility. However, Hypothesis 1 was not supported. Contrary to expectations, HL of AI was negatively associated with perceived moral responsibility. This direct negative effect coexisted with a positive and sig nificant indirect effect via the sequential paths of perceived patiency and agency. This pattern constitutes inconsistent mediation (MacKinnon et al., 2000), in which the direct and indirect effects have opposite signs.



[Fig. 1] Results of structural equation model results with moderation effects

Note. All coefficients are standardized path coefficients. * p < .05, ** p < .01.

4. Moderating Effect of Emotional Unpleasantness



[Fig. 2] Moderating effect of emotional unpleasantness on the relationship between human-likeness of AI and perceived moral patiency

To examine the moderating role of emotional unpleasantness, Ping's (1996) two-step approach to latent interaction modeling was employed. In Step 1, a measurement model was estimated including the predictor variable (perceived HL of the AI) and the moderator (emotional unpleasantness). To reduce potential multicollinearity, the observed indicators for both latent constructs were mean-centered prior to interaction term construction. In Step 2, the interaction term (HL ×emotional unpleasantness) was considered a single-indicator latent variable. Its variance (1.50), error variance (fixed at 103.29), and factor loadings (10.63) were derived from the estimates obtained in Step 1, following Ping's recommended procedure. These values were fixed in the structural model to test for moderation effects. The overall model fit was acceptable: p < .001; χ^2/df = 3.025; CFI = .929; TLI = .915; RMSEA =.087 (90% CI [.076, .098]); SRMR = .065; AIC = 374.507; BIC = 504.050. As shown in Figure 1, the interaction term had a significant positive effect on perceived moral patiency (β = .26, p < .01), indicating a meaningful moderation effect.

These results support Hypothesis 3, which proposed that emotional unpleasantness moderates the relationship between HL and dyadic morality. Specifically, the positive association between AI HL and perceived moral patiency was stronger when emotional unpleasantness was high. In contrast, under low levels of emotional unpleasantness, the effect of HL on perceived patiency was relatively weak (Figure 2).

Ⅶ. Discussion

This study explored the relationship between AI HL and perceived moral responsibility by examining the mediating role of dyadic morality and the moderating role of emotional unpleasantness.

A notable finding was the emergence of a suppression effect within the sequential mediation model. Although perceived HL did not significantly predict moral responsibility in a positive direction as hypothesized (H1 was not supported), it exerted a significant negative direct effect. By contrast, the indirect path through moral patiency and moral agency was significantly positive, supporting Hypothesis 2. The total effect of HL on moral responsibility was negative but not statistically significant, suggesting that the direct and indirect effects canceled each other. The presence of a negative direct effect along with a significant positive indirect effect represents a classic suppression effect (MacKinnon et al., 2000). In this case, the inclusion of mediating variables revealed a positive pathway of moral attribution that was previously obscured by the direct effect, indicating that a true underlying association emerges only when intermediary psychological processes are considered.

Furthermore, the negative direct effect remained robust even after accounting for emotional unpleasantness as a moderator, suggesting that the inverse association between HL and perceived moral responsibility may reflect a distinct and consistent psychological response that operates independently of affective aversion. This may be explained by Złotowski et al.'s (2014) dimensional approach to anthropomorphism, which builds on Haslam's (2006) dehumanization model. According to this view, human attributes can be divided into "Human Nature" (emotion-based traits) and "Uniquely Human" characteristics (moral reasoning and intelligence). AI with a low physical embodiment may invite attributions of intelligence and moral reasoning, whereas more human-like AI may evoke emotional responses that obscure higher-order evaluations. In our study, HL was significantly associated with emotional unpleasantness (r = .27), which may have contributed to reduced moral responsibility attributions. However, the absence of a significant correlation between emotional unpleasantness and moral responsibility (r = -.01) suggests that other affective or cognitive mechanisms—such as cuteness, norm violation, or expectation conflict may be involved. Future research should explore additional mediating processes beyond affective reactions such as eeriness.

Another major finding was that emotional unpleasantness moderated the first stage of mediation. When emotional unpleasantness was low, HL had minimal influence on perceived moral patiency. By contrast, when unpleasantness was high, HL substantially increased perceptions of moral patiency, which in turn activated the downstream path from agency to moral responsibility. In other words, only when human-likeness elicits sufficient emotional discomfort does the empathic-cognitive sequence ("This AI can feel pain → This AI can act intentionally → This AI is responsible") become salient.

This finding aligns with prior research on the uncanny valley (Laakasuo et al., 2021; Shin et al., 2019), suggesting that the discomfort triggered by human-like robots can significantly influence perception and moral evaluation. A plausible explanation for this amplification effect is that participants may have interpreted their discomfort as a signal of empathic relevance, thereby attributing greater emotional capacity to the AI. Prior studies support this possibility: Wang and Krumhuber (2018) found that the social function of robots increased ascriptions of emotion, highlighting the role of affective functions in mind perception, and Rosenthal-von der Pütten et al. (2013) showed that observers' emotional reactions toward robots enhanced empathic attributions. Overall, these findings suggest that affective responses, even those with negative valence, can serve as heuristic cues for inferring an entity's capacity to feel.

Nonetheless, the reason why high unpleasantness amplifies the effect of HL on patiency remains unclear. Given that patiency entails the capacity to feel pain or joy, one possibility is that discomfort is projected onto the AI, enhancing the perception that it could experience emotions. Further empirical work is needed to test this possibility and clarify whether the observed amplification is unique to unpleasant affect or whether it reflects the broader role of affective responses in shaping patiency judgments.

The results of this study have several theoretical implications. First, they show that individuals can attribute moral responsibility to AI, suggesting that questions on AI rights and accountability are not merely philosophical or legal, they are also psychological. Second, the negative direct effect of HL on moral responsibility contradicts prior studies and supports Złotowski et al.'s (2014) argument that anthropomorphism is best conceptualized as a multidimensional construct. Future research should distinguish between its subcomponents, particularly when investigating moral judgments. Third, our findings support the broader framework of dyadic morality (Gray et al., 2007; Stuart & Kneer, 2021) by demonstrating that both perceived patiency and agency contribute to moral judgment. Although this theory does not specify a directional sequence, this study is the first to empirically propose and validate a sequential pathway in which patiency precedes agency in the formation of moral responsibility attributions. However, further research in various contexts is necessary to generalize this mechanism. Finally, in contrast to previous studies that treated emotional unpleasantness as an outcome of perceived experience (Gray & Wegner, 2012b), this study offers a novel contribution by positioning it as an independent moderator that influences how HL interacts with moral cognition.

1. Limitations and Future Research

This study has several limitations. First, this study presented only AI facial images, following the methods used in previous studies. However, even small changes in AI's appearance can lead to different perceptions of its mental attributes and abilities (Powers & Kiesler, 2006). Therefore, future research should explore various approaches, such as using full-body images, to further re-examine the findings. Second, this study measured perceived moral responsibility toward AI using a single-item scale. Studies suggesting that single-item measures may be equally or even more valid than multiple-item scales for heuristic judgments such as responsibility perception (Fisher et al., 2016) support the appropriateness of this measurement approach for interpreting the current findings. Future research should develop and apply multiitem scales to reassess responsibility judgments, particularly instruments that directly capture moral responsibility judgments to provide a more precise understanding of the construct. Third, the scenario used in this study involved a government AI system that caused significant losses. In contrast, a study using low-stakes scenarios such as food-serving errors (Yam et al., 2021) found that a greater perceived experience with robots was associated with greater forgiveness. Additionally, in a study using digital therapeutics scenarios (Choi, Sung, & Chung, unpublished data), perceived moral agency mediated the effect of HL of AI on the intention to use, rather than perceived moral patiency. These findings suggest that additional research is required to generalize the results of this study, emphasizing the need to vary the degrees of loss, application domains, and types of victims in future research. Fourth, this study measured only the dimension of unpleasantness as an emotional response to AI, based on prior research on the uncanny valley. However, low unpleasantness does not necessarily indicate a pleasant state, highlighting the need to measure positive and negative emotions separately. Studies have also examined positive emotions toward AI (Smith et al., 2020; Otaka et al., 2024), and future research should assess both negative and positive emotions to clarify their specific role of positive emotions. Finally, future studies should consider the temporal dynamics of emotional unpleasantness. During the initial interactions, emotional unpleasantness may amplify the relationship between the AI's HL and perceived moral patiency. However, with repeated interactions and learning, emotional unpleasantness may decrease or increase, potentially altering the relationships. Therefore, future studies should systematically explore how emotional unpleasantness toward AI evolves over time and how evaluations of AI failures change with repeated exposure.

Ⅷ. Conclusions and Practical Implications

This study contributes to research on AI and robot anthropomorphism by elucidating the mechanisms through which HL of AI influences the perception of its moral responsibility. Additionally, it expands existing uncanny valley research by demonstrating that emotional unpleasantness toward an AI plays a critical role in evaluating the AI's capacity for experience.

The practical implications of this study are as follows: First, in today's society where AI services are rapidly expanding across various sectors, AI development companies and governments must seriously consider how to address the legal and moral responsibilities arising from AI service failures. This study, along with previous research, consistently demonstrates that people tend to attribute responsibility to AI, highlighting the need for governments to consider developing future policies regarding accountability for AI service failures. Second, AI and robot developers should recognize that an AI's appearance influences perceptions of experience and agency, which in turn shape consumer responses. Although the research in this area is insufficient, ongoing studies should accumulate detailed data on how the appearance of AI affects consumer reactions to inform AI development. Finally, this study suggests that the degree of HL of AI and robots, together with emotional unpleasantness, plays an important role in how people make judgments about AI. Therefore, minimizing unpleasantness may be more important than simply increasing HL in AI and robot designs. Thus, developers should consider this when designing AI and robotic systems.

References

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological bulletin, 126(*4), 556. https://doi.org/10.1037/0033-2909.126.4.556

Bartneck, C., Kanda, T., Ishiguro, H., & Hagita, N. (2009, September). My robotic doppelg nger-A critical look at the uncanny valley. In *RO-MAN 2009-The 18th IEEE international symposium on robot and human interactive communication* (pp. 269-276). IEEE. https://doi.org/10.1109/ROMAN.2009.5326351

Berscheid, E., & Hatfield, E. (1969). *Interpersonal attraction* (Vol. 69, pp. 113-114). Reading, MA: Addison-Wesley.

Broadbent, E., Kumar, V., Li, X., Sollers 3rd, J., Stafford, R. Q., MacDonald, B. A., & Wegner, D. M. (2013). Robots with display screens: a robot with a more humanlike face display is perceived to have more mind and a better personality. *PloS one, 8*(8), e72589. https://doi.org/10.1371/journal.pone.0072589

Byrne, D. (1971). T*he Attraction Paradigm Academic Press. New York, NY, USA.*

Cheetham, M., Suter, P., & Jäncke, L. (2011). The human likeness dimension of the "uncanny valley hypothesis": behavioral and functional MRI findings. *Frontiers in human neuroscience, 5,* 126. https://doi.org/10.3389/fnhum.2011.00126

Cheetham, M., Suter, P., & Jancke, L. (2014). Perceptual discrimination difficulty and familiarity in the uncanny valley: more like a "Happy Valley". *Frontiers in psychology, 5,* 1219. https:/doi.org/10.3389/fpsyg.2014.01219

Choi, Y.-B., & Jang, D. (2022). Is Mr. AI more responsible? The effect of anthropomorphism in the moral judgement toward AI's decision making. *Korean Journal of Cognitive Science, 33(*4), 169-203. https://doi.org/10.19066/cogsci.2022.33.4.001

Diel, A., Weigelt, S., & Macdorman, K. F. (2021). A meta-analysis of the uncanny valley's independent and dependent variables. *ACM Transactions on Human-Robot Interaction (THRI), 11*(1), 1-33.

https://doi.org/10.1145/3470742

Dodd, J. (2024, November 19). Expert warns of AI chatbot risks after teen user's suicide. *PEOPLE.* https://people.com/expert-warns-of-ai-chatbot-risks-after-recent-suicide-of-teen-user-8745883

Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and autonomous systems, 42*(3-4), 177-190. https://doi.org/10.1016/S0921-8890(02)00374-3

El Atillah, I. (2023, March 31). Man ends his life after an AI chatbot encouraged him to sacrifice himself to stop climate change. *Euronews.* https://www.euronews.com/next/2023/03/31/man-ends-his-life-after-an-ai-chatbot-encouraged-him-to-sacrifice-himself-to-stop-climate-

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: a three-factor theory of anthropomorphism. *Psychological review, 114*(4), 864-886. https://doi.org/10.1037/0033-295X.114.4.864

Eyssel, F., & Kuchenbrandt, D. (2012). Social categorization of social robots: Anthropomorphism as a function of robot group membership. *British Journal of Social Psychology, 51*(4), 724-731. https://doi.org/10.1111/j.2044-8309.2011.02082.x

Ferrey, A. E., Burleigh, T. J., & Fenske, M. J. (2015). Stimulus-category competition, inhibition, and affective devaluation: a novel account of the uncanny valley. *Frontiers in psychology, 6,* 249. https://doi.org/10.3389/fpsyg.2015.00249

Fink, J. (2012). Anthropomorphism and human likeness in the design of robots and human-robot interaction. In I*nternational conference on social robotics* (pp. 199-208). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-34103-8_20

Fisher, G. G., Matthews, R. A., & Gibbons, A. M. (2016). Developing and investigating the use of single-item measures in organizational research. *Journal of occupational health psychology, 21*(1), 3. https://doi.org/10.1037/a0039139

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and machines, 14,* 349-379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and autonomous systems, 42*(3-4), 143-166. https://doi.org/10.1016/S0921-8890(02)00372-X

Gerrans, P., & Kennett, J. (2010). Neurosentimentalism and moral agency. *Mind, 119*(475), 585-614. https://doi.org/10.1093/mind/fzq037

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *science, 315*(5812), 619-619. https://doi.org/10.1126/science.1134475

Gray, K., & Wegner, D. M. (2009). Moral typecasting: divergent perceptions of moral agents and moral patients. *Journal of personality and social psychology, 96*(3), 505-520. https://doi.org/10.1037/a0013748

Gray, K., & Wegner, D. M. (2012a). Morality takes two: Dyadic morality and mind perception. In M. Mikulincer & P. R. Shaver (Eds.), *The social psychology of morality: Exploring the causes of good and evil* (pp. 109-

127). American Psychological Association. https://doi.org/10.1037/13091-006

Gray, K., & Wegner, D. M. (2012b). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition, 125*(1), 125-130. https://doi.org/10.1016/j.cognition.2012.06.007

Greene, J. D. (2007). Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Trends in cognitive sciences, 11*(8), 322-323. https://doi.org/10.1016/j.tics.2007.06.004

Gunkel, D. J. (2012). *The machine question: Critical perspectives on AI, robots, and ethics.* mit Press.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review, 108*(4), 814.

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and social psychology review, 10*(3), 252-264. https://doi.org/10.1207/s15327957pspr1003_4

Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent?. *Ethics and Information Technology, 11,* 19-29. https://doi.org/10.1007/s10676-008-9167-5

Kahn Jr, P. H., Ishiguro, H., Friedman, B., Kanda, T., Freier, N. G., Severson, R. L., & Miller, J. (2007). What is a human?: Toward psychological benchmarks in the field of human-robot interaction. *Interaction Studies, 8*(3), 363-390. https://doi.org/10.1075/is.8.3.04kah

Kahn Jr, P. H., Kanda, T., Ishiguro, H., Gill, B. T., Ruckert, J. H., Shen, S., ... & Severson, R. L. (2012, March). Do people hold a humanoid robot morally accountable for the harm it causes?. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction* (pp. 33-40). https://doi.org/10.1145/2157689.2157696

Kätsyri, J., Förger, K., Mäkäräinen, M., & Takala, T. (2015). A review of empirical evidence on different uncanny valley hypotheses: support for perceptual mismatch as one road to the valley of eeriness. *Frontiers in psychology, 6,* 390. https://doi.org/10.3389/fpsyg.2015.00390

Kim, B., de Visser, E., & Phillips, E. (2022). Two uncanny valleys: Re-evaluating the uncanny valley across the full spectrum of real-world human-like robots. *Computers in human behavior, 135,* 107340. https://doi.org/10.1016/j.chb.2022.107340

Kline, T. J. (2005). *Psychological testing: A practical approach to design and evaluation.* Sage publications.

Kühne, R., & Peter, J. (2023). Anthropomorphism in human-robot interactions: a multidimensional conceptualization. *Communication Theory, 33*(1), 42-52. https://doi.org/10.1093/ct/qtac020

Laakasuo, M. (2023). Moral Uncanny Valley revisited-how human expectations of robot morality based on robot appearance moderate the perceived morality of robot decisions in high conflict moral dilemmas. *Frontiers in Psychology, 14,* 1270371. https://doi.org/10.3389/fpsyg.2023.1270371

Laakasuo, M., Palomäki, J., & Köbis, N. (2021).

Moral uncanny valley: A robot's appearance moderates how its decisions are judged. *International Journal of Social Robotics, 13*(7), 1679–1688. https://doi.org/10.1007/s12369-020-00738-6

Ladak, A., Harris, J., & Anthis, J. R. (2024, May). Which artificial intelligences do people care about most? A conjoint experiment on moral consideration. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (pp. 1–11). https://doi.org/10.48550/arXiv.2403.09405

Linnunsalo, S., Küster, D., Yrttiaho, S., Peltola, M. J., & Hietanen, J. K. (2023). Psychophysiological responses to eye contact with a humanoid robot: Impact of perceived intentionality. *Neuropsychologia, 189,* 108668. https://doi.org/10.1016/j.neuropsychologia.2023.108668

MacDorman, K. F., & Entezari, S. O. (2015). Individual differences predict sensitivity to the uncanny valley. *Interaction Studies, 16*(2), 141–172. https://doi.org/10.1075/is.16.2.01mac

MacKinnon, D. P., Krull, J. L., & Lockwood, C. M. (2000). Equivalence of the mediation, confounding and suppression effect. *Prevention science, 1*(4), 173–181.

Malle, B. F., Scheutz, M., Forlizzi, J., & Voiklis, J. (2016, March). Which robot am I thinking about? The impact of action and appearance on people's evaluations of a moral robot. In *2016 11th ACM/IEEE international conference on human–robot interaction (HRI)* (pp. 125–132). IEEE. https://doi.org/10.1109/HRI.2016.7451743

Mathur, M. B., & Reichling, D. B. (2016). Navigating a social world with robot partners: A quantitative cartography of the Uncanny Valley. *Cognition, 146,* 22–32. https://doi.org/10.1016/j.cognition.2015.09.008

Moll, J., de Oliveira–Souza, R., & Eslinger, P. J. (2003). Morals and the human brain: a working model. *Neuroreport, 14*(3), 299–305.

Monroe, A. E., Dillon, K. D., & Malle, B. F. (2014). Bringing free will down to Earth: People's psychological concept of free will and its role in moral judgment. *Consciousness and cognition, 27,* 100–108. https://doi.org/10.1016/j.concog.2014.04.011

Mori, M. (1970). The uncanny valley. Energy, *7*(4), 33–35.

Mosakas, K. (2021). On the moral status of social robots: considering the consciousness criterion. *AI & Society, 36*(2), 429–443. https://doi.org/10.1007/s00146-020-01002-1

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., Steuer, J., & Tauber, E. R. (1994, April). Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 72–78).

Noh., Hyungrae (2022). Literalism as Non–anthropocentric Semantics. Journal of the *Future of Society, 13*(1), 1–19. https://doi.org/10.22987/jifso.2022.13.1.1

Nucci, L. P., & Gingo, M. (2010). The development of moral reasoning. *The Wiley-Blackwell handbook of childhood cognitive development,* 420–445. https://doi.org/10.1002/9781444325485

Otaka, E., Osawa, A., Kato, K., Obayashi, Y., Uehara, S., Kamiya, M., ... & Kondo, I. (2024). Positive emotional responses to socially assis-tive robots in people with dementia: pilot study. *JMIR aging, 7*(1), e52443. https://doi.org/10.2196/52443

Ping Jr, R. A. (1996). Latent variable interaction and quadratic effect estimation: A two-step technique using structural equation analysis. *Psychological Bulletin, 119*(1), 166. https://doi.org/10.1037/0033-2909.119.1.166

Powers, A., & Kiesler, S. (2006, March). The advisor robot: tracing people's mental model from a robot's physical attributes. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction* (pp. 218-225). https://doi.org/10.1145/1121241.1121280

Rosenthal-von der Pütten, A. M., Krämer, N. C., Hoffmann, L., Sobieraj, S., & Eimler, S. C. (2013). An experimental study on emotional reactions towards a robot. *International Journal of Social Robotics, 5*(1), 17-34. https://doi.org/10.1007/s12369-012-0173-8

Shank, D. B., & 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

Shin, M., Kim, S. J., & Biocca, F. (2019). The uncanny valley: No need for any further judg-ments when an avatar looks eerie. *Computers in Human Behavior, 94,* 100-109. https://doi.org/10.1016/j.chb.2019.01.016

Smith, E. R., Sherrin, S., Fraune, M. R., & Šabanović, S. (2020). Positive emotions, more than anxiety or other negative emotions, pre-dict willingness to interact with robots. *Personality and Social Psychology Bulletin, 46*(8), 1270-1283. https://doi.org/10.1177/0146167219900439

Song, S. W., & Shin, M. (2024). Uncanny valley effects on chatbot trust, purchase intention, and adoption intention in the context of e-commerce: The moderating role of avatar familiarity. *International Journal of Human-Computer Interaction, 40*(2), 441-456. https://doi.org/10.1080/10447318.2022.2121038

Stuart, M. T., & Kneer, M. (2021). Guilty artifi-cial minds. *arXiv preprint arXiv:2102.04209.* https://doi.org/10.48550/arXiv.2102.04209

Torrance, S. (2008). Ethics and consciousness in artificial agents. *Ai & Society, 22,* 495-521. https://doi.org/10.1007/s00146-007-0091-8

Tsukimoto, T. (2017). A Kansei-evaluation Study on the Uncanny Valley of Real-world Robots-From the Perspective of Mind Perception. *Transactions of Japan Society of Kansei Engineering, 16*(3), 293-298. https://doi.org/10.5057/jjske.TJSKE-D-16-00091

Waytz, A., Cacioppo, J., & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science, 5*(3), 219-232. https://doi.org/10.1177/1745691610369336

Yam, K. C., Bigman, Y. E., Tang, P. M., Ilies, R., De Cremer, D., Soh, H., & Gray, K. (2021). Robots at work: People prefer—and forgive—service robots with perceived feelings. *Journal of Applied Psychology, 106*(10), 1557. https://doi.org/10.1037/apl0000834

Yam, K. C., Goh, E. Y., Fehr, R., Lee, R., Soh, H., & Gray, K. (2022). When your boss is a robot: Workers are more spiteful to robot supervisors that seem more human. *Journal of Experimental Social* Psychology, 102, 104360. https://doi.org/10.1016/j.jesp.2022.104360

You, S., & Robert Jr, L. P. (2018, February). Human-robot similarity and willingness to work with a robotic co-worker. In *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction* (pp. 251-260). https://doi.org/10.1145/3171221.3171281

Wang, X., & Krumhuber, E. G. (2018). Mind perception of robots varies with their economic versus social function. *Frontiers in psychology, 9,* 1230. https://doi.org/10.3389/fpsyg.2018.01230

Złotowski, J., Strasser, E., & Bartneck, C. (2014, March). Dimensions of anthropomorphism: from humanness to humanlikeness. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction* (pp. 66-73). https://doi.org/10.1145/2559636.2559679

투고일자: 2025. 8. 29.
심사일자: 2025. 10. 10.
게재확정일자: 2025. 10. 14.

미래사회
Journal of Future Society
2025. Vol. 16, No. 3, 218～239.
https://doi.org/10.22987/jifso.2025.16.3.218

AI의 인간유사성이 도덕적 책임감에 미치는 영향:
도덕 쌍과 정서적 불쾌감의 역할

윤은진 정은경

강원대학교

본 연구에서는 인공지능(AI)의 인간 유사성이 AI 도덕적 책임 판단에 미치는 영향을 검증하고자 하였다. 이를 위해 첫째, AI의 인간 유사성과 AI 도덕적 책임 판단의 관계를 살펴보았고, 둘째, AI의 인간 유사성과 도덕적 책임 판단의 관계에서 지각된 도덕적 경험성과 도덕적 행위성의 순차적 매개효과를 확인하였다. 그리고 정서적 불쾌감이 AI의 인간 유사성과 지각된 도덕적 경험성의 관계를 조절하는지를 확인하였다. 이를 위해 온라인 설문조사를 통해 국내 성인 270명의 자료를 수집하였다. 연구 결과, AI의 인간 유사성은 AI의 도덕적 책임성 판단과 부적 상관을 나타냈으며 이는 AI가 인간과 닮을수록 AI의 도덕적 책임이 더 높다고 판단할 것이라는 예측과는 반대되는 결과였다. 그러나 지각된 도덕적 경험성과 도덕적 행위성을 거치는 순차적 매개 경로가 확인되어, AI의 인간 유사성이 AI의 도덕적 책임 판단에 간접적으로는 정적인 영향을 미침을 보여주었다. 또한 정서적 불쾌감은 AI의 인간 유사성과 지각된 도덕적 경험성 간의 관계를 조절하는 것으로 나타났으며, 정서적 불쾌감이 높은 경우 AI의 인간 유사성이 지각된 도덕적 경험성에 미치는 영향이 강화되었다. 이러한 연구결과는 인간-AI 상호작용에서 AI의 인간 유사성, AI 도덕적 책임 판단, 지각된 도덕적 경험성 및 행위성, 불쾌한 골짜기(Uncanny Valley) 효과 간의 관계를 이해하는 데 기여한다.

주요어 : AI, 인간 유사성, 도덕 쌍, 도덕적 책임, 정서적 불쾌감