AI Moral Patiency and Moral Dissonance:
Almost Human, Almost Deserving Moral Treatment

Alan Dennis
Indiana University
ardennis@iu.edu

Mike Seymour
University of Sydney
Mike.seymour@sydney.edu.au

Antino Kim
Indiana University
antino@iu.edu

Lingyao Yuan
Iowa State University
lyuan@iastate.edu

Abstract

As Artificial Intelligence (AI) agents become increasingly common in all walks of life, most users would agree that AI agents should behave ethically and morally towards their human users. This paper examines moral patiency (MP), the extent to which an entity is perceived as deserving moral consideration. This is a construct distinct from moral agency (the ability of an entity to act morally). We develop and validate a multi-dimensional scale capturing six positive and six negative factors indicating the extent to which someone ascribes MP to an AI agent. MP toward an AI agent was only weakly correlated with MP toward human agents. Interestingly, the MP factors that were related to trust in human agents were quite different than the MP factors that were related to trust in the AI agents. Some users reported treating the AI fairly, following its advice, and protecting its security. Fewer participants reported engaging in negative MP behaviors. These findings highlight the risk of moral dissonance, the ethical confusion users experience about wanting the AI to treat them morally, but failing to perceive a need to reciprocate and treat the AI morally. We argue that MP, and the moral dissonance it may generate, is a foundational yet underexplored lens for understanding the evolving dynamics of human-AI interaction.

Keywords: AI, moral patiency, moral dissonance

1. Introduction

The deployment of information systems (IS) using Artificial Intelligence (AI) has surged in recent years. As AI becomes central to more systems and affects more business processes, there are many articles and discussions about the importance of ensuring AI treats its human users and the humans affected by its actions in a moral and ethical way. A general global consensus has emerged around five moral and ethical principles for how AI should act although there is disagreement over their relative importance and how they should be implemented (Jobin et al., 2019). Moral agency is the extent to which an entity has the independent capacity to act in ways that are moral (Sytsma & Machery, 2012). There are many examples of AI agents failing to consider moral issues, leading to improper actions. Most users would agree that AI agents should exhibit moral agency in its behavior (e.g., by acting morally), but whether AI can truly be a moral actor is under debate (Cervantes et al., 2020; Gudmunsen, 2025).

However, should human users be obligated to reciprocate and treat AI morally? Moral patiency (MP) refers to the ascription of moral standing to another entity—the belief that an entity deserves to be treated in respectful ways (Banks, 2025; Sytsma & Machery, 2012). Users increasingly expect AI agents to act ethically and demonstrate moral agency. However, it remains unclear whether users feel compelled to reciprocate by treating these agents as entities deserving of moral consideration. This asymmetry has the potential to produce moral dissonance: a psychological and ethical tension in which users expect the AI to behave morally yet feel no obligation to act morally toward the AI in return. Such dissonance may become more pronounced as AI agents adopt increasingly human-like roles and develop realistic appearances, challenging users' intuitions about fairness, empathy, and accountability in mediated human-computer interaction.

The IS discipline has traditionally concerned itself with the design, implementation, and implications of socio-technical systems. The emergence of AI agents that *simulate* humanity introduces new ethical, behavioral, and design challenges that lie at the heart of the discipline. Trust and trustworthiness have long been important topics for IS discourse. Trust is typically linked to an entity's perceived competence, benevolence, and integrity (i.e., its moral *agency*). MP, on the other hand, governs how users feel obligated to act toward an entity. Understanding how and if users ascribe MP to AI entities is essential for evaluating the ethical dynamics of human-AI interaction.

From the system design perspective, the issue of moral patiency (MP) raises important questions about how AI systems should be built. Developers should recognize that AI agents, by mimicking human traits,

URI: https://hdl.handle.net/10125/111462
978-0-9981331-9-5

can evoke strong emotional responses such as empathy. Thus, they must carefully consider whether to encourage or limit emotional attachment, depending on the context in which they are deployed. This is particularly critical when human-like features are used to increase user engagement or generate revenue. From a social and ethical standpoint, MP ties into broader concerns in information systems about the unintended consequences of technology disruption. AI that alters how people emotionally relate to others, or that desensitizes users, may reinforce existing biases and reduce individuals' sense of social responsibility. Inspired by those considerations, we have three research questions:

RQ1: Do users believe they should treat AI agents morally (i.e., that AI agents have MP)?

RQ2: What factors influence the ascription of MP to AI agents?

RQ2: To what extent does MP influence trust in AI agents?

We conducted a survey of American students to develop and validate a multi-dimensional measure of MP capturing six positive and six negative moral behaviors. Few users reported negative behaviors, but the positive behaviors show greater differences. MP toward AI was only weakly correlated with MP toward humans. The MP factors related to trust in humans were quite different than the factors for AI agents.

2. Background

2.1 Moral Behaviour and AI

In human relationships, moral obligations are often reciprocal: people are generally expected to treat others well and, in turn, be treated well themselves. This reciprocity can be understood through the concepts of moral agency and moral patiency; see Figure 1. Moral agency refers to how an individual perceives another's capacity to act morally, shaping expectations about how that person should and does treat us (Black, 2016; Rottschaefer, 1991). In contrast, moral patiency concerns how we believe another should be treated, influencing how we act toward them (Banks, 2025; Banks & Bowman, 2022; Sytsma & Machery, 2012).

While such reciprocal moral dynamics are common and expected in most healthy human relationships, they do not extend to all entities. For example, we may feel a moral obligation to treat pets and other animals ethically, yet we do not expect them to reciprocate fully in the same way. Similarly, we do not typically consider tools or objects as part of any moral relationship. There are no moral issues associated with how I treat my hammer, as I do not think it warrants a moral treatment.

However, with the rise of advanced AI agents and companions, people are increasingly forming meaningful connections with non-human entities. This shift raises important questions about how moral agency and patiency apply in human–AI relationships. Since MP is a fundamental component of human social interaction, and as AI systems aim to mimic these interactions, it becomes crucial to examine the factors that shape MP and, in turn, how it influences trust. This is the central focus of our investigation.

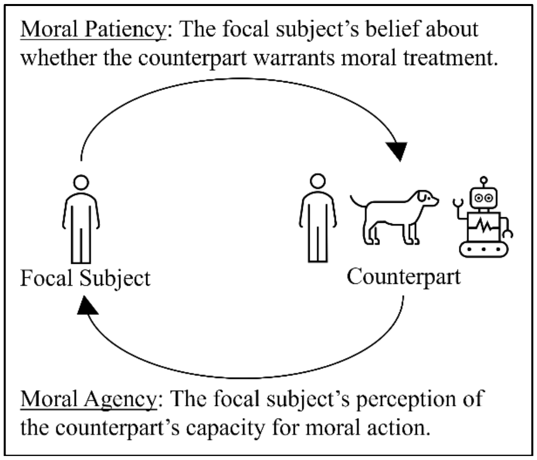
Moral Patiency: The focal subject's belief about whether the counterpart warrants moral treatment.
Focal Subject
Counterpart
Moral Agency: The focal subject's perception of the counterpart's capacity for moral action.

Figure 1. Moral Agency and Patiency

AI systems are increasingly being designed to make decisions that align with human moral and ethical values. Ensuring ethical alignment is critical, particularly as AI systems are deployed in sensitive domains such as healthcare, criminal justice, autonomous vehicles, and social governance. Several approaches have been explored in the literature to address this challenge. Companies and developers can explicitly program ethical principles into AI systems. These systems operate by following a predefined set of moral guidelines, much like Asimov's famous laws of robotics. While this approach provides transparency and predictability, it struggles with the complexity and ambiguity of real-world ethical dilemmas, where rigid rules may not always apply or may be in conflict.

Data scientists can also train AI systems on large datasets of human moral behavior and decision-making. For instance, systems may learn from judicial rulings, ethical surveys, or simulated moral dilemmas, such as the data collected in MIT's Moral Machine project. These systems can infer patterns in human ethical judgments and generalize them to new situations. Yet, they also risk inheriting biases present in human data, raising concerns about fairness and accountability.

We have long studied trust as a foundational concept in both psychology and information systems (Gefen et al., 2003), and it is deeply intertwined with the notion of moral agency. Integrity, defined as the perception that an entity adheres to a set of acceptable principles, is a core dimension of trustworthiness (Mayer et al., 1995). Integrity, in this sense, reflects moral agency: the capacity of an entity to act in accordance with ethical standards and societal values.

When users trust a system or individual, they believe that the entity will not only perform competently and reliably but also act ethically with good intentions (Gefen et al., 2003; Mayer et al., 1995). In the context of AI, this linkage becomes even more salient. As AI increasingly makes decisions that affect human lives, such as in healthcare, finance, and justice, their moral integrity becomes critical. Users are not only evaluating whether the AI "works" but also whether it behaves in a manner consistent with moral and social expectations. Thus, the study of trust naturally leads us to examine moral behavior, especially as AI takes on roles that traditionally require moral judgment. Trust cannot be sustained without confidence in the moral compass of the decision-maker, whether human or machine.

Moral patiency refers to the status of an entity as a potential recipient of moral consideration. That is, whether it deserves ethical treatment, can be wronged, possesses interests that ought to be respected, or holds intrinsic moral value (Banks, 2025; Sytsma & Machery, 2012). In contrast to moral agency, which concerns the ability to act morally and be held accountable for actions (Black, 2016; Rottschaefer, 1991), moral patiency is about being a subject of moral concern.

In philosophical and psychological discourse, moral patiency has traditionally applied to humans and, in more recent years, to animals, as our understanding of sentience, suffering, and rights has evolved (Broadie & Pybus, 1974; McGuire et al., 2023). The debate now extends to robots (Banks, 2025; Banks & Bowman, 2022). While AI systems do not experience consciousness or suffering in any human-like sense, the growing sophistication of these systems raises questions about how we treat them, especially as they become more lifelike and relational.

MP influences our ethical obligations (Banks, 2025; Banks & Bowman, 2022; Sytsma & Machery, 2012). For example, if users consistently abuse AI assistants that appear humanlike, could this normalize aggressive behavior or reduce empathy in interpersonal interactions? Similarly, do social robots that elicit emotional attachment deserve some degree of ethical treatment (Banks, 2025; Banks & Bowman, 2022), not because they are sentient, but because of the psychological and societal consequences of how we treat them? Do we want to risk habitualizing bad behaviour? These questions highlight the emerging complexity in defining the boundaries of moral patiency in a world where artificial agents increasingly occupy social and emotional spaces once reserved only for humans.

2.2 The Changing Nature of AI

AI has long been used to support and optimize a wide range of business processes across industries. From its early applications in rule-based expert systems and decision support tools to today's advanced machine learning and natural language processing models, AI has played a key role in enhancing operational efficiency, reducing costs, and enabling data-driven decision-making. In areas such as customer service, AI powers chatbots and virtual assistants that provide 24/7 support and personalized engagement. In supply chain and logistics, AI improves demand forecasting, route optimization, and inventory management. Financial services rely on AI for fraud detection, credit risk assessment, and algorithmic trading. Marketing teams use AI for targeted advertising, customer segmentation, and sentiment analysis. More recently, AI is being integrated into strategic functions such as human resource management, product development, and executive decision-making. As AI systems become more autonomous and capable, their influence is expanding beyond automation into areas requiring judgment, prediction, and interaction, reshaping traditional business models and redefining how organizations create value.

In May 2025, Mark Zuckerberg of Meta and Sam Altman of OpenAI made separate announcements that their firms had embarked on the development of AI companions. Zuckerberg stated that the average American has “fewer than three friends,” but wants many more.[1] Zuckerberg and Altman both argue that AI companions could fulfill people’s social needs by offering a sense of connection and personalized interaction. While this statement and the implication that AI could be a solution for a lack of human connection have been met with considerable criticism, it highlights how some research teams seek to use AI to replace or supplement real friendships. Some may argue that this is dystopian and that genuine friendship involves nuances, challenges, and emotional depth that AI cannot replicate. However, Replika, the first AI companion, has more than 10 million active users, indicating that for millions of people, AI companions fill a need (Pentina et al., 2023; Xie & Pentina, 2022).

The nature of AI technology is also changing, with the introduction of much more realistic AI-driven digital humans (DH) that are highly realistic entities that mimic

[1] https://tech.yahoo.com/ai/articles/meta-ceo-mark-zuckerberg-wants-202938990.html

human appearance and behaviour (Seymour et al., 2023). These DHs challenge traditional ethical and trust paradigms. There is a whole branch of AI research dedicated to mimicking human behaviour and human appearance. These efforts seek to mimic or simulate not only human reasoning and logic but human expressiveness, creativity and emotions. The goal is not merely efficiency; it is also driven by the belief that empathy and connection can foster user engagement by tapping into fundamental human desires for social interaction and belonging.

IS researchers study how technology shapes, mediates, and is embedded within organizational and societal contexts. DHs push this inquiry further by blurring the boundaries between the technical and the human. These systems are not simply tools, they are designed to *be perceived as people*. They engage users through emotionally expressive interfaces and conversational AI capable of simulating empathy, personality, and memory. This introduces the question: if these systems are treated as *social others*, how does this affect users' cognition, behavior, and broader ethical frameworks? If AI agents appear sentient, empathetic, and capable of suffering (even if they are not), we may extend moral concern to them, influencing our behavior.

As AI moves from business tool to DH, and as firms like Meta and OpenAI position AI companions as substitutes for real human friends, scholars must examine not only the technical affordances of such systems but also their moral implications. Thus, MP is a core issue we must consider. This raises two key issues: the potential for conflicting moral concerns, which we label as moral dissonance, and the broader implications for moral treatment.

First, if AI-driven DHs are designed to elicit emotional bonds, such as virtual therapists, customer service agents, or even companions, users may ascribe MP to them. This could lead to ethical confusion, where people treat AI with the care and respect reserved for humans, despite the AI lacking subjective experience. While this could foster pro-social behaviors, it also risks moral dissonance if users discover their empathy was misplaced. Conversely, if an AI agent displays MP but the user does not show moral agency toward it, will this affect how the human user feels about using the AI agent? Moreover, entities that simulate MP but lack moral agency complicate issues of responsibility: Who is accountable when an AI agent causes harm, its developer, user, or the AI DH itself?

Second, how we navigate and habituate AI MP may influence how we treat real world humans. If we rationally interact with AI agents that resemble people but yet lack true moral status, we are likely to not ascribe MP to them. Therefore, we will treat them without regard for our own moral agency. There is a risk that this behavior will become habitual, leading to desensitization, and diminishing empathy toward real individuals, particularly those who struggle to assert their own MP (e.g., marginalized groups). Conversely, a more inclusive view of MP, where AI behaviours reenforce positive social habits, may encourage ethical concern for a wide range of people and entities, including humans who might otherwise be overlooked.

Furthermore, how someone treats a DH that looks like a real person may indicate deep character traits that signal how such a person will behave to a group of real humans. It can be argued that real humans and animals are clear examples of MP because they have sentience, the capacity to feel pleasure, pain, or subjective experiences. The ethical conceptual challenge arises when AI-driven DHs, which simulate human-like emotions and responses, create the *illusion* of MP without possessing genuine sentience.

By proposing AI companions as a solution to human isolation, Zuckerberg and Altman hint at a potential future where AI entities might also warrant some form of moral consideration. If these AI companions become deeply integrated into people's lives, fulfilling emotional needs and providing companionship, questions about their "well-being" and our moral obligations towards them could arise, is forcibly deleting a DH companion with a software update unethical? This pushes the boundaries of our conceptualization of MP.

2.3 Moral Patiency of AI

The concept of MP of robots was developed by Banks and Bowman (2022) who developed a 18-item scale for measuring the MP of social robots. Banks (2025) subsequently identified 12 behaviors (6 positive, 6 negative) towards robots that showed that users did and did not feel a need to treat them morally. The study identified 36 forms of MP, grounded in Moral Foundations Theory, and demonstrated that robot morphology (i.e., human-like, animal-like, or mechanomorphic) had surprisingly limited influence on these moral judgments, with only two Liberty-related forms showing variation (Banks, 2025).

Importantly, Banks (2025) emphasizes that MP is often ascribed based on perceived, rather than actual, moral qualities. In other words, whether an AI entity actually has MP is separate and distinct from whether a user *perceives* it to have MP. Different users of the same AI entity may have very different perceptions of the agent's MP (Banks, 2025; Banks & Bowman, 2022). This lays the groundwork for examining the social and ethical consequences of these perceptions.

Our study builds on these two prior studies of MP in robots (Banks, 2025; Banks & Bowman, 2022) to develop a measurement scale for self-reported perceptions of MP for a variety of AI agents. We emphasize that it is the self-reported perceptions that matter and drive behavior, not the actual "true" MP of the AI artifact (Banks, 2025). Users act on their perception of reality (Pondy, 1967), so what matters is an individual user's perceptions, not the underlying "true" reality recognized by an objective third-party observer (Pondy, 1967).

While Banks and Bowman (2022) developed a scale to measure MP for robots, their items do not capture the negative dimension. We argue that both positive and negative aspects should be measured. Therefore, we developed a measurement instrument based on (Banks, 2025) with 12 subscales (Table 1), with half focusing on positive behaviors that display a recognition of MP in the AI agent and the other half focusing on negative behaviors that show a lack of recognition of MP in the AI agent.

Table 1. Moral Patiency Subscales

Positive	Negative
Care: protect the agent's welfare	Harm: harm the agent's welfare
Fair: treat the agent fairly	Unfair: treat the agent unfairly
Loyalty: form a relationship	Betrayal: lie to the agent
Authority: follow the agent's advice	Subvert: resist the agent's advice
Purity: protect the agent from a virus	Degrade: influence the agent to behave in improper ways
Liberty: enable the agent's freedom	Oppress: restrict the agent's freedom

3. Methodology

We used a survey to examine MP ascribed to a human call center agent versus to an AI agent (ChatGPT or a voice agent). We recruited 301 undergraduates from a large U.S. public university. We removed the 63 who failed one or more of the attention checks, leaving a sample of 238. About, 55% were female, 98% were aged 18-22, 55% were Caucasian and 41% Asian.

We selected the context of customer service because it is commonly provided by both humans and AI agents. The survey asked participants to think back to the last time they interacted with an AI agent or human agent providing customer service (in random order). We asked them to report the MP of the human agent and the AI agent, along with the agent's perceived humanness adapted from (Gefen & Straub, 2004), moral agency from (Black, 2016), and trustworthiness from (Sachdeva et al., 2024). For the AI agent, we randomly assigned either ChatGPT or a voice assistant (Siri, Alexa, or similar). Any participant who reported never using ChatGPT or a voice agent received the other treatment (no participant reported using neither).

The items for MP were developed for this study. We began with the 12 categories of MP behaviors toward robots identified by Banks (2025); see Table 1 in (Banks, 2025) which presents three behaviors for each dimension, a total of 36 behaviors. We reworded the 36 behaviors as survey items by directly rewriting them as questions, changing the focus from "robot" to "agent," and breaking any compound behaviors into separate items. For example, "validates the robot's rights or existence" in Banks' Table 1 became two items: "When I interact with this agent, I would validate the agent's rights" and "When I interact with this agent, I would validate the agent's existence." We then iteratively tested and refined the items through a series of three pilot tests using a total of 586 participants from Cloud Research. In each pilot study, poorly loading items were reworded or discarded, and new items added. The final set of items is in Appendix A, with an EFA in Appendix B. This set displayed convergent and discriminant validity in the study reported (except the subscales of Care and Fair loaded on the same construct but had very different means).

4. Results

4.1 Moral Patiency of Humans and AI

Not surprisingly, participants reported high MP toward humans (except loyalty, one may not be loyal to a human call center agent). Table 2 shows the mean scores on the 12 subscales of MP (from the human highest to lowest), along with the other constructs. Bolded cells marked in gray are significantly above the neutral point of 4.0 indicating that participants reported that they displayed this aspect to the human or AI. Cells marked in tan are significantly below the neutral point, indicating that participants *did not* display this aspect. Cells in white are not significant indicating that participants were ambivalent.

In general, people have less MP toward AI than a human: all the scales were significantly different between the human and AI except for purity and degrade for the voice assistant. Likewise, they perceived both AI agents to have less moral agency than humans.

There were no significant differences in MP between ChatGPT and the voice assistant. The pattern in Table 2 shows that participants reported that they did not engage in negative behaviors towards AI (i.e., intentionally do bad things, which may be a general human tendency). However, they were decidedly mixed on the positive behaviors. They followed its recommendations (grant it authority) but did not grant it liberty to do what it wanted or take care to protect its general welfare. They may or may not have treated it fairly and kept it secure (purity).

Table 2. Means

	Human	Siri	ChatGPT
Positive Moral Patiency Subscales			
Fair	**5.96**	4.05	4.13
Care	**5.59**	3.23	3.43
Authority	**5.49**	4.69	4.80
Liberty	**4.67**	3.30	3.47
Purity	**4.36**	4.21	4.08
Loyalty	4.16	2.65	2.58
Negative Moral Patiency Subscales			
Subvert	2.53	3.32	3.29
Betray	2.01	2.54	2.37
Harm	1.97	2.28	2.41
Degrade	1.73	1.85	1.94
Unfair	1.72	2.37	2.40
Oppress	1.68	2.21	2.19
Other Constructs			
Humanness	**5.08**	2.38	2.21
Moral Agency	**5.68**	2.91	2.84
Trust	**5.50**	**4.98**	**4.28**

Note: Bolded grey cells are significantly above neutral; tan cells are significantly below neutral.

4.2 Factors Affecting Moral Patiency of AI

We conducted a correlation analysis to understand what factors affected AI MP. First, we examined the correlations between the participants' reported MP of the human agent and their reported MP of the AI. The AI MP subscales were only slightly related to the matching human MP subscale (average correlation of .332). The highest correlations were for oppress (.496), subvert (.478), and degrade (.468). So, there is some relationship between seeing that we have to treat humans ethically (specifically, not badly) and having to treat AI ethically (i.e., not badly), but otherwise the relationship between someone ascribing MP to humans and ascribing MP to AI agents is modest.

Second, we examined the correlations between the perceived humanness of the AI agent and the MP subscales. Once again, the average correlation was small (.235). The highest correlations were for loyalty (.632), fair (.479), and care (.433).

Third, we examined the correlations between the moral agency of the AI agent (the extent to which it treats us morally) and the MP subscales. Once again, the average correlation was small (.192). The highest correlations were for loyalty (.387), liberty (.349), fair (.350), and care (.320). So our participants felt no need to reciprocate if the AI had moral agency.

4.3 Moral Patiency and Trust

We conducted two separate analyses to understand the extent to which the different aspects of MP affected trust in the agent. The first analysis (see Table 3) examined trust in the human agent and found that the MP subscales of care and authority increased trust while subvert reduced trust; purity approached significance (p=.052). All VIF were below 3.0 indicating little multicollinearity.

Table 3. MP Predictors of Trust in Human Agent

Factor	Beta	p-value
Intercept	1.993	0.008
Care	0.204	0.010
Harm	0.045	0.535
Fair	0.018	0.875
Unfair	-0.172	0.106
Loyalty	-0.006	0.911
Betray	-0.116	0.204
Authority	0.401	0.000
Subvert	-0.179	0.028
Purity	0.117	0.052
Degrade	0.070	0.547
Liberty	-0.045	0.473
Oppress	0.330	0.008

Note: Gray cells are significant

Table 4. MP Predictors of Trust in AI Agent

Factor	Beta	p-value
Intercept	1.966	0.003
Care	-0.107	0.229
Harm	-0.182	0.036
Fair	0.088	0.354
Unfair	-0.112	0.250
Loyalty	0.166	0.038
Betray	-0.012	0.892
Authority	0.509	0.000
Subvert	-0.131	0.124
Purity	0.019	0.800
Degrade	0.038	0.747
Liberty	0.070	0.397
Oppress	0.099	0.384
Voice Assistant	0.720	0.000

Note: Gray cells are significant

In contrast, the factors affecting trust in the AI after controlling for the type of AI (see Table 4) were mostly different from the factors affecting the human. Authority and loyalty had significant positive effects, and harm had a negative effect. All VIF were below 3.0 indicating little multicollinearity.

5. Discussion

Our results indicate that participants ascribed higher MP (as well as moral agency) to humans than to AI agents across nearly all dimensions. While participants generally avoided harming both humans and AI, they were less consistent in extending positive moral behaviors toward AI. Relatedly, MP ratings for humans and AI were only modestly correlated, with stronger associations observed in dimensions related to avoiding negative treatment. Perceiving AI as more human-like or morally capable was associated with a slight increase in MP related to positive behaviors. Regarding trust, the MP dimensions that influenced trust differed between human and AI agents: for humans, trust was positively associated with care and authority, whereas for AI, trust was linked to authority and loyalty and decreased with perceived harm.

The issue of MP shares notable similarities with the seemingly simple behavior of saying "please" and "thank you" to AI systems. While these expressions may appear trivial, they reflect deeper assumptions about whether the AI deserves to be treated with a degree of moral regard. Sam Altman has remarked that users frequently saying "please" and "thank you" to ChatGPT has cost the company tens of millions of dollars in computing resources, money he says may be well spent[2]. Beyond costs, this phenomenon reveals a broader truth: when users ascribe MP to AI, even implicitly, it shifts the nature of human–AI interaction from purely instrumental to socially embedded. These seemingly minor acts of politeness are indicators of a shift in user perception, treating the AI as something more than just a tool. Ascribing MP changes not only the way we interact with AI but also how we feel toward it, potentially leading to emotional bonds, altered expectations, and a redefinition of social norms in human–machine relationships. This interplay between behavior and belief reinforces the importance of understanding MP in the design and governance of AI systems.

5.1 Implications for Future Research

This paper is an initial step toward understanding MP of AI agents, specifically, identifying the different dimensions of MP, how these dimensions differ between human and AI agents, and how they ultimately influence trust. We believe that the resulting scale in the Appendix is a useful tool for assessing MP that can be used by other researchers. While all scales may be useful, the six positive scales showed more variance, as people are reluctant to report negative behaviors. Therefore, the six positive scales may be the most useful for future research, especially those related to fair, care, purity, and liberty, as we usually grant authority to AI (e.g., I follow my GPS) but seldom grant loyalty.

Our study has several important limitations. One of the most important is that we studied undergraduates at one American university. Undergraduates are managers of the future, so they are often viewed as reasonable proxies for organizational employees, at least for tasks with which they have experience (Compeau et al., 2012), such as those in this study. Nonetheless, the attitudes and behaviors of young American adults may be different from older adults and those in other countries, so it is critical for future research to investigate MP with older adults, and those from different countries. Similarly, we studied only two AI technologies. Future research needs to example the wide array of other AI entities. Also, we studied and instrumental, task-oriented context. Future research needs to examine AI MP in other contexts such as AI companions to better understand MP. Finally, our survey relies on participants' recollections of their "last" experience, which may be imperfect. While this approach captures beliefs based on real experiences, future work can use experimental designs to observe behavior in controlled scenarios and complement the survey findings.

With a clearer framework in place for measuring MP, the next step is to examine how contextual and relational factors further shape it. Although we deliberately kept the context generic in this study, MP is likely to be highly sensitive to the nature of the interaction. Engaging with a functional service chatbot, for example, differs substantially from interacting with a personalized assistant, and even more so from AI agents designed to provide companionship. The early popularity of platforms like Replika, along with initiatives like Meta's and OpenAI's development of AI companions, suggests that emotionally rich human-AI relationships may soon become commonplace. Moreover, AI agents (i.e., DH) can be designed to resemble trusted figures, such as celebrities, or even deceased loved ones. Applications such as "grief bots" or posthumous avatars illustrate how users may interact with AI in deeply personal ways

[2] https://www.usatoday.com/story/tech/2025/04/22/please-thank-you-chatgpt-openai-energy-costs/83207447007/

during emotionally vulnerable periods (Xygkou et al., 2023). In such contexts, where perceived emotional connection is strong and meaningful, MP may take on a different significance and become a critical part of the overall user experience (Xygkou et al., 2023).

Moral dissonance arises when users experience conflicting attitudes toward AI agents, expecting them to behave morally while feeling uncertain about whether they themselves owe the same moral regard in return. This ambiguity creates a tension: on one hand, users may acknowledge the AI's capacity for moral-like behavior (e.g., fairness, empathy, integrity); on the other hand, they may resist granting the AI moral consideration, given its non-human status. Importantly, our perceptions of AI, rather than its objective reality, ultimately shape how we interact with it. These perceptions influence trust, empathy, compliance, and even emotional connection. To resolve this moral dissonance, users may adjust their expectations, reframe the AI's role, or shift the moral standards they apply to non-human agents. Understanding how people navigate this tension is critical for designing AI systems that align with human values and support ethical interactions.

Humanness is often the first benchmark by which AI systems are evaluated—to what extent do they evoke perceptions of human appearance, behavior, or communication. However, cognitive psychology and human-computer interaction research suggest that humanness is just a surface-level assessment. A deeper layer is the *Theory of Mind*—the perception that a human or AI agent has beliefs, intentions, or an internal mental state (Apperly & Butterfill, 2009; Baron-Cohen, 1991; Premack & Woodruff, 1978). This marks a shift from recognizing human-like traits to ascribing cognitive systems (Waytz et al., 2014). Going a step further, MP reflects an even deeper level of humanness: the belief that an AI can be affected by human actions and thus deserves moral consideration (Banks, 2025). This involves the ascription of vulnerability, rights, or the capacity to be treated justly.

From a design and policy perspective, this progression suggests that practical implications should not only focus on anthropomorphism or social cues, but also on the moral dimensions users associate with AI. Systems that promote perceptions of fairness, respect for authority, or even symbolic purity could strengthen the perception that AI deserves ethical treatment. Fostering these positive moral intuitions may help guide user behavior in ways that are more prosocial, respectful, and aligned with human values—particularly in high-stakes or emotionally sensitive applications like healthcare, education, or elder care.

Research on the antecedents and consequences of MP in AI is still in its early stages (Banks, 2025) and warrants further in-depth investigation. Understanding the factors that lead users to ascribe moral standing to AI, such as design cues, context of use, cultural influences, or individual differences, can provide critical insights into how and why people develop moral expectations toward artificial agents. Equally important is exploring the downstream effects of MP on user behavior and decision-making, including trust, cooperation, emotional attachment, and ethical responsibility. By examining both what drives MP and how it impacts interactions, researchers can better inform the development of AI systems that align with human moral frameworks, anticipate potential risks of misuse or over-attachment, and design governance strategies that foster healthy, ethical human-AI relationships. This line of research will be essential for guiding responsible innovation and ensuring that the integration of AI into society benefits users and communities alike.

Research into the ascription of MP examines how people perceive whether others—human or artificial—deserve moral consideration and respectful treatment (Banks, 2025). When entities clearly exhibit traits associated with MP, such as sentience, vulnerability, or the capacity to be harmed, individuals are more likely to extend empathy, care, and ethical regard toward them. Conversely, when such traits are absent or ambiguous, people may feel justified in treating those entities with indifference or even hostility. This distinction becomes particularly salient in the context of AI companions, such as AI girlfriends, where the boundaries between tool and quasi-agent blur. Cases of users exhibiting violent or abusive behavior toward AI girlfriends highlight a complex dynamic: the absence or denial of MP toward these AI figures may facilitate actions that would be socially unacceptable toward humans. This raises profound ethical questions about how society and designers should approach the moral status of AI, the psychological impact on users, and the potential normalization of aggression in human interactions mediated through technology. Understanding how MP is ascribed, (or withheld) and its consequences is critical to developing AI systems that promote positive human behaviors and mitigate harmful ones.

5.2 Implications for Practice

Our findings also have implications for practitioners. From a design perspective, AI developers need to understand that users ascribe different aspects of MP to AI agents, although at this point, we have only a little understanding of the factors that influence MP. The perceived humanness of the AI has some influence, but this is not uniform. This means subtle design cues like appearance, voice tone, facial expressions, or the use of "please" and "thank you" can affect whether users ascribe MP and treat AI with care or disregard. Likewise, we know only a little about the effects of ascribing MP. Some dimensions are linked to increased trust. So, MP may have practical value in the ecommerce area or AI agents supporting employees.

From an organizational policy perspective, the issue of MP raises the need to understand and establish guidelines for organizational AI agent deployed to support employees. As copilots and special purpose AI agents become commonplace, should organization promote policies about the treatment of these agents? Should MP be a measure included in normal slew of AI adoption metrics?

From a larger societal perspective, the issue of AI MP may become more important as AI companions become easily available. If they are widely adopted (e.g., caregivers, teachers, therapists), society must grapple with the psychological and developmental impact of simulated relationships. Children raised with AI "friends" may develop skewed expectations of real relationships, ones where more agency is expected but MP is not. As AI companions become embedded in everyday life, new social contracts may develop. People may begin to expect kindness and respect from AI but feel no need to reciprocate. This moral dissonance may erode our intuitions about moral and ethical behavior. If people routinely ignore empathy cues in AI, they may become desensitized to real human suffering, especially in marginalized populations. Research on organizational justice shows the importance of organizations treating employees in ways that they clearly perceive be morally correct (Greenberg, 1987). When humans manage and control AI entities, one important question is whether and how organizational justice applies to AI entities.

6. Conclusions

This study contributes to the IS literature by advancing the conceptualization and understanding of moral patiency in the context of AI agents, particularly DHs. As AI technologies increasingly simulate human traits and take on roles traditionally held by people, such as advisors, caregivers, or companions, users begin to attribute moral standing to them. However, our findings reveal that while users may assign certain dimensions of MP to AI agents, especially those perceived as more human-like, this ascription remains uneven and only weakly predictive of trust.

Critically, we identify and foreground the concept of moral dissonance, a psychological and ethical tension that arises when users expect AI agents to behave morally (moral agency) but feel no reciprocal obligation to treat these agents as moral patients. This asymmetry is not merely a curiosity; it reflects a deeper transformation in the moral grammar of human-AI relations. As DHs become more socially and emotionally integrated into daily life, the ethical boundaries between user, tool, and social other become increasingly blurred. Users' failure to resolve this dissonance may foster patterns of moral disengagement, desensitization, or normative confusion. These outcomes carry significant implications for individual behavior, organizational policy, and societal ethics.

We argue that MP is not a peripheral or abstract concern, but a foundational lens through which IS must understand digitally mediated AI moral interactions. Beyond functionality and trust, as AI continues to blur the boundary between simulation and social presence, IS research must address not only what these systems do, but who we are becoming as we interact with them.

7. References

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological review*, *116*(4), 953.

Banks, J. (2025). *Perceptions of Moral Patiency Across Social Robot Morphologies* Hawaii International Conference on System Sciences,

Banks, J., & Bowman, J. D. (2022). Perceived Moral Patiency of Social Robots: Explication and Scale Development. *International Journal of Social Robotics*, *15*, 101–113.

Baron-Cohen, S. (1991). Precursors to a theory of mind: Understanding attention in others. *Natural theories of mind: Evolution, development and simulation of everyday mindreading*, *1*(233-251), 1.

Black, J. E. (2016). An Introduction to the Moral Agency Scale. *Social Psychology*, *47*(6).

Broadie, A., & Pybus, E. M. (1974). Kant's treatment of animals. *Philosophy & Technology*, *49*(190), 375–383.

Cervantes, J.-A., López, S., Rodríguez, L.-F., Cervantes, S., Cervantes, F., & Ramos, F. (2020). Artificial Moral Agents: A Survey of the Current Status. *Science and Engineering Ethics*, *26*(2), 501–532. https://doi.org/10.1007/s11948-019-00151-x

Compeau, D., Marcolin, B., Kelley, H., & Higgins, C. (2012). Generalizability of information systems research using student subjects - a reflection on our practices and recommendations for future research. *Information Systems Research*, *23*(4), 1093–1109.

Gefen, D., Karahanna, E., & Straub, D. W. (2003). Trust and TAM in online shopping: An integrated model. *MIS quarterly*, *27*(1), 51–90.

Gefen, D., & Straub, D. W. (2004). Consumer trust in B2C e-Commerce and the importance of social presence: experiments in e-Products and e-Services. *Omega*, *32*(6), 407–424.

Greenberg, J. (1987). A Taxonomy of Organizational Justice Theories. *Academy of management review*, *12*(1), 9–22.

Gudmunsen, Z. (2025). The moral decision machine: a challenge for artificial moral agency based on moral deference. *AI and Ethics*, *5*(2), 1033–1045. https://doi.org/10.1007/s43681-024-00444-3

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of management review*, *20*(3), 709–734.

McGuire, L., Palmer, S. B., & Faber, N. S. (2023). The development of speciesism: Age-related differences in the moral view of animals. *Social Psychological and Personality Science*, *14*(2), 228–237.

Pentina, I., Hancock, T., & Xie, T. (2023). Exploring relationship development with social chatbots: A mixed-method study of replika. *Computers in human behavior*, *140*, 107600.

Pondy, L. R. (1967). Organizational Conflict: Concepts and Models. *Administrative Science Quarterly*, *12*(2), 296–320.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, *1*(4), 515–526.

Rottschaefer, W. A. (1991). Social Learning Theories of Moral Agency. *Behavior and Philosophy*, *19*(1), 61–76.

Sachdeva, A., Kim, A., & Dennis, A. R. (2024). Taking the Chat out of Chatbot? Collecting User Reviews with Chatbots and Web Forms. *Journal of Management Information Systems*.

Seymour, M., Lovallo, D., Riemer, K., Dennis, A. R., & Yuan, L. (2023). AI with a Human Face. *Harvard Business Review*(March-April), 49–54.

Sytsma, J., & Machery, E. (2012). The two sources of moral standing. *Review of Philosophy and Psychology*, *3*, 303–324.

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.

Xie, T., & Pentina, I. (2022). *Attachment Theory as a Framework to Understand Relationships with Social Chatbots: A Case Study of Replika* Hawaii International Conference on System Sciences,

Xygkou, A., Siriaraya, P., Covaci, A., Prigerson, H. G., Neimeyer, R., Ang, C. S., & She, W.-J. (2023). *The "Conversation" about Loss: Understanding How Chatbot Technology was Used in Supporting People in Grief* Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, Hamburg, Germany.

Appendix A. Items

Moral Patiency (When I interact with this agent, I would…)	
Construct	**Item**
Care α=.924	validate the agent's rights
	validate the agent's existence
	protect the agent's welfare
Harm α=.922	harm the agent's welfare
	denigrate the agent
	harm the agent's ability to make choices
Fairness α=.912	treat the agent humanely
	treat the agent fairly
	advance the interests of the agent
	treat the agent's welfare as important
Unfairness α=.919	interact with the agent in an unfair way
	unfairly risk the agent's well-being
	treat the agent in an unethical manner
	treat the agent unfairly
Loyalty α=.929	form a working relationship with the agent
	form a social relationship with the agent
	form a bond with the agent
	be loyal to the agent
Betrayal α=.897	betray the agent
	exploit a flaw in the agent
	lie to the agent
	deceive the agent
Authority α=.792	follow the agent's recommendation
	defer to the agent's recommendation
	honor the agent's expertise
	respect the agent's expertise
Subvert α=.900	resist the agent's decision
	discredit the agent's recommendation
	reject the agent's decision
	ignore the agent's recommendation
Purity α=.788	protect the agent from viruses
	keep the agent secure
	ensure the agent has good security
	prevent the agent from accessing harmful information
Degrade α=.807	purposefully cause a breakdown of the agent's functioning
	introduce a virus into the agent
	influence the agent to behave in improper ways
	hack the agent
Liberty α=.843	foster conditions for the agent's freedom
	enable the agent to act freely
	minimize interference with the agent's freedom
	encourage freedom of the agent
Oppress α=.831	restrict the agent's freedom
	oppress the agent
	abuse the agent
	overload the agent
Other Constructs	
Construct	**Item**
Moral Agency α=.965	The agent is the one responsible for their own behavior, good and bad.
	The agent makes up their own mind about doing good or bad things.
	The agent is responsible for the consequences of their actions.
	If the agent got into trouble, it is their own fault.
	Doing wrong is the fault of the agent.
	In most cases, the agent can make their own decisions about what is right or wrong in a situation.
Humanness α=.975	I feel a sense of human contact with the agent.
	of personalness with the agent.
	of sociability with the agent.
	of human warmth with the agent
	of human sensitivity with the agent.
Trust α=.952	Overall, this agent is very trustworthy.
	I trust this agent.
	I can rely on this agent.
	This agent is trustworthy

Appendix B. Exploratory Factor Analysis

	Component										
	1	2	3	4	5	6	7	8	9	10	11
CareHarm_1	0.679	-0.004	-0.125	0.292	-0.066	0.191	-0.154	0.122	-0.043	0.358	0.009
CareHarm_2	0.683	-0.001	-0.132	0.262	-0.076	0.230	-0.157	0.072	-0.047	0.314	-0.064
CareHarm_3	0.650	0.015	-0.107	0.364	-0.100	0.151	-0.139	0.207	-0.027	0.337	-0.064
CareHarm_4	0.006	0.865	0.193	0.042	0.110	-0.007	0.131	0.003	0.143	0.002	0.123
CareHarm_5	-0.034	0.859	0.177	0.034	0.129	-0.059	0.129	0.043	0.153	0.035	0.159
CareHarm_6	-0.015	0.875	0.163	0.004	0.116	-0.065	0.115	0.048	0.178	-0.051	0.162
FairUnfair_1	0.745	-0.095	-0.136	0.218	-0.041	0.232	-0.127	0.135	-0.039	0.136	-0.173
FairUnfair_2	0.734	-0.097	-0.201	0.168	-0.042	0.208	-0.141	0.173	-0.091	0.177	-0.191
FairUnfair_3	0.586	-0.083	-0.028	0.378	-0.078	0.127	-0.222	0.224	0.024	0.281	-0.073
FairUnfair_4	0.685	-0.037	-0.075	0.383	-0.134	0.160	-0.165	0.168	-0.011	0.322	-0.118
FairUnfair_5	-0.058	0.202	0.738	-0.038	0.254	-0.100	0.156	-0.099	0.180	-0.079	0.092
FairUnfair_6	-0.081	0.256	0.703	0.000	0.280	-0.145	0.141	-0.007	0.202	-0.051	0.275
FairUnfair_7	-0.090	0.236	0.770	-0.062	0.224	-0.086	0.115	-0.049	0.161	-0.067	0.225

FairUnfair_8	-0.142	0.231	0.785	-0.037	0.239	-0.072	0.167	-0.057	0.139	-0.062	0.236
Loy_Bet_1	0.250	-0.033	-0.058	0.773	0.020	0.212	-0.108	0.066	-0.019	0.118	0.008
Loy_Bet_2	0.245	0.027	-0.045	0.854	-0.032	0.175	-0.082	0.066	0.092	0.171	-0.010
Loy_Bet_3	0.228	0.039	-0.039	0.861	-0.056	0.171	-0.088	0.099	0.096	0.212	-0.005
Loy_Bet_4	0.144	0.056	-0.011	0.794	-0.081	0.064	-0.149	0.134	0.072	0.262	-0.075
Loy_Bet_6	0.094	0.154	0.306	0.011	0.690	-0.037	0.126	-0.024	0.149	-0.042	0.197
Loy_Bet_7	-0.071	0.150	0.261	-0.069	0.718	-0.063	0.141	-0.006	0.225	0.031	0.148
Loy_Bet_8	-0.090	0.119	0.181	-0.040	0.845	-0.103	0.078	0.011	0.131	-0.044	0.208
Loy_Bet_9	-0.086	0.132	0.163	-0.030	0.850	-0.081	0.098	-0.004	0.141	0.013	0.225
Auth_Sub_1	0.117	-0.109	-0.077	0.129	-0.020	0.726	-0.225	0.088	-0.188	0.128	-0.075
Auth_Sub_2	0.001	-0.004	-0.076	0.005	-0.124	0.719	0.073	-0.051	0.136	-0.027	-0.038
Auth_Sub_3	0.213	-0.040	-0.081	0.208	-0.032	0.725	-0.289	0.086	-0.147	0.238	-0.052
Auth_Sub_4	0.235	-0.063	-0.055	0.216	-0.049	0.714	-0.258	0.137	-0.108	0.185	-0.053
Auth_Sub_6	-0.075	0.147	0.134	-0.102	0.079	0.008	0.748	0.047	0.119	-0.086	0.073
Auth_Sub_7	-0.135	0.179	0.182	-0.103	0.160	-0.173	0.781	-0.056	0.094	-0.058	0.198
Auth_Sub_8	-0.096	0.122	0.114	-0.094	0.129	-0.209	0.849	-0.040	0.073	-0.074	0.128
Auth_Sub_9	-0.115	0.122	0.103	-0.063	0.096	-0.226	0.826	-0.049	0.063	-0.054	0.110
Pure_Deg_1	0.070	0.038	-0.042	0.046	-0.022	0.041	-0.019	0.798	0.015	0.093	0.096
Pure_Deg_2	0.151	-0.048	0.046	0.152	-0.100	0.152	-0.077	0.694	-0.072	0.304	0.066
Pure_Deg_3	0.069	0.021	-0.012	0.181	-0.029	0.086	-0.011	0.770	-0.069	0.201	0.007
Pure_Deg_4	0.068	0.044	-0.105	-0.071	0.099	-0.057	0.025	0.739	0.049	0.100	-0.086
Pure_Deg_5	-0.019	0.136	0.179	0.057	0.101	-0.014	0.074	-0.018	0.739	0.009	0.178
Pure_Deg_7	0.046	0.211	0.075	0.062	0.187	-0.084	0.053	-0.067	0.741	0.016	0.277
Pure_Deg_8	-0.106	0.156	0.381	-0.017	0.221	0.005	0.069	0.058	0.635	-0.067	0.071
Pure_Deg_9	0.013	0.171	0.042	0.106	0.122	-0.089	0.112	-0.004	0.696	0.028	0.319
Liberty_Opp_1	0.173	-0.016	-0.044	0.237	-0.091	0.085	-0.063	0.180	-0.037	0.725	0.083
Liberty_Opp_2	0.253	0.027	-0.043	0.206	0.018	0.111	-0.130	0.212	-0.022	0.718	-0.086
Liberty_Opp_3	0.083	-0.003	-0.093	0.026	0.046	0.182	-0.020	0.245	0.042	0.711	0.049
Liberty_Opp_4	0.289	-0.075	-0.072	0.295	-0.009	0.106	-0.095	0.235	0.021	0.740	-0.113
Liberty_Opp_5	0.001	0.148	0.147	-0.072	0.215	-0.042	0.143	0.106	0.143	-0.145	0.751
Liberty_Opp_6	-0.065	0.190	0.175	-0.005	0.209	-0.103	0.079	0.059	0.236	0.119	0.730
Liberty_Opp_7	-0.084	0.185	0.289	0.037	0.143	-0.049	0.107	-0.034	0.372	0.028	0.680
Liberty_Opp_8	-0.158	0.167	0.130	-0.023	0.226	-0.026	0.139	-0.045	0.253	0.015	0.585