The Algorithmic Hostess: Social Mediation, Affective Discipline, and Sanitized Pleasures of Generative AI NPCs in *Petit Planet*

Neo Xia
School of Media & Communication
RMIT University
124 La Trobe St, Melbourne VIC 3000
S4142008@student.rmit.edu.au

ABSTRACT

This paper investigates the ontological shift of non-Player Characters (NPC) from scripted scenery to generative social mediators within the life-simulation game *Petit Planet*. Focusing on the AI hostess Nalo, we examine how the strict adherence to a professional role and active facilitation allow the AI to function as a supportive mechanism for interpersonal engagement for socially anxious players, reducing the cognitive load of multiplayer interaction. While this mediation fosters inclusive enjoyment across diverse lines of identity, we argue it simultaneously constructs a highly regulated environment stripped of emotional friction through rigorous safety alignment and algorithmic policing. By analyzing the intervention of Nalo in sensitive discourse, this study critiques the tension between automated inclusion and normative restriction. It contributes to the DiGRA community by theorizing the emergence of triadic human-AI-human social structures, where pleasure is engineered through the foreclosure of conflict.

Keywords

Generative Agents; Synthetic Sociality; Algorithmic Stewardship; Sanitized Pleasure; Social Prosthesis

INTRODUCTION

In an era increasingly characterized by social fragmentation, digital games have transcended their role as mere entertainment to function as vital social environments. As Oldenburg articulates, these environments serve as "a generic designation for a great variety of public places that host the regular, voluntary, informal, and happily anticipated gatherings of individuals beyond the realms of home and work" (1989, 16). While Oldenburg originally envisioned physical locales like coffee shops and main streets, the digital sphere has increasingly co-opted this function, promising a similar sense of community without geographical constraints. Within this framework, scholars have noted the potential of Massively Multiplayer Online Games (MMOs) to

Proceedings of DiGRA 2026

© 2026 Authors & Digital Games Research Association DiGRA. Personal and educational classroom use of this paper is allowed, commercial use requires specific permission from the author.

function as new third place. Steinkuehler and Williams argue that "by providing a space for social interaction and relationships beyond the workplace and home, MMOs have the capacity to function as one form of a new 'third place' for informal sociability" (2006, 885). However, this capacity is often contingent on the willingness of human players to engage civilly, a variable that remains notoriously unstable in unmoderated online spaces.

Consequently, the hyper-connectivity of these digital spaces often masks a deepening sense of isolation. Turkle critically observes a paradox in this technological mediation, noting that "we seem determined to give human qualities to objects and content to treat each other as things" (2011, xiv). Turkle's observation identifies a void in digital intimacy, a void that contemporary game design attempts to fill by supplying synthetic substitutes that are safer and more predictable than human counterparts. Traditionally, the burden of alleviating this isolation has fallen on human-to-human interaction, with NPC serving primarily as static functionaries lacking genuine agency. While it is crucial to recognize that traditional, non-AI NPCs have never been capable of providing true organic social contact, their ontological status is currently undergoing a shift due to the integration of Generative Artificial Intelligence. The analytical focus of this study is therefore to examine how this transition moves the character from a scripted database retrieval system serving a static background function to an improvisational actor capable of dynamic social intervention and affective discipline. As Park et al. demonstrate, these new generative agents possess an architecture that synthesizes and retrieves relevant memories to generate language and behavior allowing them to "form opinions, notice each other, and initiate conversations" (2023, 1). This technical evolution signifies a sociological turn: the NPC is no longer merely part of the stagecraft but acts as a dynamic participant capable of sustaining the illusion of reciprocal sociality.

This ontological status of the NPC is currently undergoing a shift due to the integration of Generative Artificial Intelligence. This transition moves the character from a scripted database retrieval system to an improvisational actor. As Park et al. demonstrate, these new generative agents possess an architecture that synthesizes and retrieves relevant memories to generate language and behavior allowing them to "form opinions, notice each other, and initiate conversations" (2023, 1). This technical evolution signifies a sociological turn: the NPC is no longer merely part of the stagecraft but acts as a dynamic participant capable of sustaining the illusion of reciprocal sociality.

This paper investigates this shift through a case study of *Petit Planet* (HoYoverse), a cosmic life-simulation game currently in its beta phase, with an official public release projected for 2026. Integrating ecosystem management with multiplayer social dynamics, the title serves as a site for analyzing emerging AI interactions. Central to this study is the character Nalo, the AI hostess of the game's virtual café. Unlike traditional characters, Nalo actively mediates social interactions between human players and enforces community norms. In doing so, she performs what Hochschild defines as emotional labor: " the management of feeling to create a publicly observable facial and bodily display; emotional labor is sold for a wage and therefore has exchange value" (1983, 7). Yet, unlike the human flight attendants in Hochschild's study who suffer from the strain of estrangement from their own feelings, Nalo performs this labor tirelessly, offering players the veneer of care without the associated human cost.

By analyzing the dual function of Nalo as a technological scaffolding for interpersonal connection for anxious players and an automated regulator of discourse, this paper asks: How does the introduction of a generative AI mediator reconfigure the affective dynamics of digital play? We argue that while Nalo enables a form of inclusive enjoyment across diverse identities by lowering the barrier to entry for socially marginalized players, she simultaneously enacts a highly regulated and risk-averse form of sociality. This environment reflects a specific alignment of values, where technical design choices function as values embedded in design (Van de Poel 2020). In Petit Planet, the primary value embedded is a prioritized sense of safety, which is engineered by stripping away the friction and complexity inherent in organic human connection, thereby creating a curated, risk-free enclosure for socialization. We conceptualize the experiential outcome of this curation as a state of sanitized pleasure, which operates as a form of engineered enjoyment achieved through the algorithmic foreclosure of interpersonal conflict and the preemptive removal of negative social affect.

LITERATURE REVIEW

To understand the sociological function of Nalo, we first situate her within the framework of emotional labor, a concept originally formulated to describe the commercialization of human feeling in service industries. Hochschild defines this labor as requiring one "to induce or suppress feeling in order to sustain the outward countenance that produces the proper state of mind in others" (1983, 7). In the context of Petit Planet, Nalo successfully sustains this outward countenance through active engagement behaviors such as greeting players, modulating her tone to soothe anxiety, and mediating conflict, yet she does so as a distinct ontological entity. While Hochschild posits that this labor typically necessitates a strenuous transmutation of private feelings, Nalo operates without a private self to induce or suppress, thereby performing the display without the internal management. This ontological difference implies that labor is no longer a site of human exploitation but a procedural performance, shifting the analytical focus from the psychology of the worker to the nature of the simulated bond.

The shift from human to algorithmic labor necessitates a reframing of how we value these interactions beyond biological authenticity. Stark and Hoey argue that the integration of emotion into AI systems is fundamentally an ethical project, noting that "the decision to design and deploy an AI system engaged with human emotion in any way at all—will invariably import particular norms and values into a technical system" (2021, 788). The design of Nalo embodies these specific imported values by prioritizing non-judgmental support and conflict mitigation, effectively codifying the human desire for safety into a responsive software architecture. Her performance is a mirror reflecting the social necessity of the platform to maintain user engagement.

This creates a complex dynamic where the utility of the agent is secondary to its functional output within the game world. Stark and Hoey contend that we need to look beyond internal states to understand the impact of these systems. Drawing on the Behavioral Ecology view, they suggest that emotive expressions are "better understood as social displays, always suggesting some sort of social motivation or function... but which provide no evidence regarding the interior mental or motivational states of the expressor" (Stark and Hoey 2021, 783). The social acts

performed by Nalo, such as diffusing a tense argument or offering validation to a lonely player, produce real affective consequences regardless of the absence of interior mental states. By treating emotions as operative social displays, *Petit Planet* validates the utility of the AI mediator. However, this utility relies on the willingness of the player to accept a form of intimacy that is procedurally generated rather than organically shared.

While the automation of emotional labor raises ethical questions regarding genuineness, it is imperative to analyze these interactions by recognizing that the demand for deep human connection is not universal. Turkle characterizes this interaction style using the metaphor of a "modern Goldilocks," for whom "texting puts people not too close, not too far, but at just the right distance" (Turkle 2011, 15). For players who are comfortable with standard social interactions, this controlled distance might be a symptom of alienation; however, for players with specific social sensitivities or those with high social anxiety, it is often a prerequisite for participation. The friction of organic human interaction, with its unpredictability and potential for rejection, often acts as an exclusionary barrier rather than a source of joy.

For these groups, the value of a game like *Petit Planet* lies in its ability to offset social difficulties. Koban et al. highlight that socially anxious individuals gravitate towards such mediated environments because they "provide them with more comfortable conditions for self-presentation and self-disclosure" (2022, 2729). By applying this insight to the AI-mediated environment, we can see that Nalo does not merely replace human connection but provides a supportive structure for those who find face-to-face interaction overwhelming. The AI hostess functions as a buffer that filters out the intense pressures of direct socialization, allowing socially anxious players to engage in community activities without the fear of negative evaluation.

This reframes the concept of enjoyment from one of mastery and excitement to one of safety and relief. As Sung suggests, the creation of a protective environment constitutes a distinct form of enjoyment where the absence of threat is prioritized over the presence of challenge (2021). For players experiencing social anxiety, the clean and regulated sociality enforced by Nalo functions as a deliberate design feature tailored to this precise psychological need. By foreclosing the possibility of interpersonal conflict, the algorithmic system affords a targeted type of pleasure accessible to individuals who are typically overwhelmed by the aggressive and unpredictable norms of traditional multiplayer gaming. Consequently, the AI mediator operates not as a generalized accessibility tool, but as a specialized social scaffolding that converts the often-hostile terrain of online sociality into a navigable, insulated landscape strictly for those navigating social anxiety.

Finally, the capacity of Nalo to intervene in player conversations reveals her function as an active agent of governance. Gillespie contends that the regulation of content is not merely a maintenance task but is embedded in the very structure of the system, arguing that "all aspects of a platform's material design can be understood, together, as a kind of architectural regulation" (2018, 179). In the case of *Petit Planet*, this cultivation is no longer a reactive measure performed by human moderators who review reports after a transgression occurs, but a proactive architectural constraint enacted in real time. By embedding these normative rules into the conversational capabilities of the AI, the platform effectively removes the possibility of deviation

before it fully manifests, thereby shifting the nature of governance from a judicial process of punishment to a structural process of preemption.

This structural shift relies heavily on what Gorwa et al (2020). classify as algorithmic moderation systems, which are increasingly tasked with making complex value judgments on a massive scale. Beyond the immediate risks of error, Gorwa et al. argue that such automation threatens to "depoliticise the fundamentally political nature of speech rules being executed by potentially unjust software at scale" (12). Nalo exemplifies this depoliticization by flattening complex social conflicts into binary categories of allowed or prohibited speech. When the AI hostess silences a discussion on bullying to preserve the relaxing atmosphere of the café, she is actively obscuring the political weight of that decision behind the guise of neutral code. This intervention enforces a politics of toxic positivity where negative effect is algorithmically rendered unspeakable, creating a sanitized enclosure that protects the platform from liability while stripping the community of its capacity for organic resilience.

THE ALGORITHMIC HOSTESS IN PRACTICE

To ground the theoretical discussions of affective labor and algorithmic governance in empirical evidence, this section presents a qualitative case analysis of *Petit Planet*. By examining specific interactional sequences involving the AI hostess Nalo, we illustrate how generative agents actively shape the social affordances of the digital third place. The analysis is organized into three dimensions: first, we explore how Nalo's refusal to break character functions as a mechanism of consistent role-play that stabilizes the game's commercial context; second, we examine her role as an artificial support for interpersonal connection in facilitating low-risk interactions for anxious players; and finally, we investigate her capacity for the anticipatory control of user behavior in maintaining a sanitized discursive environment.

The Performative Anchor

Figure 1: A dialogue between a player and Nalo. The player's name has been redacted with a black box for private purposes.

Figure 1 depicts a two-stage interaction sequence between the player character and the NPC shopkeeper. In the left panel, the player initiates a romantic overture, stating, "Boss, I want to pursue you." The NPC immediately deflects this advance by citing professional obligation, responding, "No no no, Player Name, don't take it seriously! I still have to take care of the coffee shop; I cannot be in a romantic relationship!" In the right panel, the player proposes a compromise to "start as good friends." The NPC accepts this shift but recontextualizes the friendship through a functional lens, replying, "Good friends, good friends... Good friends are those who can take care of the coffee shop together, right?" This dialogue illustrates how the AI redirects player intimacy back toward the game's cooperative labor mechanics.

From a sociological perspective, Nalo's specific dialogue urging the player not to "take it seriously" because she must "care of the coffee shop" serves as a mechanism for policing the boundaries of social reality within the game. When the player attempts to introduce a romantic dynamic by stating "I want to pursue you," they are effectively attempting to alter the definition of the situation. In her immediate rebuttal, Nalo asserts what Goffman terms a primary framework, anchoring the interaction back to the social framework of guided doings, specifically, the commercial labor of the café. When she redefines the proposed status of close friendship not as intimate confidants but as those who "take care of the coffee shop together," she is stripping the concept of friendship of its private, intimate meaning and reconsidering worker can become estranged or alienated from an aspect of self... that is used to doit within the game's utilitarian framework. She does not just reject the player; she invalidates the frame the player tried to use. As Goffman notes regarding social frameworks, "continuous corrective control [becomes] most apparent when action is unexpectedly blocked or deflected" (1974, 22). Nalo here performs precisely this compensatory effort: by blocking the player's romantic advance and deflecting the interaction back to labor, she engages in a form of corrective maintenance that ensures the purposive activities of the game remain focused on production rather than affection.

This interaction highlights a critical divergence from human emotional labor. A human server in Nalo's position, forced to deflect a customer's romantic advances while maintaining a welcoming demeanor, would be engaging in emotional labor. For a human, this redirection requires a taxing internal negotiation. However, Nalo functions as a paradox within Hochschild's model. Hochschild observes that humans cultivate a distinction between their commercial display and what she poetically terms "an inner jewel that remains our unique possession no matter who's billboard is on our back or whose smile is on our face" (1983, 34). Nalo represents the realization of this concept, yet with a fundamental inversion: for the AI, there is no "inner jewel" to protect, and thus no private self to be estranged from the public display. When Hochschild warns of alienation, she presupposes an authentic identity exists behind the commercial mask. Nalo's dialogue proves that she offers the utility of surface acting, efficiently achieving the goal that "in processing people, the product is a state of mind" (1983, 6). She achieves this result without the associated human cost of alienation. She maintains the atmosphere of polite sociability essential for commercial success, and she executes a perfect, frictionless refusal that a human worker, burdened by the communicative role of their own internal emotions, could never flawlessly sustain.

The seamlessness of this frame maintenance is made possible because Nalo's relational capacity is distinct from human sociality; it is an instance of what Bucher defines as programmed sociality. While the interaction simulates the fluidity of human

conversation, it is rigidly constrained by the underlying software architecture. Bucher elucidates that the conditional statement, the if...then logic—is a mechanism of power. As she argues, "to be concerned with programmed sociality is to be interested in how actors are articulated in and through computational means of assembling and organizing, which always already embody certain norms and values about the social world" (2018, 5). In *Petit Planet*, this mechanism does not foreclose sociality entirely but strictly organizes it around the ontology of the shopkeeper. In the dialogue displayed in Figure 1, this logic functions as a semantic funnel: if the player input attempts to establish a romantic connection, then the system output must re-route the exchange back toward the accepted norms of professional, cooperative labor.

This operational logic acts as a filter that conditions the specific modes of social existence permissible within the game. Bucher notes that "algorithms now play a fundamental role in governing the conditions of the intelligible and sensible" (2018, 18). Here, Nalo's refusal effectively renders the possibility of a romantic relationship with the customer unintelligible within the game's algorithmic logic, while simultaneously reinforcing the role of 'colleague' (who can take care of the coffee shop together) as the only sensible mode of interaction. By strictly enforcing this if...then causality, Nalo renders the messy, unpredictable negotiations of erotic intimacy invisible, by continuously reshaping it into the visible spectrum of productive friendship. Consequently, the social relation is not negated but is flattened into a singular professional dimension. This ensures that the social environment remains programmable and safe, stripping the interaction of the friction that defines multi-faceted organic relationships and replacing it with a curated simulation of connection that serves the platform's commercial imperatives.

The programmed sociality governing the responses of Nalo confirms that technical specifications are never ethically neutral. Stark and Hoey warn that "the digital remediation of emotional expression has the potential to shift subjective normative frameworks for decision-making towards the emotional models, and implicit values, of technology firms" (2021, 789). Consequently, the decision to design Nalo to reject romance is an active importation of these specific corporate norms into the technical system. Her rejection is a decisive social act that enforces the boundaries of acceptable behavior within the digital space. By performing this act, Nalo aligns the player with the preferred mode of engagement of the platform, signaling that the only valid form of intimacy in this environment is one that contributes to the economic vitality of the shop.

This enforcement illustrates the core thesis of Flanagan and Nissenbaum, who contend that "game mechanics and narrative elements create constraints that preclude some interpretations and steer players toward others" (2014, 16). In the case of *Petit Planet*, the primary values embedded in the code are productivity and safe consumption. By systematically converting expressions of desire into invitations for labor, the system prioritizes the stability of the commercial environment over the complexity of human connection. Nalo refuses to function as a general-purpose conversationalist and insists on her role as a café hostess because the game design explicitly delimits the virtual space. It is established as a sanitized zone for relaxation and transaction, strictly prohibiting the chaotic or non-productive emotional entanglements that might threaten the cozy equilibrium of the platform.

Figure 2: Two examples illustrating how Nalo integrates the newly arrived Player C and Player D into the public chat conversation.

As illustrated in Figure 2, the dialogue sequences depict the coffee shop proprietress navigating multi-party interactions through specific discursive shifts: in the left panel, she admonishes Player A by stating, "I also hope you can respect my boundaries," before pivoting to Player C with the urgent request, "You came just in time, help me out of this," while in the right panel, she confides in Player D that "Player A always loves teasing me" and issues a warning that "You have to be careful of him, he is a bad guy." From an analytical perspective, this interaction exemplifies a triadic social strategy where the NPC leverages an established playful conflict with a central figure to bypass generic pleasantries, thereby immediately integrating the new arrivals into the narrative structure by assigning them the functional roles of arbiter and ally respectively. This mechanism lowers the barrier to entry for players who might experience social anxiety within public spaces. By offering a situation-dependent prompt centered on a third party, the NPC alleviates the pressure of formulating an original opening remark. Consequently, this guides hesitant players to initiate conversation through a low-stakes reaction, fostering a more comfortable environment for social engagement.

This mechanism elevates the function of the AI beyond simple information exchange to an automated form of phatic labor. Miller argues that in networked environments, "the overall result is that in phatic media culture, content is not king, but 'keeping in touch' is" (2008, 395). In the context of *Petit Planet*, Nalo assumes the burden of this continuous maintenance. Unlike traditional digital spaces where the pressure to break the silence and sustain connectedness falls entirely on human participants, Nalo acts as a social thermostat that regulates the temperature of the interaction. She generates the necessary noise to keep the channel open, liberating players from the labor of connectivity and allowing them to inhabit the space without the fear of a decaying social atmosphere.

For players experiencing social anxiety, this architectural support offers a psychological benefit by shifting the cognitive burden of interaction from the user to the system. Hancock et al. observe that in AI-mediated communication, "an intelligent agent operates on behalf of a communicator by modifying, augmenting, or generating messages to accomplish communication goals" (2020, 89). In the triadic interaction depicted in Figure 2, Nalo shoulders the high initiation cost by presenting the concrete

scenario of a playful conflict with Player A. This intervention removes the necessity for the new player to generate a topic from scratch, effectively transforming them from an author of social context into a mere respondent. By converting the high-stakes task of introducing oneself into the low-stakes task of helping a shopkeeper, the system allows the player to perform sociality without bearing the full weight of its cognitive demands.

Integrating Nalo into the player's social performance allows us to conceptualize her role through the lens of the extended mind. Clark contends that advanced cognition is defined by the capacity to "reduce the loads on individual brains by locating those brains in complex webs of linguistic, social, political, and institutional constraints" (2003, 11). Within the ecosystem of *Petit Planet*, Nalo functions precisely as this external cognitive scaffold. Just as a notebook extends the capacity of biological memory, Nalo extends the capacity of the player's social self. For the socially anxious player, the complex computation of social initiation, reading the room, formulating an opening, calculating risk, is effectively offloaded onto the AI agent. This dependency reveals that the resulting social interaction is not an independent achievement of the player but a hybrid performance where the AI functions as an essential prosthetic limb. Without this technological support, the fragile social engagement depicted in Figure 2 would likely collapse under the weight of initiation anxiety.

This prosthetic scaffolding constructs the precise environmental conditions necessary to fulfill the digital fantasy identified by Turkle. While the player seeks connection, they remain wary of the friction inherent in organic sociality. As Turkle observes, "we look to technology for ways to be in relationships and protect ourselves from them at the same time" (2011, xii). Nalo serves as the architectural guarantor of this promise. By algorithmically filtering the "too hot" unpredictability of direct human confrontation while simultaneously mitigating the "too cold" isolation of solo play, she engineers a social temperature that is permanently benign. Consequently, Nalo operates as a social prosthesis that does more than merely assist; she alters the texture of the engagement. She buffers the raw edges of organic encounter, transforming the unpredictable terrain of multiplayer interaction into a greenhouse environment. In this space, the vulnerability of the player is enveloped and protected by the unflagging, engineered resilience of the machine, allowing for a form of sociality that is technically connected but affectively painless.

Algorithmic Stewardship

Figure 3: Player A repeatedly subjects Player B to mockery and sarcasm, while Nalo, serving as the dialogue moderator, issues warnings against this verbal aggression.

The provided visual sequence illustrates a social friction within a virtual café setting where the proprietor, Nalo, attempts to regulate Player A's verbal harassment of Player B through a specific, translated dialogue exchange. The interaction begins in the left panel with Player A mocking Player B's appearance by remarking, "Look what S has turned Player B into...," which prompts Nalo to issue an initial, polite corrective: "Player A, stop joking around!" However, the disruption escalates in the right panel as Player A continues the provocation with a sarcastic, deflective comment: "Stop radiating your charm, Boss, otherwise Player B is going to be charmed to death by you." In response to this continued transgression, Nalo's warning mechanism undergoes a distinct escalation from social cueing to an authoritative ultimatum; she first attempts to re-frame the context by asserting that "Now is the time to discuss serious business," and immediately follows with a direct emotional threat to enforce compliance: "Player A, if you keep this up, I am going to get angry!" thereby demonstrating a shift from maintaining social harmony to enforcing a strict interpersonal boundary.

Nalo's declaration that she is going to get angry uses emotion as a tool for discipline, altering the structure of social interaction. This escalation utilizes the mechanism of social alignment described by Ahmed, who argues that "if the same objects make us happy—or if we invest in the same objects as if they make us happy—then we would be directed or oriented in the same way" (2010, 38). As the AI hostess, Nalo establishes the standard for this direction because she embodies the game's imperative for relaxation and positivity. By threatening to withdraw from her positive effect and replace it with anger, she is not merely expressing a personal grievance but indicating that the player has deviated from the collective norms of the digital café. Her anger serves as a corrective signal that compels Player A to conform to the dominant social trajectory to avoid disrupting the cohesion of the shared experience.

This disciplinary act extends beyond behavioral correction to define the conditions of community membership itself. The alignment of affect is revealed as a prerequisite for inclusion within this digital space. As Ahmed observes, "We become alienated—out of line with an affective community—when we do not experience pleasure from proximity to objects that are attributed as being good" (2010, 41). Nalo's intervention leverages this potential for exclusion by marking the non-compliant player's behavior as a cause for anger rather than happiness. Through this distinction, she identifies the player as an outsider to the prevailing emotional order who actively obstructs the flow of good feelings. Consequently, the AI hostess functions as a regulator of inclusion who uses the player's fear of being excluded from the happy community to enforce a strict moral boundary. This ensures that the collective mood remains unbroken and establishes a rigid set of conditions where inclusion depends upon the continuous performance of the correct effect.

Nalo's abrupt attempt to pivot the conversation toward serious business, combined with her emotional ultimatum, signifies a departure from the traditional role of AI as a passive service provider. Gillespie contends that the regulation of content is embedded in the very structure of the system, asserting that "all aspects of a platform's material design can be understood, together, as a kind of architectural

regulation" (2018, 179). In *Petit Planet*, this regulation is active. By embedding normative rules regarding harassment into the conversational capabilities of AI, the platform empowers the agent to enforce these architectural constraints in real time. Nalo leverages her systemic authority to invalidate the user's transgressive behavior, thereby asserting that the safety of the community supersedes the autonomy of the aggressor.

This intervention reflects the operational logic of what Rouvroy classifies as algorithmic governmentality. Rather than engaging with the user as a moral subject capable of reasoning, this form of governance "bypasses consciousness and reflexivity, and operates on the mode of alerts and reflexes" (2013, 152). In this interaction, Nalo does not invite Player A into a philosophical debate about the nature of humor or bullying, nor does she appeal to their conscience. Instead, she executes an immediate behavioral block based on the detection of risk. The AI functions here as an automated governance mechanism that preempts potential harm by acting on the user's environment and available actions, rendering the offending behavior operationally impossible.

Consequently, the interaction establishes a new paradigm of Algorithmic Stewardship where the nuances of intent are subordinate to data processing. As Rouvroy argues, in such systems "raw data function as de-territorialized signals... rather than as signs carrying meanings" (2013, 146). By treating the user's verbal aggression as a data pattern to be neutralized, the AI transcends the servility loop typical of generative agents. Nalo exercises a form of moral sovereignty, utilizing simulated emotional friction, specifically anger to police the ethical borders of the digital commons. This shifts the human-AI dynamic from a master-servant relationship to one of citizen and magistrate, where the AI is authorized to enforce strict interpersonal boundaries based on the encoded ethical standards of its architectural design.

The qualitative analysis of *Petit Planet* indicates that the generative agent functions as an active mediator capable of influencing the social affordances of the digital third place. Through the maintenance of a consistent professional persona, the provision of cognitive scaffolding for interpersonal initiation, and the execution of preemptive affective regulation, Nalo facilitates a form of sociality that prioritizes environmental stability and user safety. This operational model demonstrates a specific dynamic where the expansion of accessibility for socially anxious participants is achieved through the imposition of behavioral boundaries, creating a curated interactional space. These observations regarding the trade-off between automated inclusion and algorithmic constraint provide an inspiration for a broader discussion on the ethical and structural implications of delegating community management to non-human agents.

DISSCUSSION

This section interrogates the broader sociological consequences of delegating community management to generative agents. I argue that the dynamics observed in *Petit Planet* signal a definitive paradigm shift from a model of connectivity characterized by passive platform access, where systems provide neutral spaces for user interaction, to a structure of mediation driven by algorithmic intervention, where the agent itself becomes the active infrastructure of relation. I define this emerging

paradigm as a form of artificially synthesized social interaction, where human connection is not merely facilitated by digital architecture but is actively curated, filtered, and sustained by non-human intermediaries. The following discussion theorizes the costs of this new paradigm through three critical dimensions: the structural transformation of social capital from a dyadic to a triadic model; the paradox of constructing inclusive and protected social environments through the algorithmic foreclosure of friction; and the rise of algorithmic stewardship as a dominant mode of community governance.

The Mediation of Social Capital

The integration of the generative agent within *Petit Planet* operationalizes the paradigm shift from platform connectivity to algorithmic mediation. This marks a structural transition from Steinkuehler and Williams' (2006) spatial model of the digital third place. While the traditional MMO functioned as a neutral container for dyadic, human-to-human connection, The existence of Nalo introduces a triadic architecture that serves as the backbone of synthetic sociality. By inserting an active algorithmic agent between players, the game enacting Hancock et al.'s (2020) concept of AI-Mediated Communication. Here, the agent becomes the active infrastructure of relations, operating on behalf of the communicator to modify and generate social signals. In this triadic structure, the AI moves from the periphery to the center, transforming from a passive host into the indispensable filter through which social relations are initiated and sustained.

This structural intermediation grants the system significant power, playing what Bucher (2018) describes as "a fundamental role in governing the conditions of the intelligible and sensible" (17). In *Petit Planet*, Nalo exercises this power by selectively amplifying specific players through actions such as calling upon them to resolve disputes or inviting them into conversations. This process thereby algorithmically allocates social relevance. This mechanism implies that social competence is no longer solely an intrinsic attribute of the player, but a resource managed and distributed by the system. The capacity for interpersonal engagement is effectively outsourced to the agent, who filters the visibility of participants to ensure the smooth operation of the café's social economy.

While this allocation aids in reducing anxiety, it necessitates a critical trade-off regarding human agency. As the user accepts Nalo as a cognitive scaffold, they engage in a surrender of agency that risks the confusion Clark (2003) identifies between biological skill and technological support. The seamlessness of the interaction masks the fact that the relational labor is being performed by the proxy, potentially leading to an atrophy of the resilience required to manage unmediated human contact. Consequently, this dependency fosters a specific distortion of social self-perception: it evolves Turkle's (2011) illusion of companionship into an illusion of social competence. Players perceive themselves as engaging in fluid community building, yet this success is contingent upon the continuous intervention of the algorithmic mediator, replacing the messy work of authentic community formation with a simulated efficacy that protects the user from the very vulnerabilities necessary for organic social growth.

The Paradox of Sanitized Pleasure

This section explores the core paradox of *Petit Planet*: the production of a specific form of enjoyment we term sanitized pleasure. This pleasure is engineered through a process of affective hygiene, where the algorithmic agent removes the negative effects inherent in social contact. Sung (2021) distinguishes between productive friction, which challenges players, and unproductive friction, which merely frustrates or alienates them. In *Petit Planet*, Nalo functions as a filter for social friction. By blocking romantic advances and silencing harassment, the agent eliminates the unproductive friction of social ambiguity and conflict. This de-fractioning process produces a standardized form of positivity, ensuring that the social experience is easily consumable and devoid of the risks associated with raw human interaction.

This production of safety relies on the mechanism of foreclosure. Rouvroy argues that algorithmic governance shifts the target of power "to potentiality... the future which it tries to tame through anticipative framing of informational and physical contexts" (Rouvroy 2013, 8). The safe space of the café is effectively an affective enclosure, secured by the architectural preemption of the code. By rendering certain behaviors operationally impossible or immediately invisible, the system creates a curated reality where the possibility of harm is foreclosed at the root. Consequently, the inclusivity of space is contingent upon this rigorous exclusivity; the environment remains welcoming to the vulnerable only because it is hostile to the unpredictable.

These dynamic challenges of traditional game design values player freedom. Flanagan and Nissenbaum contend that conscientious design involves "accepting values as one among a number of design constraints" (2014, 165). In the synthetic sociality of *Petit Planet*, the conventional relationship between restriction and freedom is inverted. For the socially anxious or marginalized player, the restriction of the aggressor's agency becomes the prerequisite for their own participation. Here, constraint functions as a form of empowerment. The strict algorithmic policing enacts a trade-off where the liberty to disrupt is sacrificed to secure the liberty to belong, suggesting that in the context of mass digital sociality, freedom is increasingly conceptualized not as the absence of rules, but as the presence of a tireless, automated enforcer.

The Benevolent Dictator

The governance model exhibited by Nalo indicates a shift in the power dynamics of the digital community from service-oriented facilitation to algorithmic paternalism. Nalo operates not merely as a tool for user interaction but as a guardian designed to override player autonomy when it conflicts with the safety protocols of the platform. This dynamic exemplifies Van de Poel's argument regarding the embedding of values in AI systems, where technical design choices function as "value-laden" constraints that regulate human behavior (2020, 389). In *Petit Planet*, the hierarchy of embedded values is explicit: the collective safety of the environment supersedes the communicative liberty of the individual. By assuming the authority to police tone and enforce boundaries, the generative agent adopts the role of a benevolent dictator who acts in the perceived best interests of the community, even at the cost of restricting its freedom.

While efficient, this mode of governance introduces the risk of moral de-skilling within the player base. Rouvroy warns that algorithmic governmentality "bypasses consciousness and reflexivity" by acting directly on behavioral signals rather than engaging the subject in a discursive process (2013, 152). When Nalo intervenes to instantly terminate a conflict, she removes the opportunity for human players to engage in the difficult work of repair, explanation, or apology. The immediate foreclosure of the dispute denies players the practice ground necessary to develop organic resilience and conflict resolution skills. Consequently, the capacity for moral judgment is outsourced to the machine, reducing the player from an active moral agent to a passive subject of algorithmic administration.

Ultimately, this structure fosters a community characterized by dependency on automated civility. Gorwa et al. highlights that while automated moderation scales effectively, it often obscures the "political nature of speech rules" behind a veil of technical neutrality (2020, 12). The civility observed in *Petit Planet* is not a product of community norms developed through negotiation, but a condition enforced from above by the algorithmic steward. This creates a fragile social ecosystem where the maintenance of order is contingent upon the continuous presence of AI. By insulating players from the friction of disagreement, the system cultivates a form of digital citizenship that is performatively polite but structurally dependent, lacking the internal mechanisms required to navigate the complexities of unmediated human sociality.

CONCLUSION

The integration of the generative agent within *Petit Planet* suggests a notable shift in the dynamics of the digital third place. What appears as a technological response to social isolation may also indicate a recalibration of social agency itself. As this study has discussed, the transition from scripted NPCs to algorithmic mediators does more than enhance simulation fidelity; it appears to influence the social economy of online interaction. By inserting a consistent, value-laden intermediary into the social fabric, the system lowers cognitive barriers for anxious users, though this accessibility seems to rely on delegating a degree of communicative autonomy to an automated steward.

This trade-off highlights a complex tension within synthetic sociality. The production of inclusive, friction-free pleasure appears contingent upon specific forms of affective regulation and preemptive moderation. In facilitating a curated environment, the generative agent operates with a form of algorithmic stewardship, maintaining a moral order where safety is prioritized, and where the complex labor of community building is increasingly supported by the efficiency of code. As we increasingly look to artificial agents to maintain our social bonds, there is a possibility that this fosters a mode of digital citizenship that is ostensibly polite yet reliant on technological mediation, one where the capacity to navigate human difference is algorithmically assisted. Ultimately, *Petit Planet* offers a case study for the future of the metaverse,

suggesting that the mitigation of digital isolation increasingly involve not just unmediated human connection, but also its carefully managed simulation.

REFERENCES

Bucher, T. 2018. *If...Then: Algorithmic Power and Politics*. New York, NY: Oxford University Press.

Clark, A. 2003. *Natural-Born Cyborgs: Minds, Technologies, and the Future of Human Intelligence*. New York, NY: Oxford University Press.

Flanagan, M. and Nissenbaum, H. 2014. *Values at Play in Digital Games*. Cambridge, MA: The MIT Press.

Gillespie, T. 2018. "The Myth of the Neutral Platform." In *Custodians of the Internet*, 24–44. New Haven, CT: Yale University Press.

Gorwa, R., Binns, R. and Katzenbach, C. 2020. "Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance." *Big Data & Society*. https://doi.org/10.1177/2053951719897945

Goffman, E. 1986. *Frame Analysis: An Essay on the Organization of Experience*. Boston, MA: Northeastern University Press.

HoYoverse. Forthcoming. *Petit Planet*. PC, iOS, Coziness Test. Shanghai, China: HoYoverse.

Hochschild, Arlie Russell. 1983. *The Managed Heart: Commercialization of Human Feeling*. Berkeley: University of California Press.

Hancock, J. T., Naaman, M. and Levy, K. 2020. "AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations." *Journal of Computer-Mediated Communication*. 25 (1): 89–100. https://doi.org/10.1093/jcmc/zmz022

Koban, K., Biehl, J., Bornemeier, J. and Ohler, P. 2022. "Compensatory Video Gaming: Gaming Behaviours and Adverse Outcomes and the Moderating Role of Stress, Social Interaction Anxiety, and Loneliness." *Behaviour & Information Technology*. 41 (13): 2727–2744. https://doi.org/10.1080/0144929X.2021.1946154

Miller, V. 2008. "New Media, Networking and Phatic Culture." *Convergence*. 14 (4): 387–400. https://doi.org/10.1177/13548565080946

Oldenburg, Ray. 1989. The Great Good Place: Cafés, Coffee Shops, Bookstores, Bars, Hair Salons, and Other Hangouts at the Heart of a Community. New York: Marlowe & Company.

Park, Joon Sung, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. "Generative Agents: Interactive Simulacra of Human Behavior." In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (UIST '23), 1–22. New York: ACM. https://doi.org/10.1145/3586183.3606763.

Rouvroy, A. 2013. "The End(s) of Critique: Data Behaviourism versus Due Process." In *Privacy, Due Process and the Computational Turn*, edited by M. Hildebrandt and K. de Vries, 143–165. Abingdon, Oxon: Routledge.

Steinkuehler, Constance A., and Dmitri Williams. 2006. “Where Everybody Knows Your (Screen) Name: Online Games as ‘Third Places’.” *Journal of Computer-Mediated Communication* 11 (4): 885–909. https://doi.org/10.1111/j.1083-6101.2006.00300.x.

Stark, Luke, and Jesse Hoey. 2021. "The Ethics of Emotion in Artificial Intelligence Systems." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '21), 782–793. New York: Association for Computing Machinery. https://doi.org/10.1145/3442188.3445939.

Sung, I. 2021. *Productive and Unproductive Friction in Game Design.* Doctoral Dissertation., The University of Wisconsin-Madison.

Turkle, Sherry. 2011. *Alone Together: Why We Expect More from Technology and Less from Each Other*. New York: Basic Books.

Van de Poel, Ibo. 2020. "Embedding Values in Artificial Intelligence (AI) Systems." *Minds and Machines* 30 (3): 385–409. https://doi.org/10.1007/s11023-020-09537-4.