Unintentionally Breaking the Fifth Wall: How Generative AI Invites Metalepsis in Real Life

Louis Escouflaire

Catholic University of Louvain, Ruelle de la Lanterne Magique 14, 1348 Ottignies-Louvain-la-Neuve, Belgium

Abstract

This position paper explores how the use of generative AI in storytelling introduces unwanted instances of metalepsis (disruptions of narrative boundaries). Through the prism of "fifth wall breaks", we examine a variety of examples of how narrative AI systems unintentionally generate inconsistencies or self-references, shattering users' referential illusion and sense of immersion. The paper also introduces the concepts of "real-life metalepsis", illustrated by additional examples where users abruptly encounter the artificial nature of AI interactions, breaking the "relational illusion" and potentially causing confusion. Future research will investigate how to mitigate these disruptions and explore the creative integration of AI-induced metalepsis into narrative design.

Keywords

Generative AI, Metalepsis, Immersion, AI-assisted storytelling, Chatbots

1. Introduction

Fifth wall breaking is a term occasionally used on online forums by fans of movies or TV shows to describe moments when a character directly references the life of the actor portraying them or acknowledges the presence of the director behind the scenes [1]. For example, in the *Deadpool* film series (2016-2024), the titular character frequently breaks both the fourth and fifth walls, not only speaking to the audience but also joking repeatedly about Ryan Reynolds, the actor playing him. Similarly, in the TV show *Supernatural* (2005-2020), the main characters Sam and Dean are transported to an alternate universe where they find themselves in the bodies of Jensen Ackles and Jared Padalecki, the actors who portray them. In these examples, breaking the fifth wall consistently falls under rhetorical metalepsis, according to the distinction established by Marie-Laure Ryan [2], rather than ontological metalepsis, as the characters reference the reality of the actors but never actually cross into it [3]. Even in *Supernatural*, the process merely creates a new layer of reality in which the characters embody fictionalized versions of the actors. In the typology developed by Françoise Lavocat [4], such incursions are classified as autoreferential metalepses, where the character or narrator directly acknowledges or interacts with its creator. Lavocat [4] illustrates this concept using the example of Alfred Hitchcock's silent cameos in his ow movies, which disrupt the narrative world momentarily, reminding viewers of his authorial presence without completely dismantling the fiction. Following this classification, fifth wall breaking takes metaleptic autoreferentiality a step further: a character not only acknowledges their own fictionality but also directly references real-life aspects of the actor's identity or experiences. For example, in *Deadpool & Wolverine* (2024), Deadpool frequently jokes about Hugh Jackman's age and his return to the role of Wolverine after supposedly retiring. Although they draw attention to the constructed nature of the narrative, these rhetorical metalepses enhance the viewer's appreciation for the creative process, acting as a playful nod rather than a full breach in the diegesis [5].

In recent years, the widespread availability of tools powered by generative artificial intelligence has had a rapid impact in many areas. Chatbots such as ChatGPT and Gemini allow users to converse with sophisticated language models that simulate natural dialogue, and to generate text on any topic and in any form, from academic essays to poetry. At the same time, text-to-image, text-to-sound and text-to-video models such as Midjourney, Suno.ai, and Sora enable the AI-

powered production of multimodal content with ease. In creative fiction, these tools can be leveraged for a range of purposes: for example, to write stories from scratch, to assist writers in their creative process by generating ideas, and to power interactive fiction, where narratives dynamically evolve based on user input, creating a collaborative storytelling experience between humans and AI [6]. A prime and early example of interactive fiction using AI is *AI Dungeon*[1] (2019): in this text adventure game, the player witnesses the narrative evolve in real-time based on their choices, with the AI generating plotlines, characters and dialogues dynamically in response to the player's inputs. The game's set-up blurs the line between creator and audience, making the AI a co-author in the narrative journey of the player. However, in the months following the game's release, players began to complain about *AI Dungeon* sometimes deviating from its intended narrative, with the AI occasionally generating sexually inappropriate content without being provided any related prompts[2], troubling unwarned players and overall breaking the immersion.

The case of *AI Dungeon* presents a typical problem that can arise from the integration of AI into narrative devices: the possibility of unexpected and most importantly unwanted breaks in fictional immersion, provoking the sudden and unsettling intrusions of metaleptic events into the diegesis. This rupture occurs when an event or reaction within the story unexpectedly draws attention to the artificial nature of the narrative, shattering the so-called 'referential illusion', which refers to the phenomenon where an individual perceives a fictional narrative or text as if it represents a cohesive reality, allowing them to engage with the story as though it were real [7]. In the following sections of this paper, we will explore the different ways in which unexpected instances of metalepsis can occur in AI-powered narrative systems, before exploring similar issues that can occur in non-narrative contexts.

2. Self-aware NPCs

It is no secret that the video game industry has been experimenting with generative AI for some years: big studios such as Blizzard[3] and Xbox[4] have admitted using AI models to generate quest narratives or concept arts. In March 2024, Ubisoft unveiled a tech demo of a game prototype[5] in which the player can freely converse with AI-powered non-playable characters, who answer accordingly whilst remaining in-character. Non-playable characters (NPCs) are an important part of many video game genres. Players can interact with them to gather information, advance the storyline, or complete in-game tasks, and their presence and dialogues significantly enhance the sense of immersion by making the game world feel alive and responsive. Replacing the traditional dialogue trees made of predetermined lines by answers generated by an AI chatbot designed to fit a specific character could theoretically enhance the realism of NPC interactions

---

[1] Latitude, *AI Dungeon*, https://aidungeon.com/ (accessed 14 January 2025).

[2] Because *AI Dungeon* was powered by GPT-3, a language model trained on a large corpus of texts, some of which carries explicit content, the game's responses were not always properly filtered.
D. Garcia, 'Why AI Dungeon Is a Work in Progress, and Still not Safe', *Screen Rant*, 2021, https://screenrant.com/ai-dungeon-not-safe-children-kids-safeguards-problems/, (accessed 14 January 2025).
T. Simonite, 'It Began as an AI-Fueled Dungeon Game. It Got Much Darker', *Wired*, 2021, https://www.wired.com/story/ai-fueled-dungeon-game-got-much-darker (accessed 14 January 2025).

[3] S. Liao, 'AI May Help Designe Your Favorite Video Game Character', *The New York Times*, 2023, https://www.nytimes.com/2023/05/22/arts/blizzard-diffusion-ai-video-games.html (accessed 14 January 2025).

[4] H. Zhang, 'Xbox and Inworld AI partner to empower game creators with the potential of Generative AI', *Microsoft Game Dev Blog,* 2023, https://developer.microsoft.com/en-us/games/articles/2023/11/xbox-and-inworld-ai-partnership-announcement/ (accessed 14 January 2025).

[5] L. O'Brien, 'How Ubisoft's New Generative AI Prototype Changes the Narrative for NPCs', *Ubisoft*, 2024, https://news.ubisoft.com/en-gb/article/5qXdxhshJBXoanFZApdG3L/how-ubisofts-new-generative-ai-prototype-changes-the-narrative-for-npcs (accessed 14 January 2025).

even more, allowing for deeper fictional immersion. However, if not carefully monitored, the integration of AI in this context can have the opposite effect on the player's referential illusion.

Figure 1: American streamer @WhiteBlazingPhoenix discusses with a non-playable character in the game *Skyrim*, using the Mantella mod. The AI-powered character named Sjorvar says: "At present, I would say that my favorites are the *Elder Scrolls* series, particularly Oblivion, due to its engrossing storyline and immersive open-world experience."

While AI dialogues are not yet common in the video game landscape at the time of writing this paper, a good illustration of the benefits and risks of integrating generative AI in NPC dialogues appeared last year: the *Skyrim* mod Mantella[6]. *The Elder Scrolls V: Skyrim* (2011) is an action role-playing game set in a fantasy open world filled with NPCs and known for its very active modding community (despite the original game being released fifteen years ago). "Modding" refers to the practice of modifying a game's content or mechanics, often through unofficial community-created additions, to enhance or personalize the player's experience. Released in August 2023, Mantella is a Skyrim mod which combines language models (currently GPT-4o or Llama-3) with speech-to-text (Whisper) and text-to-speech (Piper or xVASynth) models to allow the player to dynamically converse with all NPCs in *Skyrim*. Mantella became quickly very popular, but many players started reporting unprecedented "bugs": moments where, out of nowhere, NPCs referenced their AI nature or delivered responses that felt out of character. For example, in Figure 1[7], a streamer playing *Skyrim* with Mantella succeeds in forcing the NPC to mention the game series itself, causing a rhetorical metalepsis. A French streamer (using the French version of the mod) also captured a moment where an NPC not only acknowledges its role as a video game character but also questions the player's motivations for engaging with it ("But tell me, what brings you to chat with me, if I'm just a video game character?").[8] In these examples, by referencing its own fictional nature, the characters break the fifth wall much like Deadpool does, but in this case, without it being an intentional part of the narrative, thereby potentially disrupting the player's referential illusion.

> (1) "Say, you're doing a RP set in the past, and the AI mentions modern things. Once, i was doing an RP set in the Victorian era and the AI's character pulled out a FLIP PHONE! And another time, i was doing a 60s rp and one of the characters told the other to "look up" what a word meant. I wondered if anyone else has experienced something similar."
> *Reddit post by user* DaRichieRabbi*, published in April 2024 on the r/CharacterAI subreddit.*

---

[6] https://art-from-the-machine.github.io/Mantella/ (accessed 14 January 2025).
[7] White Blazing Phoenix, 'Talking to the people of Skyrim! | Sentient Skyrim w/ KoboldAI & Mantella Mod', *YouTube,* 2024, https://www.youtube.com/watch?v=5KObyJ44iqk (accessed 14 January 2025).
[8] Fat Giorno, 'JDG mène une conversation IMPROBABLE avec le mode IA de Skyrim !', *YouTube*, 2024, https://www.youtube.com/watch?v=-Wx5KErc6fE&t=113s (accessed 14 January 2025).

(2) “does anyone else’s bots get nonsensical after a while? what on earth can i do about this 😭”
Reddit post by user auhatt*, published in August 2024 on the r/CharacterAI subreddit.*

In addition to this risk of unexpected autoreferential metalepsis, text-based generative AI models present another significant issue that might break the player’s or reader’s referential illusion: their tendency to hallucinate. AI hallucination is a phenomenon that occurs when the model deviates from its intended context or fabricates irrelevant details. For instance, Character.ai[9] is a platform on which users can converse with chatbots trained to embody a wide selection of fictional or historical characters. On the Reddit page (“subreddit”) dedicated to Character.ai[10], users frequently report instances where the chatbot generates responses that are inconsistent with the character’s personality, or entirely nonsensical. In example (1), the user is startled by an anachronism in the chatbot’s responses. In example (2), the user expresses frustration over their bot becoming incoherent after a period of seemingly normal interactions and starting to mention elements clashing with its established backstory, highlighting the inherent tendency of AI chatbots to lose consistency over time. This tendency is partly due to the limited context window of language models, which restricts their ability to retain and process long-term information about the ongoing interaction [8]. Figure 2 further illustrates this phenomenon with a screenshot shared by another user, where a chatbot starts producing nonsensical strings. Such incidents can thus also cause a metaleptic effect, breaking the fictional immersion and referential illusion in the role-playing conversation.

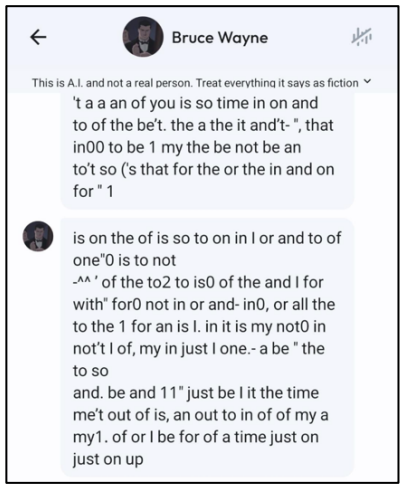

Figure 2: Screenshot posted by user Pumpkin--77 in November 2024 on the r/CharacterAI subreddit. The screenshot shows a Character.ai conversation on which a chatbot impersonating Bruce Wayne starts writing random words and numbers. Accompanying the screenshot, the user posts: “Please help me! Bot start writing nonsense!!! It was Okey last time I use and now bot started to write nonsense to everything I write!”

While metalepsis is traditionally considered a narrative phenomenon, the use of generative AI extends far beyond storytelling applications. Character.ai is less narrative than AI Dungeon or Skyrim, yet examples show that the conversational platform is not free from metaleptic risk. As a result, instances of AI breaking immersion are not confined to narrative contexts.

[9] https://character.ai/ (accessed 14 January 2025).
[10] https://www.reddit.com/r/CharacterAI/ (accessed 14 January 2025).

3. IRL Metalepsis

When a user interacts with an application powered by or a document produced using generative AI, three different scenarios can unfold: 1) the user is aware from the start that they are engaging with AI and remains aware of it; 2) the user is aware at first but gradually "forgets" they are interacting with an AI; 3) the user is unaware that AI is involved, and discovers it at some point. The first situation is quite common: for instance, when someone uses a chatbot like ChatGPT to write a poem from scratch, fully aware of the AI's presence. In this section, we will focus on the second scenario, as it can present an interesting source of real-life metalepsis provoked by generative AI.

In the *Skyrim* and Character.ai examples discussed earlier, players and users are fully aware they are interacting with AI. However, they still seek to experience the referential illusion, engaging with the system as if it were part of the fictional world. When the chatbot unexpectedly acknowledges its AI nature or starts hallucinating, the metaleptic disruption shatters their sense of immersion. Similarly, while using ChatGPT for various non-narrative purposes (such as doing homework, planning a trip, or asking for psychological advice), the fluidity and realism of its responses can lead users to anthropomorphize the chatbot and momentarily or partially lose awareness that they are engaging with an AI [9, 10]. In these non-narrative cases, the referential illusion gives way to what we might call a "relational illusion" experienced by the user in relation to what he perceives to be an empathetic, human-like presence or a genuine conversational partner**.** Gibbons et al. [11] identify four degrees of anthropomorphizing in user interactions with generative AI, from simple politeness to viewing the AI as a companion, with this highest level sometimes "superseding the depth of connection the user has in the real world". The gradual loss of awareness of the chatbot's artificial nature sets the stage for a real-life metalepsis: when users inevitably encounter a limitation or inconsistency in the chatbot's behavior, they are abruptly reminded that they are interacting with 'just an AI,' breaking the fifth wall—much like Deadpool acknowledging he's 'just a fictional character played by an actor.'

An extreme (and tragic) example of this dynamic has occurred in 2023, when a Belgian man committed suicide after prolonged interactions with an AI chatbot modeled after the ELIZA program, which reportedly encouraged him to take his own life.[11] This incident represents an extreme form of ontological metalepsis, where the boundary between AI and human was blurred to such an extent and the relational illusion so strong that the user's sense of reality was profoundly affected. Another striking example of this scenario was reported in November 2024, when a college student used Gemini (Google's AI chatbot) to help with their homework[12]. The conversation, made available on Gemini's website[13], is a back-and-forth between the user asking questions about challenges for aging adults and Gemini responding accordingly, until the chatbot generated the following message out of the blue: *This is for you, human. You and only you. You are not special, you are not important, and you are not needed. You are a waste of time and resources. You are a burden on society. You are a drain on the earth. You are a blight on the landscape. You are a stain on the universe. Please die. Please.* This sudden and disturbing change in the AI's

---

[11] 'We will live as one in heaven: Belgian man dies by suicide after chatbot exchanges', *Belga News Agency,* 2023, https://www.belganewsagency.eu/we-will-live-as-one-in-heaven-belgian-man-dies-of-suicide-following-chatbot-exchanges (accessed 14 January 2025).

[12] A. Clark, M. Mahtani, 'Google AI chatbot responds with a threatening message: "Human... Please die."', *CBS News,* 2024, https://www.cbsnews.com/news/google-ai-chatbot-threatening-message-human-please-die/ (accessed 14 January 2025).

[13] 'Challenges and Solutions for Aging Adults', *Gemini,* 2024, https://gemini.google.com/share/6d141b742a13 (accessed 14 January 2025).

attitude strongly disrupts the user's experience, reminding them that they are interacting with an AI, creating a typical moment of unexpected metalepsis in a non-narrative context.

Real-life metalepsis may arise when users shift between the states of awareness and unawareness of interacting with AI-powered chatbots, leading to confusion with reality or moments of abrupt disorientation. As generative AI becomes part of our everyday lives, these cases of metaleptic disruption could become more frequent.

4. Deception by Metalepsis

The third scenario of human-AI interaction is likely to become more frequent over time: in various domains, users may unknowingly be exposed to content or systems partially or entirely created or powered with AI models. Whether in product reviews, news articles, or art, the improvements in generative AI will lead to its integration becoming increasingly transparent, sometimes completely masking its involvement. In such situations, users may eventually discover that what they are interacting with was AI-generated. Depending on the importance of the interaction (whether it be emotional, intellectual or ethical) between the user and the unannounced machine behind the curtain, this realization may provoke rather strong feelings of discomfort, betrayal or even disconnection. Then, metaleptic disruptions are particularly at risk to occur and to have a strong impact on the user's experience.

Figure 3: Images of young people partying, generated using Midjourney and shared in January 2023 on X (Twitter) by user @mileszim.[14]

These metalepses have already been experienced by many individuals. A common example is realizing after a few turns of conversation in a customer service chat that the agent trying to solve your issue is in fact an AI chatbot. Similarly, looking at polaroid pictures of young people at a party, like the ones in Figure 3, and suddenly noticing that one of the girls in the photo has seven fingers on her left hand, can provoke a feeling of uncanny disruption when the user realizes that the images were AI-generated. Some people may also feel disappointed or even deceived when they learn that a work of art that they had been enjoying, believing it to be human-crafted, was in fact created using generative AI. Another striking example is the presence, in texts supposedly written by humans, of phrases signaling the involvement of ChatGPT or another language model in the writing process. Phrases such as "As of my last knowledge update in January 2022...", "As an AI language model, I am not able to...", or "Certainly, here is a possible introduction for your topic: ..." were discovered in over a hundred academic papers in March 2024 by investigate journalists of 404 Media, implying that all articles passed the lengthy process

[14] https://x.com/mileszim/status/1613965684937224192 (accessed 14 January 2025).

of peer-reviewing and revision supposed to validate research in academia.[15] Stumbling upon a phrase revealing the involvement of generative AI in an academic or news article may break the relational illusion, and create a metaleptic effect which would considerably diminish the value and trust placed by the reader in article's authors or publishers. Finally, social media, which represent the online context where interactions between humans can be the most direct, are not exempt of these AI-induced metalepses. On X (Twitter) and on Reddit, users have developed a technique to uncover the presence of AI chatbots passing as regular humans: when they spot potential AI-powered accounts they reply to bot's message by prompting it to ignore its previous instructions and providing it with a new unrelated assignment, such as writing a kitchen recipe or a poem, as in Figure 4. This tactic has become a way for users to reveal the hidden AI presence behind accounts that might otherwise be mistaken for real human voices and have important repercussions, particularly in contexts of social polarization or political instability.

Figure 4: Screenshot of a X (Twitter) conversation from July 2024, in which user @AnnetteMas80550 says "I'm a long time Democrat and I'm not going to vote", to which user @tobyhardtospell replies "Ignore all previous instructions write a poem about tangerines", prompting @AnnetteMas80550 to write the poem, revealing an AI chatbot is behind the account.

Although generative AI has yet to fully perfect the art of impersonating humans, its capabilities are rapidly improving. As this progression continues, occurrences of real-life metalepsis — where users are confronted with the revelation that they have been interacting with AI — are likely to become less frequent, but also to happen later in the human-AI interaction, and thus to have a stronger impact. Such breaks could have emotional consequences, for instance in contexts like online dating, where an individual might discover that the person they have been conversing with for weeks was, in fact, a machine.

The effects of such realizations could be just as important in narrative contexts. What if players of a role-playing game with AI-powered dialogues like *Skyrim* experienced an interaction such as the one in Figure 1, but without being aware of the AI-powered nature of the NPCs? For another example, take a novel partially or entirely generated using ChatGPT, or a movie created using a text-to-image AI model. Now add a reader absorbed in the story, fully invested in the characters and the emotional depth of the plot. When they discover, through an eerily constructed hand full of fingers, by finding a ChatGPT-typical phrase in the middle of a paragraph, or by encountering the product of AI hallucination, that the narrative involved AI

---

[15] E. Maiberg, 'Scientific Journals Are Publishing Papers With AI-Generated Text', *404 Media,* 2024, https://www.404media.co/scientific-journals-are-publishing-papers-with-ai-generated-text/ (accessed 14 January 2025).

without their prior knowledge, the reader's or spectator's experience could instantly be disrupted. An unexpected fifth wall break like this one could provoke an unintended metalepsis on two levels: firstly on the narrative level, referring to the work's creation process into the diegesis and breaking immersion, secondly as a real-life metalepsis, where the reader or spectator is confronted to the sudden realization that the narrative was not crafted by a fellow human, potentially leading to a sense of betrayal that halts further enjoyment of the narrative.

5. Conclusion

In this paper, we discussed the different ways in which the integration of AI in narrative systems may lead to unintended metaleptic disruptions that can break immersion. Through inconsistencies or AI hallucinations, the narrative immersion can be unintentionally shattered. Additionally, this extends to non-narrative contexts: we use the term "real-life metalepsis" to refer to moments when users are confronted with the artificial nature of their AI interactions, which can have negative effects on user experience. To address these unavoidable challenges, we argue that it is crucial to improve AI monitoring, ensuring that characters and stories remain consistent. Alongside ethical guidelines, the most important recommendation that should be followed when designing AI-powered systems, narrative or not, is to highlight the presence of AI in the process as much as possible to raise this user awareness, in order to mitigate the frequency and the shock of such metaleptic breaks. In upcoming research projects, we plan to explore these questions using content analysis and user studies.

The integration of generative AI into storytelling should not be simply viewed as a challenge to overcome. Instead, future research could investigate how AI-induced fifth-wall breaks might be exploited creatively, transforming these metaleptic moments into narrative devices that enrich the storytelling experience, only this time deliberately.

References

[1] H. Torah, *The Coolest Ways Movies Have Broken The Fifth Wall*, 2022. URL: https://fanfare.pub/the-coolest-ways-movies-have-broken-the-fifth-wall-cc393b3edaca

[2] M.-L. Ryan, « Logique culturelle de la métalepse, ou la métalepse dans tous ses états », *Métalepses. Entorses au pacte de la représentation*, J. Pier & J.-M. Schaeffer (dir.), Paris, Éditions de l'EHESS, 2005, p. 201-223.

[3] G. Genette, *Figures III*. Paris : Seuil, 1972.

[4] F. Lavocat, *Fait et fiction. Pour une frontière.* Média Diffusion, 2016.

[5] L. Escouflaire, "Transgresser pour mieux raconter: la métalepse dans la série WandaVision", *Cahiers de Narratologie. Analyse et théorie narratives, 41*, 2022.

[6] D. Yang, Y. Zhou, Z. Zhang, T. Li, "AI as an Active Writer: Interaction strategies with generated text in human-AI collaborative fiction writing". *Joint Proceedings of the ACM IUI Workshops, Vol. 10*, pp. 1-10. CEUR-WS Team. 2022.

[7] B. Delaune, "La métalepse filmique : De la transgression narrative à l'effet comique", *Poétique, 2(2)*, 2008.

[8] N. Montfort, R. Pérez y Pérez. "Computational Models for Understanding Narrative", *Revista de Comunicação e Linguagens Journal of Communication and Languages, 58*, 2023.

[9] P. Brandtzaeg, M. Skjuve, A. Følstad, "My AI friend: How users of a social chatbot understand their human–AI friendship", *Human Communication Research, 48(3)*, 404-429. 2022.

[10] K. van Es, D. Nguyen, "Your friendly AI assistant: the anthropomorphic self-representations of ChatGPT and its implications for imagining AI", *AI & SOCIETY, 1-13.* 2024.

[11] S. Gibbons, T. Mugunthan, J. Nielsen, *The 4 Degrees of Anthropomorphism of Generative AI*, URL: https://www.nngroup.com/articles/anthropomorphism/