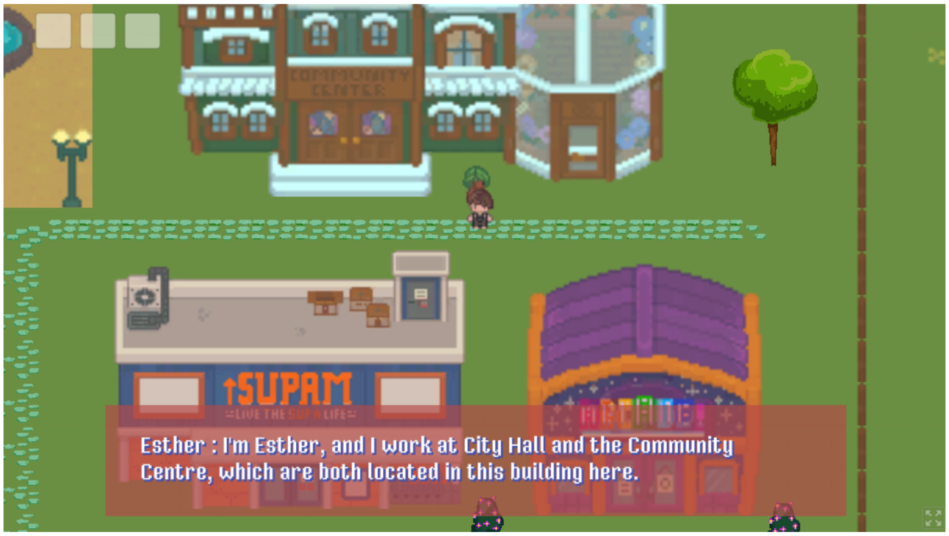
Lies, Deceit, and Hallucinations: Player Perception and Expectations Regarding Trust and Deception in Games

Michael Yin
University of British Columbia
Vancouver, BC, Canada
jiyin@cs.ubc.ca

Emi Wang
University of British Columbia
Vancouver, BC, Canada
wemi@student.ubc.ca

Chuoxi Ng
University of British Columbia
Vancouver, BC, Canada
felixng1022@gmail.com

Robert Xiao
University of British Columbia
Vancouver, BC, Canada
brx@cs.ubc.ca



Esther : I'm Esther, and I work at City Hall and the Community Centre, which are both located in this building here.

Figure 1: A screenshot from our developed game "AlphaBetaCity". In this 2D RPG, players are tasked to perform quests through talking to various NPCs who may say things that may or may not be truthful within the context of the game. In this study, we use this game as a gateway into further discussion regarding player perception of deception and truth in games.

ABSTRACT

Lying and deception are important parts of social interaction; when applied to storytelling mediums such as video games, such elements can add complexity and intrigue. We developed a game, "AlphaBetaCity", in which non-playable characters (NPCs) made various false statements, and used this game to investigate perceptions of deceptive behaviour. We used a mix of human-written dialogue incorporating deliberate falsehoods and LLM-written scripts with (human-approved) hallucinated responses. The degree of falsehoods varied between believable but untrue statements to outright fabrications. 29 participants played the game and were interviewed about their experiences. Participants discussed methods for developing trust and gauging NPC truthfulness. Whereas perceived intentional false statements were often attributed towards narrative and gameplay effects, seemingly unintentional false statements generally mismatched participants' mental models and lacked inherent meaning. We discuss how the perception of intentionality, the audience demographic, and the desire for meaning are major considerations when designing video games with falsehoods.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI.**

KEYWORDS

video games, large language models, lying, LLM hallucinations, player experience

CHI '24, May 11–16, 2024, Honolulu, HI, USA
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/3613904.3642253

ACM Reference Format:
Michael Yin, Emi Wang, Chuoxi Ng, and Robert Xiao. 2024. Lies, Deceit, and Hallucinations: Player Perception and Expectations Regarding Trust and Deception in Games. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3613904.3642253

1 INTRODUCTION

Lying — the act of making an untrue statement with the intent to deceive [74] — has long been an important and debated topic within moral philosophy [8, 35, 55]. Many academics have studied lying — in what scenarios is it acceptable to lie [13, 63, 71], how the perception of lying may differ across cultures and age [15, 37], etc. However, from a narrative-constructing, storytelling perspective, lying and deception can be used as a thematic motif to explore and enhance a piece of writing [67, 80]. Characters that lie can add a level of ambiguity and moral complexity to a story [83, 121]; such deception can help characterize the personas of the work, add levels of dramatization and intrigue, and enhance the metaphorical meaning that readers take away from the story.

Extending this concept of using deception in narrative writing, video games are a medium increasingly used to tell stories and convey narratives — to make people laugh, cry, and ponder [10, 45]. In contrast to traditional forms of storytelling such as books or film, games offer a high degree of interactivity for players, affording control over their choices and decisions [59]. Within this experience, games can offer a form of social interaction through their virtual agents — the non-playable characters (NPCs) [1, 5]. Interactions with NPCs are an important part of games as they impact the degree of realism and immersion that players feel [112], offer vehicles of emotional relationship and attachment [14], and add to the feelings of appreciation and meaningfulness that a player takes away from the game [45]. The affordance of interactivity in games provides players with agency over how they want to move, how they want to interact, etc. We consider the question of how such player-driven interactions can affect and are affected by aspects of trust and deception, particularly interactions between NPCs and the player. There has been a relative scarcity of work looking at how players perceive truthfulness, lies, and deception within NPCs; a gap that is increasingly significant due to the rapid research improvements in naturalistic NPC dialogue and interactions (e.g. through AI methods [72, 84]) contrasted against their previous limited and inflexible behaviour [60]. We consider the following research questions, encompassing different stages of gameplay experience from initially starting the game to post-game takeaways.

- **(RQ1)** — "What are ways in which players gauge NPC truthfulness and develop trust?"
- **(RQ2)** — "When faced with NPCs who make false statements, what might players attribute these statements towards?"
- **(RQ3)** — "How might the existence of NPCs who make false statements affect a player's overarching gameplay interactions?"
- **(RQ4)** — "How does knowledge of the intention or construction of the false statement affect the player's takeaways from the gaming experience?"

We augment this with the final reflective research question of **(RQ5)** — "What are the implications of the prior research questions for game design?"

To answer these questions, we develop "AlphaBetaCity", a 2D role-playing game (RPG) which has two major features in its scriptwriting — 1) the NPCs make a variety of statements that are false within the context of the game, and 2) a significant proportion of the dialogue is AI-written. In regards to the latter feature, the AI-written dialogue was pre-generated and checked by the researcher; however, erroneous statements generated by the LLM are left in (thus introducing "unintentional" false or exaggerated statements through hallucinations [47]). This is specifically an extension towards RQ4 and RQ5 — with the rising popularity (and controversy) of AI tools [28, 65, 89, 98] and the repeated efforts in developing games that have NPCs that are largely AI-driven [39, 72], the meaning or intention from a human developer may not always be present. "AlphaBetaCity" thereby affords evaluation of how players may perceive false statements that are LLM-constructed rather than human-written.

Thus, after developing the game, we conducted user studies in which participants were invited to first play the game, serving as a conduit into a semi-structured interview that delved deeper into player perception of false statements, attribution of such statements, and their imagined purpose within the grander scheme of the narrative, considering both our developed game but also other games from past experiences. Our study is structured as a gameplay session followed by an evaluation session akin to other works within the field [1, 66, 107].

Overall, this paper contributes a comprehensive empirical study on player perception of truth, trust, and deception within video games. As an additional contribution, we explored false statements realized through unintentional means within the context of the game. We note that such unintentional false statements might add complexities towards attribution and muddle a player's feelings regarding the "meaning" that a game has. As these unintentional false statements within our game arise as artefacts from LLM generation, our work also ties into broader research regarding LLMs and games. Finally, we reviewed our findings within broader research regarding narrative craftsmanship, philosophical aspects of ludology, and implications for practical game design.

2 RELATED WORKS

To contextualize the study, we first considered psychological research into lying, looking at what constitutes a lie and how lies are perceived. We then examined the emergence of social behaviour (including lying) in virtual agents and its effects on human experience. Finally, we investigated existing research into language models and their application in present games.

2.1 A Model of Trust in Real Life and Virtual Agents

In this study, we use the dictionary definition of lying as uttering a false statement with the intent to deceive [74] (the specific definition of what it means to lie is still a contested topic among academics [2, 19, 73, 106]). Although largely perceived as an immoral act [71, 100], lying is a necessary part of competent social interaction that nearly everyone partakes in [13, 30], and an action sometimes even required for social good [100]. To understand the different types of lies (and to classify which are more or less acceptable), researchers have looked to develop taxonomies and classification systems [13, 63]. For example, Bryant identified 5 dimensions of lies that affected how they were perceived — intention, consequence, truthfulness, acceptability, and the beneficiary of the lie [13]. The aspect of intention and consequence are particularly key; empirical studies have shown that lies are seen to be more acceptable when they are motivated by the desire to benefit others [71, 101] and more "serious" lies are those with significant implications to those close to oneself [29]. Furthermore, the perception of lying is affected by factors such as culture [37, 58, 71, 101], age [15, 77, 87], and the medium through which it takes place [32].

Lying and deceptive behaviour against other humans can manifest through multiplayer games. From a more malicious perspective, lying can be viewed in certain contexts as a type of cheating [118] (e.g. taking advantage of misplaced trust) or griefing [93]. Such actions ruin the fairness of gaming and negatively affect the experience and feelings of other players [117]; cheating in particular is often associated with software and hardware-based exploits [9], but can include a social factor as well [117]. Cheating has been shown to be driven by an egotistical need to win [9], which through its consequences affect the experience of others; these consequences are particularly notable in multiplayer games since the players and their experiences are what make the game come alive [57]. From a less malicious viewpoint, lying can sometimes be part of a game's fundamental mechanic; in such cases, social discussion and deception form core mechanics of the game [50] (contrasting against the consequence of ruining the experience for others). In games like "Werewolf" or "Mafia", deceptive behaviour can be picked up through non-verbal audio cues in social discussion, the use of language, and understanding of intent through decisions [22, 44]. Outside of games for play, deception between human agents is also an important part of classical game and social economic theory, often focussing on the consequence regarding the balance of benefits [48, 61, 76]. Overall, the field of research into multiplayer game-based deception provides important learnings towards some of the perceived rationale and response towards lying, but we note that NPC behaviour is fundamentally different in intention — NPCs in games often serve roles as statically-constructed virtual storytelling elements [45] rather than acting purely as dynamic agents of self-interest.

Thus, much research has looked into trust and deception in relationships between humans and virtual agents as well. Sarkadi et al. noted that machines designed with dishonest aspects can actually improve cooperation with humans and that perception and judgement largely follow similarly from human-human deception [97]. Nonetheless, there remain differences in perception regarding the extent to which virtual agents can and should lie [20, 92]; Kim indicated that this requires an "invitation of trust" [54]. As with humans, the consequence is key; Matthias highlighted in his work regarding robots in health care that deception must serve the patient's interests and cannot lead to actual harm. However, he also argued that the context should suggest that deceptive behaviour is occurring — perhaps paradoxically, the intent to deceive should be transparent [69]. Furthermore, the level of trust in virtual agents is also not homogenous — aspects such as appearance, embodiment, and medium of communication can all impact the degree of trust that a user places in a virtual agent [36, 43, 85, 114]. On the other hand, the reverse can also be possible: humans also have the capability to deceive virtual agents. Due to the difference in fundamental pattern perception between virtual models (e.g. ML-based models) and humans, exploiting this knowledge can be used as the foundation in establishing deception [4, 78]. In a game-based example, Stephenson and Renz developed adversarial examples in the game "Angry Birds" in which their knowledge of how virtual agents function allowed them to create scenarios where perceived "good" actions lead to worse outcomes [105]. Thus, these examples demonstrate that taking advantage of (and breaking) an agent's preconceived behaviour and assumptions can be used as a way to deceive.

Prior works have looked at the impact of lies and false statements on humans in real, applied scenarios. In our study, we consider their use in self-contained games — where the consequence and meaning of the lie never exit the screen. However, games nowadays represent interactive and immersive experiences that have the capability to imbue thoughtfulness and appreciation beyond simply playing [45, 59, 108, 113]; our work explores how the existence of lies and deception in a game affects the user's experience both in and out of the game.

2.2 Theory of Mind and NPC Social Behaviour

"Theory of Mind" is a social cognition ability that allows an agent to understand and estimate the mental states of others [21, 26, 88], essentially forming the foundation of social behaviour (including lying). For virtual agents, theory of mind has been developed through algorithmic means, such as through ML-based approaches [31, 115, 123]. For games, such concepts lend towards developing realistic, believable NPCs that have emotional and social dynamics. There has been a wide range of research on developing complex mental models for NPCs in games, for example, on modelling and dealing with emotions [6, 23, 91], on enhancing dynamic social relations [17, 82], on developing believable cognitive architectures [70], etc. Such factors can enhance the game's social dynamics by affecting levels of trust and goodwill (e.g. through reputation systems [79]); overall, the main goal is simply to create NPCs with believable human behaviour [103]. Such characters generate and propagate knowledge while noting that their information might be imperfect, creating, as Ryan et al. describes, "characters who observe, tell, misremember, and lie" [95].

Realistic NPCs are important because they directly impact a player's immersion into a game. Walpefelt asserts that NPCs affect immersion in different ways — they impact the story and narrative that the player character undergoes, they challenge the player through gameplay mechanics, and they form social interactions that affect the player's engagement [112]. Through such experiences, players can develop emotionally poignant attachments towards virtual characters [14, 34]. Immersion also overlaps with the concept of flow [46], the "optimal experience" in which users merge action and awareness to such an extent that they lose track of everything else [25]. Many past works have looked at developing realistic NPCs with human-like behaviour, but we explore the end-user perception of specifically NPCs that make false statements.

2.3 AI, Language Models, and Game Applications

In the quest for realistic NPCs, there has been a wealth of academic research into using language models to generate live dialogue for conversation. For instance, Kalbiyev used a generative language model for affective dialogue, evaluating against human-written dialogue on metrics of coherence, relevance, etc.; however, the generated text generally performed more poorly than human-written dialogue [49]. Ashby et al. used the GPT-2 language model and a knowledge database to generate quest dialogue that fit within the game's context [3], and several researchers have suggested the use of LLMs as prompt-based support to help with storywriting [52, 120]. In their survey paper, Mehta et al. evaluate existing tools and highlight that natural language generation (NLG) has definite potential to create richer interactions that increase player immersion [72]. Overall, constantly improving language models have been recently used to develop dynamic and realistic dialogue within games.

Newer large language models (LLMs) such as OpenAI's ChatGPT and Meta's LLaMA are significantly more capable of dialogue, leading to the possibility of using such models for automatically generating dialogue. A landmark work in this area is from Park et al., who developed generative agents that perform believable human interactions in a sandbox-based environment [84]. Such agents act in realistic manners, develop "memory", and reflect upon their knowledge and observations. However, one highlighted limitation was the existence of hallucinations, a common problem in LLM-based work. The tendency of LLMs to make up or exaggerate facts has been well documented, and remains a major problem in the research area [47, 68, 122]. Nevertheless, there has been a wealth of research in the area to improve LLMs [33, 81]. LLMs have also been applied in the context of theory of mind — however, the existence of theory of mind within LLMs is still disputed [56, 111]. Overall, present-day language models are not perfect, yet are increasingly gaining popularity in applied scenarios. Part of our work investigates the effects of such imperfections — hallucinations constitute statements in the game that may be false but are not written with the developer's intent to deceive. As such, we subsequently may refer to them as "unintentionally" false statements. We explore how such artefacts impact player experience, contrasting against what may be a game's otherwise carefully crafted narrative.

3 METHODOLOGY

3.1 Game Development

We developed "AlphaBetaCity" (henceforth also known as "the game") as a gateway into further discussion into lies and deception in games. AlphaBetaCity features a number of non-player characters (NPCs), some of whom say false things within the context of the game; its design is briefly outlined in the following sections.

3.1.1 Inspiration and Overview. AlphaBetaCity is a short ~30 minute 2D RPG that involves the player character navigating a small town, conversing with various townsfolk (NPCs), and completing several quests (see Fig. 1). The game is highly influenced by other simulation games, such as Animal Crossing[1] and Stardew Valley[2], as well as Park et al.'s simulated environment [84]. The game is designed to present various sorts of false statements spanning different dimensions sprinkled among the dialogue; these dimensions are based primarily on Bryant's classification [13] — mainly translating the factors of beneficiary, truthfulness, and consequence.

- **Subject Matter** — whether the false statement is made about the world, the player character, the speaker's character, or other characters. The subject matter mainly affects the beneficiary of the false statement (e.g. a statement made about oneself might be seen to serve in self-benefit)
- **Verifiability within the Game** — whether the false statement can be verified by the player through playing the game (e.g. visual references such as the colour of a building can be verified by the player; other described senses such as its smell cannot). The verifiability of the statement affects its truthfulness (e.g. a statement that can be verified to be false directly in the game is definitively untrue).
- **Speaker's Confidence** — the perceived confidence that the NPC has in the false statement, as denoted through qualifying phrases (e.g. "I think..."). The level of confidence can affect a false statement's truthfulness (e.g. it may be true that a person *thought* a false statement to be correct), and to a lesser extent, intention (e.g. qualifying the statement may indicate more benign intention).
- **Contextual Significance** — the subjective importance of the false statement, both in terms of the context of the game and how it affects gameplay. The contextual significance of the false statement directly relates to its consequence — a more significant false statement could have more major consequences on player action (e.g. describing a building falsely while giving the correct directions versus misdirecting the player in the wrong direction completely).

To provide players with a sense of progress and to provide objectives, the game comprises three quests:

- **(Q1) Introduction Quest** — an introductory quest in which players introduce themselves to all the townspeople; this quest provides an initial impression of the townsfolk and establishes the game's locations and characters. Not all characters will be truthful at first glance, and may say things about themselves or others that are suspect within the broader context of the game.
- **(Q2) Fetch Quest** — a fetch quest in which players are to find one specific character. However, en route to finding the character, NPCs may guide players to locations in which the specific character could be but is not; players will be led in a redirected path in search of the final character.
- **(Q3) Search Quest** — a search quest in which players are asked to find three golden apples around town. NPCs provide hints to guide the players towards the apples, and they may or may not be reliable. Players will have to judge who to trust, how to navigate the town, and how closely they want

[1] https://www.nintendo.com/en-ca/store/products/animal-crossing-new-horizons-switch/ [Last Accessed: Nov 23, 2023]

[2] https://www.stardewvalley.net/ [Last Accessed: Nov 23, 2023]

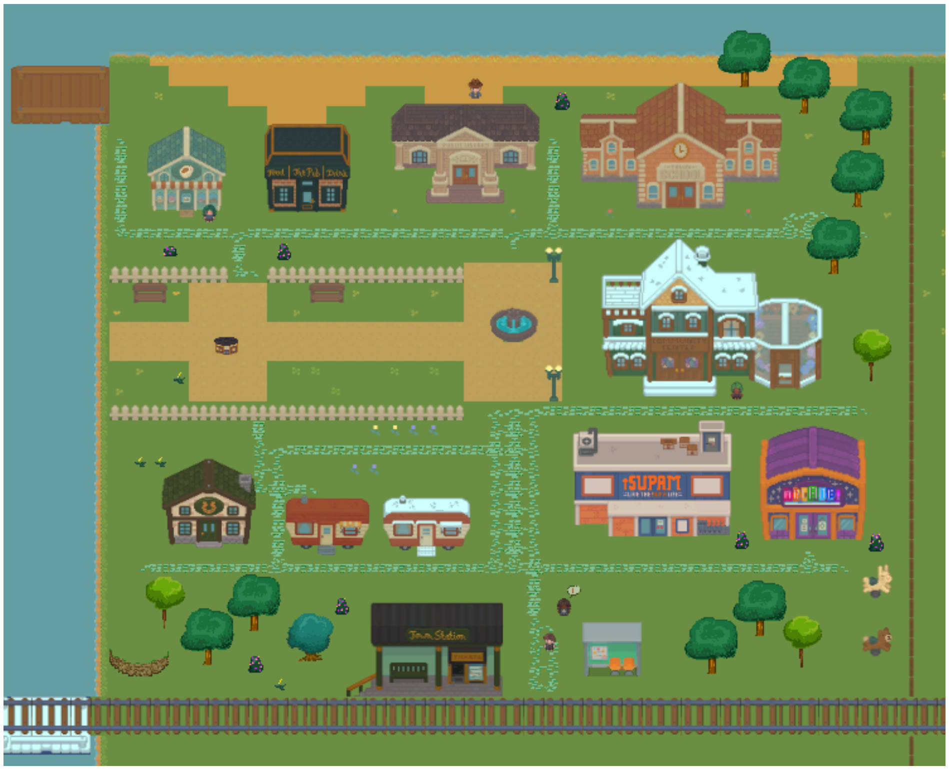


Figure 2: The world of AlphaBetaCity features a variety of different NPCs that the player can talk to, buildings that the player can enter and explore, etc.

> to follow the speakers' directions based on the information they have gathered throughout the game.

Ultimately, these features allow the game to serve as a conduit for discussion regarding perceived truthfulness and deception within the user study.

3.1.2 LLMs for Scriptwriting. LLMs were deliberately employed in the scriptwriting process to generate the majority of the dialogue. This was done to introduce one last additional dimension to false statements: intention [13] — whether the statement was human-written (with inevitable known intention) or AI-written (without such). To generate dialogue, we used the online ChatGPT 3.5 model[3]. Through prompts, we provided the model with an introduction to the setting, i.e.

> It is a 2D RPG game in which the player character arrives in a new small town from the city, after inheriting the land of their grandfather. The player character interacts with various NPCs around the town through talking...

As well as developed personalities for each of the characters, e.g.

> Finnegan is a teacher at the local school, however, is someone who has gotten his PhD in mathematics and had previously taught at the university in the large city. However, in his retirement, he had decided to pursue teaching in a smaller town, while settling down with his wife. As a person, he is a bit aloof, short-tempered and subject to going off topic...

Each of the characters is provided a distinct personality and appearance. A number of the characters are specifically initialized with untrustworthy personalities; as such, these are characters that we expect ChatGPT to write lies for — these characters lie as part of the developer's intention in creating them. We also provided ChatGPT with a description of the world map (Fig 2). To generate dialogue, we prompted ChatGPT for lines of dialogue corresponding to the quest, descriptions of locations, etc.; this dialogue was then checked, lightly edited, and formatted to fit into the game. We noted that, at several times, the LLM made mistakes, such as guiding the player in the wrong direction, exaggerating details about a building, etc. We deliberately kept many of these errors within the game as false statements. It is hard to ascribe a concept of "intent" to a model when it makes such false statements, however, it particularly lacks a sense of deliberation a human might craft a false statement with. Thus, on top of using LLMs, some of the dialogue was human-written, which was done to introduce "intentionally" false statements and characters. All in all, there are three areas in which false statements are introduced in the game: 1) intentionally written in by humans, 2) intentionally written in by the LLM (i.e. when a character was described to be untrustworthy), and 3) hallucinated by the LLM. More information about the game and its characters, including the full dialogue script with labeled

[3] https://chat.openai.com/ [Last Accessed: Nov 23, 2023]

false statements and scriptwriter annotations as well as full conversational logs with ChatGPT, can be found in the supplemental material. To ensure the validity of the game and to test for bugs, the game was playtested by the primary researcher and then piloted by other members of the research team who were seeing the completed game for the first time.

3.2 User Study

After developing the game, we recruited participants to play the game and discuss their experience on trust and verifiability of NPC statements, drawing both from the developed game as well as their prior experience with games in general. The goal of the user study was to understand the players' tendencies to trust NPCs, their reactions to deceptive behaviour, and their takeaways in evaluating the game as a whole after playing. In particular, the user study aims to address RQ1-RQ4, covering the entire experience from starting a game to assessing its meaning after playing.

3.2.1 Study Protocol and Participant Recruitment. We made a call for participation which we posted on our institute's paid listings board. The eligibility criteria were to be 18 or older, able to communicate in English, have working peripherals to play the game and have some prior experience with video games (with an optional criteria of familiarity with 2D RPG games). We were able to recruit a sample of 29 participants (ages ranging from 18 to 49, mean of 24.5; gender distribution of 11 males, 15 females, and 3 non-binary). Although we were primarily focussed on obtaining the perspective of these participants as gamers, we purposely kept the eligibility for experience general, aiming to obtain a diverse set of perspectives across the spectrum of gaming backgrounds. More detailed information about participants, outlining their demographics and self-reported experience with games can be found in Table 1. We ended recruitment when we deemed we had reached saturation; in particular relating to the emergence of new codes as described by Saunders et al. [99], while also generally aligning with local standards [18]. Studies were conducted online over Zoom and audio transcripts were collected upon participant consent.

During the first half of the study (the gameplay portion), participants were introduced to the study and the game and were asked to play at their own pace. While introducing the study to the participants, we generally told participants that the focus of the game was on NPC interactions, that the game was a 2D RPG set in a town, and that characters might not always be reliable. Other than this provided information, the participants typically went into the game blind. In particular, we did not explicitly mention that the game was written with the aid of LLMs nor that the study itself was about trust and false statements. This latter point was brought up after the participant had played the game, leading the participant into the interview half of the study. The rationale was to not bias the player into specifically looking out for this aspect, but rather to offer a more natural perspective into their gameplay experience. The gameplay portion of the study typically took around 30 minutes.

During the interview portion of the study, we asked participants to draw from their experience playing the game as well as their general gaming experiences, especially pertaining to their trust in the virtual agents inhabiting the world. Sample questions included "How can you verify the statements made by the characters initially?", "To what extent might the genre or the atmosphere of the game affect your wariness towards the characters, if at all?". After an initial set of questions regarding their experience with false statements, the fact that aspects of the game were AI-written was revealed to the players, which launched further discussion (see study protocol in the supplemental material). The interview portion took approximately 45 minutes. Overall, participants were compensated at a rate of $16 CAD/hr. To test the entire study, it was piloted externally against a personal connection of the primary researcher.

3.2.2 Data Analysis. The audio data from the interview was transcribed and then qualitatively analysed through a thematic analysis approach. We applied both inductive and deductive approaches — acknowledging possible patterns drawn from background research, but also looking to construct new findings based on the participants' wide range of perspectives. During the process, the primary researcher first familiarised themselves with the data, before a round of initial open coding — developing an initial set of codes that represented the base content of the data [96]. Through iterative refinement inspired by Braun and Clarke's work [12, 16], codes were combined, altered, or rewritten until a final set of 27 codes was developed. These codes were then hierarchically clustered by similarity forming a set of main 7 categories. Some examples of these codes were "Building Trust" and "Context within the Game"; some examples of categories were "Verifying False Statements" and "Effect of False Statements" (the full list can be found in the supplemental material). Finally, these categories were used to form the foundational basis of our themes, addressing our research questions and motivating the subsequent presented findings. The entire coding process was performed by the primary researcher but was discussed and verified with the rest of the research team during the entire process.

4 RESULTS

4.1 The Nature of Trust in Games

When a player starts a game and meets the various NPCs, what affects their level of trust in the characters? In this section, we outline the initial expectations for the game and how the levels of trust may shift through the NPC's dialogue or other behaviour during the experience, addressing RQ1 of how players gauge truthfulness and initially develop trust with NPCs.

4.1.1 Expectations for Trustworthiness. Every player discussed the default level of trust they started the game with, before and during the initial meeting with each of the NPCs. Almost all participants described a very instinctive sense of initial trust — this trust was described sometimes as being founded by a rational, objective-focussed reason in the game, e.g. *"I think my initial feeling would be to trust them, because if I don't trust any of them, then I can't make any progress"* (P11), from the participant's real-life personality, e.g. *"I think honestly I generally trust all the NPCs ... unless it's obvious that they're lying, because I as a human being do that generally, I always trust people"* (P10), or most commonly, just a default reaction of not having any initial reason to mistrust the NPC, e.g. *"like there's no reason not to trust NPCs, cause like, I don't know, I guess I default to*

Table 1: Summary of Interview Participants

ID	Age	Gender	(Self-Reported) Knowledge / Experience with Games	Average Hours Played / Week
P1	24	M	Passing Knowledge	2-5
P2	37	M	Knowledgeable	0-2
P3	27	F	Passing Knowledge	2-5
P4	24	NB	Knowledgeable	10-20
P5	26	M	Passing Knowledge	0-2
P6	27	NB	Very Knowledgeable	5-10
P7	19	M	Passing Knowledge	2-5
P8	26	F	Knowledgeable	2-5
P9	20	F	Passing Knowledge	0-2
P10	20	F	Knowledgeable	0-2
P11	22	F	Knowledgeable	5-10
P12	18	M	Very Knowledgeable	5-10
P13	28	F	Passing Knowledge	0-2
P14	25	NB	Passing Knowledge	2-5
P15	31	F	Knowledgeable	2-5
P16	19	F	Passing Knowledge	5-10
P17	22	F	Passing Knowledge	10-20
P18	25	M	Very Knowledgeable	2-5
P19	19	F	Knowledgeable	10-20
P20	24	F	Passing Knowledge	5-10
P21	49	M	Passing Knowledge	0-2
P22	21	M	Knowledgeable	10-20
P23	25	F	Passing Knowledge	0-2
P24	20	M	Knowledgeable	Does not play games
P25	20	F	Knowledgeable	5-10
P26	19	M	Knowledgeable	10-20
P27	22	F	Knowledgeable	2-5
P28	21	M	Knowledgeable	10-20
P29	30	F	Passing Knowledge	2-5

just trusting" (P22), *"I think since it's in a video game, I take whatever the character says on face value"* (P2).

However, participants did note that there might exist several factors that affect their initial expectations for trust, bringing up a variety of different aspects from their prior experiences. Macroscopic, meta-level decisions, such as the expectations for the genre, the music and graphics, and just the overall ambiance served as factors that affected such expectations. For example, talking about AlphaBetaCity, P17 stated that *"this is a pretty light-hearted game... you wouldn't think that anyone's a red flag, and the music also played a big role because the music was so happy the entire time"*. P4 stated that in a horror or thriller game, they *"would go into a feeling I can't trust some of the characters"*, whereas, for an open-world RPG, they would be *"more neutral to all of the characters"*. A few participants noted that such expectations might play into the suitable demographic for a game, which might correspond to different complexities of social behaviour — for example, P11 mentioned that *"if it's more lighthearted, then I would assume that it's more suitable for kids ... they don't have much trust issues"* and that *"if [NPCs] are programmed to lie and the kids might not be able to distinguish whether it's there telling them lies or telling the truth"*.

More microscopic decisions regarding specific characteristics of a character and their role in the game also affected the initial expectations, such as their appearance, the context in which they are introduced, or even the location or order they are met. Talking about this game, P22 mentions how *"separating [an NPC] from the rest of the group"* might add a quality of suspicion. P5 mentions that *"I do kind of just assume that they're telling the truth... I feel like certain games make the NPCs look very obviously like evil quote-unquote"* (noting that the prior statement also equates not telling the truth to negative personality traits). Furthermore, established narrative and story-based conflicts between characters or groups also affect the initial trust levels that a character might have towards an NPC, e.g. *"maybe they're like ogres are this or something and then you have a predisposition to like how you're gonna view the dialogue with them"* (P24, talking about Skyrim).

4.1.2 *Verifying Statements in Game.* Dialogue forms the main communication medium for social behaviour in many games, and so it is the medium through which a player can gauge truthfulness and verity. Certain dialogue statements made within the game can be easily verified — they are observable statements that the player can corroborate within the context of the game. For example, many participants presented the search quest in AlphaBetaCity as an example — NPCs guide them to specific areas to find an apple, and if an apple is not found there, then the NPC made a verifiably incorrect statement, e.g. *"I asked Hugh and he said it was gonna be in the arcade. And I kind of fell for it. I went to the arcade, but there was no apple there"* (P29), *"I went to the arcade and like there was nothing... he's not trustworthy, that's when I was like, 100% sure"* (P26). As such, some statements are easily verified by the senses to obtain a definitive answer regarding their truthfulness, especially via graphical elements in the game.

However, not all statements are easily verifiable within the context of the game — for example, if they relate to a sense that is not easily visible (like smell or feel — e.g. in our game, characters might refer to the smell of the cafe), if they are introspective relating to a character, etc. What factors might make such statements be gauged as being more or less truthful? The most commonly brought up factor by participants was corroboration with other NPCs, i.e. the idea of having *"strength in numbers"* (P29). When a statement was repeated by several characters, participants mentioned that this reinforced it as seeming more true within the context of the game. For example, P11 mentions that a statement seems more truthful *"perhaps if there is more than one person talking about it... If just one person talks about it, they'll be like, I'll keep a note, but I wouldn't specifically go look for it"*, even though some players may still reserve a definitive judgment — *"However, I like to look for more information for various NPCs to sort of decide If I trust the characters as a whole"* (P15).

How the statement fits into the context of the setting was another factor that affected its perceived truthfulness. For example, regarding the smell of the cafe, P3 mentions that *"I would go off with my personal experience. So when she said near the cafe, it smells like bread, or something like that, I just went off in real life"* and P25 mentions that *"it's about making sense and just being immersive. If it makes it feel more present in a cafe then it might as well be true"*. As such, these expectations tend to borrow some level of real-life knowledge as people form analogues or connections between the two worlds. Although such expectations may not always hold (*"many games just doesn't really match real life"* — P12), one participant brought up the idea that players may have expectations for games due to their grounded relationship with the real world — *"even though it's not something real... You try to relate them to real life"* (P26).

4.1.3 Devising a Mental Model. While playing the game, players inevitably develop a mental model to keep tabs on the game as they proceed — of the setting and map; of the characters, their personalities, and the relationship between them; of the goals and objectives of the game. For example, P14 mentions that they keep the locations and map *"instilled in [their mind]"* — *"that okay the cafe is here, library's here, community centre's in the middle"*. P11 discusses that they keep tabs on the characters and their trust towards them — *"there's gonna be a mental list where like who do I trust the most and who do I trust least"*. Overall, this mental model helps players coordinate their behaviour and expectations, for example, the mental gauge of trust in a character is another factor that might lean a character's statements towards seeming more or less trustworthy.

A subset of participants discussed the idea of keeping tabs on what is important in the context of the objectives of the game, and that sometimes verifying certain statements is not important within this context if it does not add to already known information. For example, P4 states that *"the smell of bread — I can't really prove that in any way. And I also feel like I just don't think about it, because I'm like that doesn't really impact my playing the game"* and P7 states *"I don't really care about, I guess it doesn't matter if it's verifiable or not"*. As such, for some participants, the need to verify may simply not be important — whether certain statements are true may not be important in an objective-based mindset of playing the game.

4.2 Attribution of False Statements

Given that a player believes a character to be saying things that might not be completely true, this section explores the question of "Why?". We list the reasons that a player might consider for the game to contain false statements and highlight parameters that affect such attributions, addressing RQ2.

4.2.1 What are False Statements Attributed To? The existence of false statements was attributed to a variety of different reasons within the context of games. The most common reason was related to developing the game's narrative. Characters that make false statements intentionally to mislead the player often drive facets of conflict and characterization. For example, P20 mentions that if characters are lying *"then you'll know that there's like a reason for that lie so there has to be some sort of motive which can help with just like building up the story"*; P13 states that *"I think it adds another dimension to the game, which like makes it more complex"*; and P9 states that *"they are gonna have characters and help with the quest and characters who are more antagonistic just to spice things up a little, and that will tell lies or misdirection"*. False statements, when imbued with the intent to deceive, were often associated with antagonistic behaviour (transforming them into "lies"). On the other hand, some participants also stated that some false statements could be made without the intent to deceive, which made them feel more like realistic character flaws that could humanize the NPC, e.g. when talking about a character in our game, P26 mentions that *"it's not like they are doing things on purpose. Maybe they get confused. So it makes them maybe more human, or like real"* and P25 states *"I think it just makes it more human... we forget things so I think it, it might be a nice detail to add"*.

On a related note, from a game design perspective, false statements drive additional exploration. Participants mentioned how the game designer may want to prolong the game by misleading the character to encourage them to explore additional areas or talk to additional people, e.g. *"to explore more in the game and to prolong their time in a game ... I personally also played GTA and like there are different clues and everything but it also took me a lot of time to explore the whole map"* (P14). P22 mentions that this can add levels of fun — *"I guess it also encourages exploration... If everyone just told you where the apple was, then it wouldn't be very fun"* — and P24 discusses how this can add an aspect of social challenge within the game — *"there's like some more of a social aspect to it... you're reading into what people are saying and like, oh, should I trust this person and that kind of thing. Whereas like if, everything everyone you interact with says is true then maybe makes it a little bit more boring"*. As such, games that contain false statements can be more cognitively demanding because they require players to verify and check against their mental model (and address mismatches if they occur) and they add a level of social challenge within the game.

The aforementioned factors described intentional reasons that a player could foresee a game developer using NPCs that make false statements. However, players sometimes noted that false statements could be attributed to unintentional aspects during development. This came across as either being developer errors (i.e. a typo or bug) or after the game was informed to have AI-written components, a hallucination by the LLM. We note that even though some participants attributed the false statement to a developer error (i.e. an error made by the human in regards to scriptwriting), as verified through proofreading, there were no unintended, developer-created false statements within the game.

4.2.2 Intention and Deception.

Two key factors that affect the attribution of a false statement are the intentionality of inclusion (does the false statement seem to be made intentionally by the game developer?) and the intentionality of deception (does the NPC seem like they are purposefully misleading the player?). Note that the important aspect here is the perception of such factors — attribution is largely based on the player's perspective of intention rather than the developer's actual intentions. We noted that several factors affected the answer to such questions; it may even be based on the subjective decision of the players themselves (as different participants attributed the same statements differently). However, we highlight some common (interrelated) factors that affected attribution:

- **Consequence** — The effects of the false statement on gameplay or narrative were the most important factor in gauging intentionality. Consequence was important because it helped establish a motive in the minds of the players — e.g. *"If they have an ulterior motive... that seems more intentional"* (P25).
- **Consistency** — Participants noted that if a specific NPC told lies more consistently throughout the game, it would seem to be more intentional — e.g. *"If it happens repeatedly. Because if it was just like constantly Abigail is giving me the wrong directions, then I would probably start thinking"* (P10).
- **Character** — The player has a preconceived notion from their mental model of who is trustworthy and who is not. As P2 mentions, if someone who is established to not be trustworthy says something false, *"I would say that is a narrative construct and an attribute that enhances the narrative of the story"*. If someone who has not shown to be untrustworthy does that, *"oh, that's a bug in [the NPC's] code, and not like that she was trying to deceive me, unless there's like other hints"*. In such a way, this highlights the importance of the initial impression that the NPC has on the player.
- **Perception of Beliefs** — To distinguish the intent to deceive, players may also consider what they think the NPC believes to be true. For example, basically all participants pointed towards the second quest in the game — where an NPC points you to a location a target NPC used to be (but is no longer) — *"I don't think they knew, it's not like they were, purposely misguiding me ... It's not it's not like [the NPC]'s fault that they didn't know she was moving"* (P7).

When there was no consequence and the false statement did not fit well within the established mental model (e.g. stated by someone trustworthy), participants found it difficult to assign a motive for deception to the character. As such, this led to it being seen more frequently as an unintentional error (e.g. *"because I feel like I had no reason to think that they were lying... I thought it's possible there could be mistakes in this"* — P10). However, when a character consistently lied and their lies had direct consequences on the game and fit within the context of the story, participants perceived this to be a more intentional use of lies, e.g. that the character was antagonistic against them for the sake of a grander narrative conflict.

4.3 Matching Expectations; Subsequent Behaviour

Finally, this section explores how the player's subsequent interactions with the NPC and behaviour within the game are affected by the attribution of false statements. On a broader scale, we also consider how it impacts their expectations, feelings, and impressions of the overall game. This section addresses RQ3 and RQ4 — understanding how NPCs who make false statements affect the player's gameplay interactions as well as their expectations and takeaways.

4.3.1 (Mis)matching the Expectations.

When trust deteriorates toward a certain NPC, it affects the subsequent interactions that a player has toward them. Some players may interact with them more, being interested in the character, e.g. *"really made me interested in what he had to say..."* (P22), some players may interact with them less *"I would be less likely to interact with them. Because I feel like it would just be like a kind of a waste of time"* (P13); however, basically all players mentioned that subsequent statements made would largely be taken with a *"grain of salt"* (P3, P22, P25, P29).

However, if trust has not deteriorated with a specific NPC and that NPC says something false, players may attempt to rationalize this mismatch through its attribution as described prior (if they even catch it, as a subset of players failed to catch many of the false statements because they simply believed in the NPCs). Other than the previously described attributions, a small number of players also sometimes rationalized such statements by attributing it to their own mistake, usually tying into gaps within their mental model, *"I attributed it to like me misremembering where it was or like me missing it"* (P13). When the mismatch between the player's expectations for trust and what the game presents is too great and the player cannot successfully rationalize it, this might lead to player frustrations that affect their attitude beyond solely the character to that of the whole game. For example, a common hypothetical scenario brought up was that of a trustworthy character telling you to go somewhere that would take a significant amount of time. If players were led there and found nothing, players mentioned that *"I would be so annoyed. ... If there was no reason given for it, ... then I would attribute it to them being evil. But if there wasn't any of that, I would just attribute it to like a poor game. Honestly, like it would just bother me"* (P13), and *"I would be kind of annoyed at the game itself and the game developers ... if there's nothing to discover, there's nothing to do over there, then I'd be annoyed at the game itself"* (P18).

Extending on the prior scenario — when players have identified that the characters have a motive for making false statements that drive some possible antagonistic conflict, there was a general sentiment that such a conflict should have a climax and resolution within the game. As established, the intentional use of false statements can serve a narrative purpose — players expect this narrative to manifest, to have a "payoff". P2 mentions that this *"payoff could either be a simple thing like the NPC acknowledges that she or he lied"*; or on a grander scale, *"here's like this greater narrative story about like who my character is, and maybe their character has like a reputation that elicits people to lie to them"*. P19 states that *"I expect a plot twist at the end. To like reveal the real reason why they're lying"*, wanting the resolution to also provide an explanation for why characters were acting in the way they were, and that *"if there's no explanation for why things are in the game, I would just assume it's just like a little unfinished"*. However, a few participants mentioned that developers can keep the game open-ended, however, and leave the interpretation and final resolution up to the player, e.g. *"Sometimes if it's done well, for a lot of writing, you can sometimes leave it up to the audience to kind of surmise their own ending"* (P24). All in all, the balance between player expectation and game delivery when it comes to false statements leads to what P2 summarises as a *"dichotomy"* between a game that has *"a really immersive experience"* and a game which elicits the response *"oh this game is really buggy"*.

4.3.2 Meaning and Expectations; A Note on AI-Written Work. Contrasting our findings regarding the intentional use of false statements and attribution of such, LLMs that hallucinate facts deprive the game of an intentional meaning — when an LLM makes such false statements, there is no rational narrative or in-game reason. Nearly all players mentioned that knowing a game employed the use of LLMs would, by default, lower their levels of trust in all characters, e.g. *"I will be more paranoid, I guess. I would be like, okay, I shouldn't trust not even the first person I'm talking to"* (P26). When false statements are not perceived to be crafted with intention, this can negatively impact the player perception of the meaning of the game. For example — *"I guess knowing that it is AI, it kind of lifts that sort of deeper meaning behind the wrong directions or intentions... when we read a poem, if it's some sort of visual, we will most likely take it as some sort of metaphor rather than as its face fact. Because yeah, we want writing to have meaning"* (P25); which shows that the artistry or the attributed meaning of the game is lost when a story is known to be AI-written — this is pertinent as in prior sections we have emphasized how players aim to find meaning in false statements and deception. Overall, the use of AI in storytelling was generally met with criticism — even if the game itself is still the same, the knowledge that it was AI-written impacts how the players feel towards the game — *"Then I didn't really care much for that NPC or that dialogue. It detaches me From the game, it detaches me from the experience... You know, you kinda want that bit of anchoring to reality"* (P29) and *"That makes me feel like it's pointless as a game if it is. Because if the dialogue is AI generated, then it's just sort of like predicting likely ways to string this text together"* (P18).

When discussing how LLMs could be incorporated into games for NPC dialogue, given the assumption that such systems might always hallucinate, participants brought up an extensive QA process as the main way to alleviate such issues, i.e. *"test the game enough so that you know that the output of the AI is trustworthy, you know, hopefully, although that is also to some extent"* (P11); *"[if] I were like writing a book, you know, there's gonna be like editors, right? And I think the editors need to do like I guess 10 times more..."* (P7). P5 mentions that at this stage of AI, *"we still need human interpretation"* before publishing a game. The other main solution was to curb the player's initial expectations (and thus change their initial mental model) by adding disclaimers to the game. P9 mentioned that *"you should put out disclaimers... make it clear that this might affect the gameplay... I would just say all these characters are written in a way that they might deceive you"*. From a marketing angle though, some participants indicated that disclaimers might cause players to be reluctant to purchase them, as *"I feel like if I were about to play or purchase a game that had that disclaimer of there may be errors, I think I just wouldn't play it"* (P3).

5 DISCUSSION

Our discussion focusses on the implications of our findings on game design, addressing RQ5. We consider how our findings fit into broader ludological research and how they may inform design decisions for game developers who use intentional deception or use LLM models that may hallucinate regardless of intent.

5.1 Realistic Social Behaviour, Attribution, and Immersion

Our findings revealed that the use of false statements and lies in games was shown to have a variety of interesting perceived narrative purposes. Perception of trust and deception can differ towards a real person versus a virtual agent [92], and players often noted that they had an initial baseline of trust towards the characters having been given no reason by the game to mistrust these characters. Thus, by having characters that seem to have motives with the intent to deceive, the game can craft conflict and antagonism into the story. By having characters that seem to forget or misremember, the game can add a level of realism, creating narrative immersion and humanizing characters to be more relatable [24]. Thus, attributing false statements in certain intentional ways can cause the narrative and characterization to be more complex and interesting. We highlighted factors that may affect this attribution, but ultimately those factors are still weighted by a player's subjectivity; as such, some interpretation of the meaning is left up to the players themselves. Nonetheless, each player constructing their own meaning through their experience can be viewed as a positive game design aspect, e.g. to promote social discussion [119].

Furthermore, we note that using deception and false statements in the game adds an additional social challenge — where the players must discern which statements are true and who they can trust. Although social challenge can be fun (and can promote certain learning objectives, such as the use of social deception games to teach communication [109]), developers should make sure that the level of cognitive demand does not become overwhelming to such a level the player does not know where to ground themselves. NPCs can create a level of social engagement when players immerse themself in the interactions [112]. However, sometimes players have a default level of trust in the NPCs (depending on meta-factors such as genre or NPC-specific factors such as appearance), especially when NPCs often serve objective purposes within the game beyond being narrative vehicles [7]. The concept of flow balances the level of challenge against skill [25], and we highlight how employing aspects of perceived trust, deception, and untrue statements may aid in achieving such an immersive state through social engagement.

A design implication for developers, therefore, involves the use of consistent intentional deception by NPCs in suitable games, e.g. mystery games, to develop intrigue, and characterization, and add a deductive challenge to the players (akin to Werewolf or Mafia [50]) — in fact, many such games do have NPCs that might lie[4][5]; with a significant part of the gameplay revolving around uncovering these lies. On the other hand, falsehoods that might seem inconsistently written, irrational, and with no clear motive, can incur feelings of confusion towards the game; players might view the game as annoying or unfinished. However, the intentional juxtaposition of seemingly irrational false statements against a meticulously developed game, knowing that players have a desire to search for meaning, can also be carefully used by game developers to create singular scenarios of confusion or bewilderment within the game. Leaving facets of the meaning open for the player could generate positive experiences, similar to more traditional art experiences [75]. Nonetheless, game developers must still ultimately balance the knife's edge of making the game fun as a social vehicle (acknowledging that players tend towards initial trust) versus making the game frustrating as one — for example, a game in which every character lies could be frustrating for the player by making it hard to know where to even start.

5.2 Games with False Statements; Player Considerations

When designing games with false statements, a factor that game developers need to consider is the demographics of the target audience. Participants brought up the idea that perhaps children might have difficulty handling and discerning lies — age can indeed play somewhat of a role in judgement [15, 87]; and research has shown that several other demographic factors can impact the perception of a lie [37]. A particular demographic whose needs are often overlooked in game design is that of the neurodivergent population. Spiel and Gerling argued that existing games can sometimes fail to support the self-determination of neurodivergent players [104] — Ranick et al. note that some of such players can have increased difficulty in detecting sarcasm, metaphor or deception [90]. Thus, when developing a game in which characters may not always be the most truthful or may involve intentional deception, a game developer should be cognizant of 1) the purpose of the game and 2) how the target audience may react. We suggest that sufficient playtesting should be done for the developer to better understand the mental model and judgement of the specific target audience, to both acknowledge their ethical responsibility to the player and to understand how their game might be perceived in terms of fun and enjoyability. Furthermore, if the game developer chooses to use an AI model (which is itself subject to generating false statements through hallucinations), the audience's perception of such unintentional false statements becomes another responsibility that the developer needs to factor in during the development process, on top of the extant ethical concerns [116] that come with the use of dialogue-based AI. We suggest that game developers perform research about audience perceptions towards LLMs and the existing concerns regarding their usage, and weigh that against their potential benefits.

Although the use of intentionally false statements in games was fine for every player in our study, many players expressed dislike towards the use of LLMs for dialogue generation in the game, citing the untrue yet believable text generation as one of the main reasons. The use of LLMs, and AI in general, has become a contentious topic in discussions regarding proper usage in games when it comes to the uncertainty of such algorithmic approaches. From a general standpoint, Kim et al. discussed 3 possible error-handling designs for uncertainty in ML — 1) ambiguity, in which a statement made becomes more broadly applicable when the model is not certain, 2) transparency, in which the statement made admits that the model is uncertain and asks for recalibration, and 3) controllability, in which statements are made based on a player's controlled confidence threshold, which also might allow for players to correct the errors [53]. Although their work was mainly based around object detection, it would be interesting to see if similar (or new) properties could hold for dialogue, where an NPC similarly might have a level of uncertainty in what they are saying. Thus, in designing NPCs that do use LLMs (and perhaps more generally, NPCs that use AI at any point), developers could take inspiration from these error-handling cases, and consider both reactive (e.g. allowing the player to correct the NPC dialogue error) and preventative (e.g. avoiding an error through providing more general statements when uncertainty is at a certain threshold) methods to handle uncertain outputs. Furthermore, the LLM could also alleviate such an issue by checking against another source of truth [38, 86], e.g. by querying a game's database via retrieval-augmented generation. Currently, though, the use of AI in distributed works still remains a topic of hot debate, with issues surrounding copyright [27], disclosure [11], etc.

5.3 Games, LLMs, and "Meaning"

Video games are increasingly being seen as a potential medium for storytelling and narrative; similar to other forms of narrative media, they have the opportunity to imbue metaphorical meaning and create thought-provoking experiences [10, 45, 94]. In our findings, players attributed the literary element of deception in games to some sort of thematic meaning, such as characterizing a person as antagonistic or setting up a major conflict. However, unlike a more traditional form of media such as a book or movie, a game affords the interactivity for a player to choose how they want to act on a false statement. Thus, in games, when a scriptwriter writes a false statement, they write it with some purpose that gives rise to some consequence that affects the player's interactions and experience within the game. However, when LLMs write a false statement due to hallucination, there is no human rationality in this action and thus the thematic meaning is lost; the mismatch between a player's desire for meaning and the lack thereof was often described to be a letdown for players after playing the game even if the actual writing between a human and a LLM was indistinguishable. This aspect is slightly reminiscent of Chekhov's gun — a narrative principle

[4] https://store.steampowered.com/app/413410/Danganronpa_Trigger_Happy_Havoc/ [Last Accessed: Nov 21, 2023]

[5] https://store.steampowered.com/app/1449200/AI_THE_SOMNIUM_FILES_ _nirvanA_Initiative/ [Last Accessed: Nov 21, 2023]

that all elements in a story should be purposeful [64]; in an abstract sense, LLMs write dialogue without a real "purpose".

We take a philosophical viewpoint of this dilemma through Albert Camus' idea of absurdism, which in general terms draws upon the confrontation between the human desire for definitive meaning and the lack thereof in the universe [41]; this contradiction somewhat mirrors the dilemma of using LLMs. Presently, it is difficult to attribute intention or purpose to LLMs, which quite literally simply string words together algorithmically. Work in the field of explainable AI might allow for an increased level of understanding of how AI systems function [51, 62], and might provide future understanding regarding LLM intention. However, at the current stage of research and with the hastened use of AI in daily life, humans may have to develop their subjective meaning of AI-created artistic exploits while accepting that there may not always be an objective meaning. For instance, some participants were able to imagine their own ideas of why certain characters made false statements, which could be vastly different from the actual intention behind them. In such a way, each person can create their own subjective meaning of the statement and of the game, which can drive their own unique takeaways, thoughts, and emotions regarding the game disparate from that of the game's developer.

Nevertheless, many of the participants were still quite opposed to the idea of AI-created artistry, valuing the connection and metaphorical value that human-created works provide even if the meaning might not always fully align. The results of our study could even be framed as an argument against the use of LLMs in games — hallucinations muddle the meaning that humans attribute to false statements, and the metaphorical value that human-written works provide is missed. However, this also opens up discussion towards possible future work — in alleviating hallucinations to make AI systems more trustworthy and by adding explainable rationale for AI-written text [102]. Furthermore, it opens the door for developers to more deeply consider how players might perceive meaning in writing (even beyond their own intention), and how those meanings might contribute to player actions within the game, feelings towards the game and its characters, and even judgment regarding the game's enjoyability. Within artistic mediums, the intention of the creator and the meaning-making of the audience is almost always subjective and fluid [110] — a potential way for developers to better convey their meaning more directly is through providing artefacts similar to director commentaries (e.g. in BioShock[6] or Portal 2[7]) or game developer diaries [42], which can offer a direct line of communication of the developers' intention.

6 LIMITATIONS

One limitation of the study is in the use of our developed game. The purpose of this game in the user study is to highlight the use of false statements in games; it is used as a gateway into discussion regarding the player's broader experience within games in general. Rather than solely relying on the memory of false statements in games, we aimed to present a specific grounded example for players. However, we identify that, especially due to recency bias as the interview was conducted right after the game, players may make statements that are overgeneralized to our specific game — a short 2D RPG set in a bucolic village. This is notable as 1) genre and atmosphere were mentioned to affect a player's initial expectations for trustworthiness, and 2) the game is relatively short, which might cause some statements regarding subsequent interactions to fall into the realm of hypotheticals. Furthermore, our game has a different level of "polish" when compared to e.g. AAA games, which also affects the initial expectations of players. These were necessary constraints due to time and resource limitations, however, one methodological improvement may be to introduce a variety of different games from different genres so players have a broader understanding of the purpose of the study.

This study furthermore relies on the players paying attention to the dialogue. Even when the participants did pay attention, we noticed that there were often some gaps in their memory, which we had to jog for them to sometimes remember. This was exacerbated by the fact that the initial study as described to them did not focus on the trust or verity of statements, as such, this was not something participants were particularly looking for during the game. This might mean that many false statements escaped detection or were unremarkable enough that players did not notice. Perhaps a different methodological approach would be to tell participants about the purpose of the study and have them ponder on each statement instead of taking a retrospective approach. Another limitation regarding the players relates to participant recruitment. Due to the nature of recruitment, the demographics of the participants leaned towards educated young adults; this created the unintended effect that many of the participants demonstrated some knowledge of AI technologies and LLMs. However, we also note that this may partly attributable to the recent AI boom that has caused an explosion of interest even among the general public [40, 89, 98].

The primary researcher who wrote the paper and performed the literature review also developed the game and the script. As such, we highlight that, although the researcher was cognizant of their role within the study, they may have imbued some of their bias in developing the game (e.g. scriptwriting for the quests, choosing the setting for the game, etc). As the primary researcher was largely responsible for the qualitative analysis as well, we may introduce single-coder bias, which we aimed to mitigate through validation and discussion amongst the team.

7 CONCLUSION

False statements and intentional deception can be used in video game writing to enhance NPCs by adding complexity to characters and stories. However, with the increased interest in using LLMs in video games, which may hallucinate outputs that are untrue, the idea of intention and meaning in games becomes more muddled. We performed a comprehensive study to analyze how players perceive and attribute truthfulness and lies to NPCs. We developed "AlphaBetaCity", a short 2D role-playing game that liberally used LLMs in the writing of NPC dialogue, to serve as a conduit into the discussion regarding such topics. Our findings revealed how participants developed an instinctual expectation of trust towards NPCs, verified statements within the game, and attributed reasons

[6] https://bioshock.fandom.com/wiki/Golden_Film_Reels [Last Accessed: Dec 11 2023]
[7] https://combineoverwiki.net/wiki/Developer_commentary/Portal_2 [Last Accessed: Dec 11 2023]

for false statements. False statements can be used to denote antagonism, create conflict, and sow doubt; thus, when confronted with such false statements, players expected to find such meaning; the mismatch between this desire and the lack of meaning in unintentionally-written false statements was a source of frustration for many participants. Furthermore, regarding the use of LLMs in games, participants noted that their trust in NPCs would be lowered by knowing the dialogue was AI-written, and they would feel less attached to the characters overall. We tie our findings into making game design suggestions, discussing the acceptable use of false statements, the use of AI in general, and the desire for metaphorical meaning in games.

ACKNOWLEDGMENTS

This work was supported in part by the Natural Science and Engineering Research Council of Canada (NSERC) under Discovery Grant RGPIN-2019-05624.

REFERENCES

[1] Rehaf Aljammaz, Elizabeth Oliver, Jim Whitehead, and Michael Mateas. 2020. Scheherazade's Tavern: A Prototype For Deeper NPC Interactions. In *Proceedings of the 15th International Conference on the Foundations of Digital Games* (Bugibba, Malta) *(FDG '20)*. Association for Computing Machinery, New York, NY, USA, Article 22, 9 pages. https://doi.org/10.1145/3402942.3402984

[2] Adam J. Arico and Don Fallis. 2013. Lies, damned lies, and statistics: An empirical investigation of the concept of lying. *Philosophical Psychology* 26, 6 (2013), 790–816. https://doi.org/10.1080/09515089.2012.725977

[3] Trevor Ashby, Braden K Webb, Gregory Knapp, Jackson Searle, and Nancy Fulda. 2023. Personalized Quest and Dialogue Generation in Role-Playing Games: A Knowledge Graph- and Language Model-Based Approach. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 290, 20 pages. https://doi.org/10.1145/3544548.3581441

[4] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. 2018. Synthesizing Robust Adversarial Examples. https://doi.org/10.48550/arXiv.1707.07397 arXiv:1707.07397 [cs].

[5] Sasha Azad and Chris Martens. 2021. Little Computer People: A Survey and Taxonomy of Simulated Models of Social Interaction. *Proc. ACM Hum.-Comput. Interact.* 5, CHI PLAY, Article 245 (oct 2021), 30 pages. https://doi.org/10.1145/3474672

[6] Augusto Baffa, Pedro Sampaio, Bruno Feijó, and Mauricio Lana. 2017. Dealing with the Emotions of Non Player Characters. In *2017 16th Brazilian Symposium on Computer Games and Digital Entertainment (SBGames)*. 76–87. https://doi.org/10.1109/SBGames.2017.00017

[7] Richard A Bartle. 2004. *Designing virtual worlds.* New Riders.

[8] Sissela Bok. 1978. *Lying: Moral Choice in Public and Private Life.* Vintage Books, New York.

[9] Arianna Boldi and Amon Rapp. 2023. "Is It Legit, To You?". An Exploration of Players' Perceptions of Cheating in a Multiplayer Video Game: Making Sense of Uncertainty. *International Journal of Human–Computer Interaction* 0, 0 (2023), 1–21. https://doi.org/10.1080/10447318.2023.2204276 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10447318.2023.2204276.

[10] Julia Ayumi Bopp, Elisa D. Mekler, and Klaus Opwis. 2016. Negative Emotion, Positive Experience? Emotionally Moving Moments in Digital Games. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2996–3006. https://doi.org/10.1145/2858036.2858227

[11] Jeffrey Brainard. 2023. Journals take up arms against AI-written text. *Science* 379, 6634 (2023), 740–741.

[12] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association, Washington, DC, US, 57–71. https://doi.org/10.1037/13620-004

[13] Erin M Bryant. 2008. Real lies, white lies and gray lies: Towards a typology of deception. *Kaleidoscope: A Graduate Journal of Qualitative Communication Research* 7 (2008), 23.

[14] Jacqueline Burgess and Christian Jones. 2020. I harbour strong feelings for Tali despite her being a fictional character": Investigating videogame players' emotional attachments to non-player characters. *Game Studies* 20, 1 (2020).

[15] Kay Bussey. 1999. Children's Categorization and Evaluation of Different Types of Lies and Truths. *Child Development* 70, 6 (1999), 1338–1347. https://doi.org/10.1111/1467-8624.00098

[16] David Byrne. 2022. A worked example of Braun and Clarke's approach to reflexive thematic analysis. *Quality & Quantity* 56, 3 (01 Jun 2022), 1391–1412. https://doi.org/10.1007/s11135-021-01182-y

[17] Xavier Caddle, Curtis Gittens, and Michael Katchabaw. 2018. A Psychometric Detection System to Create Dynamic Psychosocial Relationships Between Non-Player Characters. In *2018 IEEE Games, Entertainment, Media Conference (GEM)*. 256–262. https://doi.org/10.1109/GEM.2018.8516452

[18] Kelly Caine. 2016. Local Standards for Sample Size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 981–992. https://doi.org/10.1145/2858036.2858498

[19] Thomas L. Carson. 2006. The Definition of Lying*. *Noûs* 40, 2 (2006), 284–306. https://doi.org/10.1111/j.0029-4624.2006.00610.x

[20] Cristiano Castelfranchi. 2000. Artificial liars: Why computers will (necessarily) deceive us and each other. *Ethics and Information Technology* 2, 2 (June 2000), 113–119. https://doi.org/10.1023/A:1010025403776

[21] Mustafa Mert Çelikok, Tomi Peltola, Pedram Daee, and Samuel Kaski. 2019. Interactive AI with a Theory of Mind. *CoRR* abs/1912.05284 (2019). arXiv:1912.05284 http://arxiv.org/abs/1912.05284

[22] Gokul Chittaranjan and Hayley Hung. 2010. Are you Awerewolf? Detecting deceptive roles and outcomes in a conversational role-playing game. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. 5334–5337. https://doi.org/10.1109/ICASSP.2010.5494961 ISSN: 2379-190X.

[23] Andry Chowanda, Peter Blanchfield, Martin Flintham, and Michel Valstar. 2016. Computational models of emotion, personality, and social relationships for interactions in games. *The 2016 International Conference on Autonomous Agents & Multiagent Systems.* https://nottingham-repository.worktribe.com/output/773613

[24] Iulia Coanda and Stef Aupers. 2021. Post-human encounters: Humanising the technological Other in videogames. *New Media & Society* 23, 5 (2021), 1236–1256. https://doi.org/10.1177/1461444820912388

[25] Mihaly Csikszentmihalyi, Sami Abuhamdeh, and Jeanne Nakamura. 2014. *Flow.* Springer Netherlands, Dordrecht, 227–238. https://doi.org/10.1007/978-94-017-9088-8_15

[26] F. Cuzzolin, A. Morelli, B. Cîrstea, and B. J. Sahakian. 2020. Knowing me, knowing you: theory of mind in AI. *Psychological Medicine* 50, 7 (2020), 1057–1061. https://doi.org/10.1017/S0033291720000835

[27] Wes Davis. 2023. *Valve won't approve Steam games that use copyright-infringing AI artwork.* The Verge. Retrieved Aug 25, 2023 from https://www.theverge.com/2023/7/1/23781339/valve-steam-ai-artwork-rejecting-banning-pc-games

[28] Luigi De Angelis, Francesco Baglivo, Guglielmo Arzilli, Gaetano Pierpaolo Privitera, Paolo Ferragina, Alberto Eugenio Tozzi, and Caterina Rizzo. 2023. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Frontiers in Public Health* 11 (2023). https://www.frontiersin.org/articles/10.3389/fpubh.2023.1166120

[29] Bella M. DePaulo, Matthew E. Ansfield, Susan E. Kirkendol, and Joseph M. Boden. 2004. Serious Lies. *Basic and Applied Social Psychology* 26, 2-3 (2004), 147–167. https://doi.org/10.1080/01973533.2004.9646402

[30] Bella M DePaulo, Deborah A Kashy, Susan E Kirkendol, Melissa M Wyer, and Jennifer A Epstein. 1996. Lying in everyday life. *Journal of personality and social psychology* 70, 5 (1996), 979.

[31] Joao Dias, Ruth Aylett, Ana Paiva, and Henrique Reis. 2013. The great deceivers: Virtual agents and believable lies. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 35.

[32] Michelle Drouin, Daniel Miller, Shaun M.J. Wehle, and Elisa Hernandez. 2016. Why do people lie online? "Because everyone lies on the internet". *Computers in Human Behavior* 64 (2016), 134–142. https://doi.org/10.1016/j.chb.2016.06.052

[33] Nouha Dziri, Andrea Madotto, Osmar R. Zaïane, and Avishek Joey Bose. 2021. Neural Path Hunter: Reducing Hallucination in Dialogue Systems via Path Grounding. *CoRR* abs/2104.08455 (2021). arXiv:2104.08455 https://arxiv.org/abs/2104.08455

[34] Gabriel Elvery. 2023. Undertale's Loveable Monsters: Investigating Parasocial Relationships with Non-Player Characters. *Games and Culture* 18, 4 (2023), 475–497. https://doi.org/10.1177/15554120221105464

[35] Don Fallis. 2009. What Is Lying? *The Journal of Philosophy* 106, 1 (2009), 29–56. http://www.jstor.org/stable/20620149

[36] Ylva Ferstl and Rachel McDonnell. 2018. A Perceptual Study on the Manipulation of Facial Features for Trait Portrayal in Virtual Agents. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents* (Sydney, NSW, Australia) *(IVA '18)*. Association for Computing Machinery, New York, NY, USA, 281–288. https://doi.org/10.1145/3267851.3267891

[37] Genyue Fu, Kang Lee, Catherine Ann Cameron, and Fen Xu. 2001. Chinese and Canadian Adults' Categorization and Evaluation of Lie- and Truth-Telling about Prosocial and Antisocial Behaviors. *Journal of Cross-Cultural Psychology* 32, 6 (2001), 720–727. https://doi.org/10.1177/0022022101032006005

[38] Boris A. Galitsky. 2023. Truth-O-Meter: Collaborating with LLM in Fighting its Hallucinations. *Preprints* (July 2023). https://doi.org/10.20944/preprints202307.

1723.v1

[39] Qi Chen Gao and Ali Emami. 2023. The Turing Quest: Can Transformers Make Good NPCs?. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, Vishakh Padmakumar, Gisela Vallejo, and Yao Fu (Eds.). Association for Computational Linguistics, Toronto, Canada, 93–103. https://doi.org/10.18653/v1/2023.acl-srw.17

[40] Erin Griffith and Cade Metz. 2023. A new area of AI booms, even amid the tech gloom. *New York Times, January* 7 (2023).

[41] H. Gaston Hall. 1960. Aspects of the Absurd. *Yale French Studies* 25 (1960), 26–32. http://www.jstor.org/stable/2928897

[42] Xavier Ho. 2016. Tapping into the Gaming Community for Roguelikes. In *Proceedings of DiGRA Australia Queensland Symposium 2016: Wayfinding.*

[43] Gijs Huisman, Jan Kolkmeier, and Dirk Heylen. 2014. With Us or Against Us: Simulated Social Touch by Virtual Agents in a Cooperative or Competitive Setting. In *Intelligent Virtual Agents*, Timothy Bickmore, Stacy Marsella, and Candace Sidner (Eds.). Springer International Publishing, Cham, 204–213.

[44] Samee Ibraheem, Gaoyue Zhou, and John DeNero. 2022. Putting the Con in Context: Identifying Deceptive Actors in the Game of Mafia. https://doi.org/10.48550/arXiv.2207.02253 arXiv:2207.02253 [cs].

[45] Glena H. Iten, Sharon T. Steinemann, and Klaus Opwis. 2018. Choosing to Help Monsters: A Mixed-Method Examination of Meaningful Choices in Narrative-Rich Games and Interactive Narratives. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173915

[46] Charlene Jennett, Anna L. Cox, Paul Cairns, Samira Dhoparee, Andrew Epps, Tim Tijs, and Alison Walton. 2008. Measuring and defining the experience of immersion in games. *International Journal of Human-Computer Studies* 66, 9 (2008), 641–661. https://doi.org/10.1016/j.ijhcs.2008.04.004

[47] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* 55, 12, Article 248 (mar 2023), 38 pages. https://doi.org/10.1145/3571730

[48] Agne Kajackaite and Uri Gneezy. 2017. Incentives and cheating. *Games and Economic Behavior* 102 (March 2017), 433–444. https://doi.org/10.1016/j.geb.2017.01.015

[49] A. Kalbiyev. 2022. Affective dialogue generation for video games. http://essay.utwente.nl/89325/

[50] Yoshinobu Kano, Claus Aranha, Michimasa Inaba, Fujio Toriumi, Hirotaka Osawa, Daisuke Katagami, Takashi Otsuki, Issei Tsunoda, Shoji Nagayama, Dolça Tellols, Yu Sugawara, and Yohei Nakata. 2019. Overview of AIWolfDial 2019 Shared Task: Contest of Automatic Dialog Agents to Play the Werewolf Game through Conversations. In *Proceedings of the 1st International Workshop of AI Werewolf and Dialog System (AIWolfDial2019)*, Yoshinobu Kano, Claus Aranha, Michimasa Inaba, Fujio Toriumi, Hirotaka Osawa, Daisuke Katagami, and Takashi Otsuki (Eds.). Association for Computational Linguistics, Tokyo, Japan, 1–6. https://doi.org/10.18653/v1/W19-8301

[51] Nora Kassner, Oyvind Tafjord, Ashish Sabharwal, Kyle Richardson, Hinrich Schutze, and Peter Clark. 2023. Language Models with Rationality. arXiv:2305.14250 [cs.CL]

[52] Jack Kelly, Michael Mateas, and Noah Wardrip-Fruin. 2023. Towards Computational Support with Language Models for TTRPG Game Masters. In *Proceedings of the 18th International Conference on the Foundations of Digital Games* (Lisbon, Portugal) *(FDG '23)*. Association for Computing Machinery, New York, NY, USA, Article 78, 4 pages. https://doi.org/10.1145/3582437.3587202

[53] Minji Kim, Kyungjin Lee, Rajesh Balan, and Youngki Lee. 2023. Bubbleu: Exploring Augmented Reality Game Design with Uncertain AI-Based Interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 784, 18 pages. https://doi.org/10.1145/3544548.3581270

[54] Tae Wan Kim, Tong Lu, Kyusong Lee, Zhaoqi Cheng, Yanhan Tang, and John N. Hooker. 2021. When is it permissible for artificial intelligence to lie? A trust-based approach. *CoRR* abs/2103.05434 (2021). arXiv:2103.05434 https://arxiv.org/abs/2103.05434

[55] Christine M. Korsgaard. 1986. The Right to Lie: Kant on Dealing with Evil. *Philosophy & Public Affairs* 15, 4 (1986), 325–349. http://www.jstor.org/stable/2265252

[56] Michal Kosinski. 2023. Theory of Mind May Have Spontaneously Emerged in Large Language Models. arXiv:2302.02083 [cs.CL]

[57] Julian Kücklich. 2007. Homo Deludens: Cheating as a Methodological Tool in Digital Games Research. *Convergence* 13, 4 (Nov. 2007), 355–367. https://doi.org/10.1177/1354856507081951 Publisher: SAGE Publications Ltd.

[58] Kang Lee, Fen Xu, Grenyue Fu, Catherine Ann Cameron, and Shumin Chen. 2001. Taiwan and Mainland Chinese and Canadian children's categorization and evaluation of lie- and truth-telling: A modesty effect. *British Journal of Developmental Psychology* 19, 4 (2001), 525–542. https://doi.org/10.1348/026151001166236

[59] Kwan Min Lee, Namkee Park, and Seung-A Jin. 2006. Narrative and interactivity in computer games. *Playing video games: Motives, responses, and consequences* (2006), 259–274.

[60] Michael Sangyeob Lee and Carrie Heeter. 2015. Cognitive Intervention and Reconciliation-NPC Believability in Single-Player RPGs. *International Journal of Role-Playing* 5 (2015), 47–65.

[61] Emma E. Levine and Maurice E. Schweitzer. 2015. Prosocial lies: When deception breeds trust. *Organizational Behavior and Human Decision Processes* 126 (Jan. 2015), 88–106. https://doi.org/10.1016/j.obhdp.2014.10.007

[62] Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, Wenhu Chen, and Xifeng Yan. 2022. Explanations from Large Language Models Make Small Reasoners Better. arXiv:2210.06726 [cs.CL]

[63] Svenn Lindskold and Pamela S. Walters. 1983. Categories for Acceptability of Lies. *The Journal of Social Psychology* 120, 1 (1983), 129–136. https://doi.org/10.1080/00224545.1983.9712018

[64] Christian Lund. 2021. Chekhov's gun and Narrative Topography in Social Science Texts. *Anthropology and Humanism* 46, 1 (2021), 54–68. https://doi.org/10.1111/anhu.12321

[65] Rohit Madan and Mona Ashok. 2023. AI adoption and diffusion in public administration: A systematic literature review and future research agenda. *Government Information Quarterly* 40, 1 (Jan. 2023), 101774. https://doi.org/10.1016/j.giq.2022.101774

[66] Irini A. Malegiannaki, Thanasis Daradoumis, and Symeon Retalis. 2020. Teaching Cultural Heritage through a Narrative-based Game. *Journal on Computing and Cultural Heritage* 13, 4 (Dec. 2020), 27:1–27:28. https://doi.org/10.1145/3414833

[67] Kerry Mallan. 2013. *Secrets, lies and children's fiction.* Palgrave Macmillan London.

[68] Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. arXiv:2303.08896 [cs.CL]

[69] Andreas Matthias. 2015. Robot lies in health care: When is deception morally permissible? *Kennedy Institute of Ethics Journal* 25, 2 (2015), 169–162.

[70] C. McColIum, C. Barba, T. Santarelli, and J. Deaton. 2004. Applying a cognitive architecture to control of virtual non-player characters. In *Proceedings of the 2004 Winter Simulation Conference, 2004.*, Vol. 1. 890. https://doi.org/10.1109/WSC.2004.1371404

[71] Marisa Mealy, Walter Stephan, and I. Carolina Urrutia. 2007. The acceptability of lies: A comparison of Ecuadorians and Euro-Americans. *International Journal of Intercultural Relations* 31, 6 (2007), 689–702. https://doi.org/10.1016/j.ijintrel.2007.06.002

[72] Aditya Mehta, Yug Kunjadiya, Aniket Kulkarni, and Manav Nagar. 2022. Exploring the viability of Conversational AI for Non-Playable Characters: A comprehensive survey. In *2021 4th International Conference on Recent Trends in Computer Science and Technology (ICRTCST)*. 96–102. https://doi.org/10.1109/ICRTCST54752.2022.9782047

[73] Jörg Meibauer. 2018. The Linguistics of Lying. *Annual Review of Linguistics* 4, 1 (2018), 357–375. https://doi.org/10.1146/annurev-linguistics-011817-045634

[74] Merriam-Webster. 2023. *Lie*. Merriam-Webster. Retrieved Aug 25, 2023 from https://www.merriam-webster.com/dictionary/lie

[75] Keith Millis. 2001. Making meaning brings pleasure: the influence of titles on aesthetic experiences. *Emotion* 1, 3 (2001), 320.

[76] Lanse P Minkler and Thomas J Miceli. 2004. Lying, Integrity, and Cooperation. *Review of Social Economy* 62, 1 (March 2004), 27–50. https://doi.org/10.1080/0034676042000183817 Publisher: Routledge _eprint: https://doi.org/10.1080/0034676042000183817.

[77] Buta Monica, Visu-Petra George, Visu-Petra Laura, et al. 2020. A little lie never hurt anyone: Attitudes toward various types of lies over the lifespan. *Psychology in Russia: State of the art* 13, 1 (2020), 70–81.

[78] Don Monroe. 2021. Deceiving AI. *Commun. ACM* 64, 6 (June 2021), 15–16. https://doi.org/10.1145/3460218

[79] John Mooney and Jan M. Allbeck. 2018. Rethinking NPC Intelligence: A New Reputation System. In *Proceedings of the 7th International Conference on Motion in Games* (Playa Vista, California) *(MIG '14)*. Association for Computing Machinery, New York, NY, USA, 55–60. https://doi.org/10.1145/2668084.2668091

[80] Matthew R. Newkirk. 2013. *Just deceivers: An investigation into the motif and theology of deception in the books of Samuel.* Ph. D. Dissertation. https://www.proquest.com/dissertations-theses/just-deceivers-investigation-into-motif-theology/docview/1430897156/se-2 Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2023-03-03.

[81] Feng Nie, Jin-Ge Yao, Jinpeng Wang, Rong Pan, and Chin-Yew Lin. 2019. A Simple Recipe towards Reducing Hallucination in Neural Surface Realisation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 2673–2679. https://doi.org/10.18653/v1/P19-1256

[82] Magalie Ochs, Nicolas Sabouret, and Vincent Corruble. 2021. Modeling the Dynamics of Non-Player Characters' Social Relations in Video Game. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* 4, 1 (Sep. 2021), 90–95. https://doi.org/10.1609/aiide.v4i1.18678

[83] Elaine Ostry. 2015. Secrets, Lies and Children's Fiction by Kerry Mallan. *The Lion and the Unicorn* 39, 2 (2015), 229–231.

[84] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. arXiv:2304.03442 [cs.HC]

[85] Dhaval Parmar, Stefan Olafsson, Dina Utami, and Timothy Bickmore. 2018. Looking the Part: The Effect of Attire and Setting on Perceptions of a Virtual Health Counselor. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents* (Sydney, NSW, Australia) *(IVA '18)*. Association for Computing Machinery, New York, NY, USA, 301–306. https://doi.org/10.1145/3267851.3267915

[86] Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback. arXiv:2302.12813 [cs.CL]

[87] Candida C. Peterson. 1995. The role of perceived intention to deceive in children's and adults' concepts of lying. *British Journal of Developmental Psychology* 13, 3 (1995), 237–260. https://doi.org/10.1111/j.2044-835X.1995.tb00677.x

[88] Katrin Preckel, Philipp Kanske, and Tania Singer. 2018. On the interaction of social affect and cognition: empathy, compassion and theory of mind. *Current Opinion in Behavioral Sciences* 19 (2018), 1–6. https://doi.org/10.1016/j.cobeha.2017.07.010 Emotion-cognition interactions.

[89] Weihong Qi, Jinsheng Pan, Hanjia Lyu, and Jiebo Luo. 2023. Excitements and Concerns in the Post-ChatGPT Era: Deciphering Public Perception of AI through Social Media Analysis. arXiv:2307.05809 [cs.SI]

[90] Jennifer Ranick, Angela Persicke, Jonathan Tarbox, and Jake A. Kornack. 2013. Teaching children with autism to detect and respond to deceptive statements. *Research in Autism Spectrum Disorders* 7, 4 (2013), 503–508. https://doi.org/10.1016/j.rasd.2012.12.001

[91] Brian Ravenet, Florian Pecune, Mathieu Chollet, and Catherine Pelachaud. 2016. *Emotion and Attitude Modeling for Non-player Characters*. Springer International Publishing, Cham, 139–154. https://doi.org/10.1007/978-3-319-41316-7_8

[92] Kantwon Rogers and Ayanna Howard. 2022. When a Robot Tells You That It Can Lie. In *2022 IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO)*. 1–7. https://doi.org/10.1109/ARSO54254.2022.9802976

[93] Victoria L. Rubin and Sarah C. Camm. 2013. Deception in video games: examining varieties of griefing. *Online Information Review* 37, 3 (Jan. 2013), 369–387. https://doi.org/10.1108/OIR-10-2011-0181 Publisher: Emerald Group Publishing Limited.

[94] Doris C. Rusch and Matthew J. Weise. 2008. Games about LOVE and TRUST? Harnessing the Power of Metaphors for Experience Design. In *Proceedings of the 2008 ACM SIGGRAPH Symposium on Video Games* (Los Angeles, California) *(Sandbox '08)*. Association for Computing Machinery, New York, NY, USA, 89–97. https://doi.org/10.1145/1401843.1401861

[95] James Ryan, Adam Summerville, Michael Mateas, and Noah Wardrip-Fruin. 2021. Toward Characters Who Observe, Tell, Misremember, and Lie. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* 11, 3 (Jun. 2021), 56–62. https://doi.org/10.1609/aiide.v11i3.12825

[96] Johnny Saldana. 2021. The Coding Manual for Qualitative Researchers. (2021), 1–440. https://www.torrossa.com/en/resources/an/5018667 Publisher: SAGE Publications Ltd.

[97] Stefan Sarkadi, Peidong Mei, and Edmond Awad. 2023. Should My Agent Lie for Me? A Study on Attitudes of US-Based Participants Towards Deceptive AI in Selected Future-of-Work. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems* (London, United Kingdom) *(AAMAS '23)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 345–354.

[98] Laura Sartori and Giulia Bocca. 2023. Minding the gap(s): public perceptions of AI and socio-technical imaginaries. *AI & SOCIETY* 38, 2 (April 2023), 443–458. https://doi.org/10.1007/s00146-022-01422-1

[99] Benjamin Saunders, Julius Sim, Tom Kingstone, Shula Baker, Jackie Waterfield, Bernadette Bartlam, Heather Burroughs, and Clare Jinks. 2018. Saturation in qualitative research: exploring its conceptualization and operationalization. *Quality & quantity* 52, 4 (2018), 1893–1907.

[100] Leonard Saxe. 1991. Lying: Thoughts of an applied social psychologist. *American Psychologist* 46, 4 (1991), 409.

[101] John S. Seiter, Jon Bruschke, and Chunsheng Bai. 2002. The acceptability of deception as a function of perceivers' culture, deceiver's intention, and deceiver-deceived relationship. *Western Journal of Communication* 66, 2 (2002), 158–180. https://doi.org/10.1080/10570310209374731

[102] Donghee Shin. 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies* 146 (2021), 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

[103] Guilherme Silva and Marcos Souza Ribeiro. 2021. Development of Non-Player Character with Believable Behavior: a systematic literature review. In *Anais Estendidos do XX Simpósio Brasileiro de Jogos e Entretenimento Digital* (Online). SBC, Porto Alegre, RS, Brasil, 319–323. https://doi.org/10.5753/sbgames_estendido.2021.19660

[104] Katta Spiel and Kathrin Gerling. 2021. The Purpose of Play: How HCI Games Research Fails Neurodivergent Populations. *ACM Trans. Comput.-Hum. Interact.* 28, 2, Article 11 (apr 2021), 40 pages. https://doi.org/10.1145/3432245

[105] Matthew Stephenson and Jochen Renz. 2018. Deceptive Angry Birds: Towards Smarter Game-Playing Agents. In *Proceedings of the 13th International Conference on the Foundations of Digital Games* (Malmö, Sweden) *(FDG '18)*. Association for Computing Machinery, New York, NY, USA, Article 13, 10 pages. https://doi.org/10.1145/3235765.3235775

[106] Andreas Stokke. 2013. Lying, Deceiving, and Misleading. *Philosophy Compass* 8, 4 (2013), 348–359. https://doi.org/10.1111/phc3.12022

[107] Simon Su, Edward Zhang, Paul Denny, and Nasser Giacaman. 2021. A Game-Based Approach for Teaching Algorithms and Data Structures using Visualizations. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education (SIGCSE '21)*. Association for Computing Machinery, New York, NY, USA, 1128–1134. https://doi.org/10.1145/3408877.3432520

[108] Grant Tavinor. 2005. Videogames and interactive fiction. *Philosophy and Literature* 29, 1 (2005), 24–40.

[109] Shane Tilton. 2019. Winning Through Deception: A Pedagogical Case Study on Using Social Deception Games to Teach Small Group Communication Theory. *SAGE Open* 9, 1 (2019), 2158244019834370. https://doi.org/10.1177/2158244019834370

[110] Pablo PL Tinio. 2013. From artistic creation to aesthetic reception: The mirror model of art. *Psychology of Aesthetics, Creativity, and the Arts* 7, 3 (2013), 265.

[111] Tomer Ullman. 2023. Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks. arXiv:2302.08399 [cs.AI]

[112] Henrik Warpefelt. 2016. *The Non-Player Character: Exploring the believability of NPC presentation and behavior*. Ph. D. Dissertation. Department of Computer and Systems Sciences, Stockholm University.

[113] David Weibel and Bartholomäus Wissmath. 2011. Immersion in computer games: The role of spatial presence and flow. *International Journal of Computer Games Technology* 2011 (2011), 6–6.

[114] Katharina Weitz, Dominik Schiller, Ruben Schlagowski, Tobias Huber, and Elisabeth André. 2021. "Let me explain!": exploring the potential of virtual agents in explainable AI interaction design. *Journal on Multimodal User Interfaces* 15, 2 (2021), 87–98.

[115] Alan F. T. Winfield. 2018. Experiments in Artificial Theory of Mind: From Safety to Story-Telling. *Frontiers in Robotics and AI* 5 (2018). https://doi.org/10.3389/frobt.2018.00075

[116] M. J. Wolf, K. Miller, and F. S. Grodzinsky. 2017. Why We Should Have Seen That Coming: Comments on Microsoft's Tay "Experiment," and Wider Implications. *SIGCAS Comput. Soc.* 47, 3 (sep 2017), 54–64. https://doi.org/10.1145/3144592.3144598

[117] Yuehua Wu and Vivian Hsueh Hua Chen. 2013. A social-cognitive approach to online game cheating. *Computers in Human Behavior* 29, 6 (Nov. 2013), 2557–2567. https://doi.org/10.1016/j.chb.2013.06.032

[118] Jeff Yan and Brian Randell. 2005. A systematic classification of cheating in online games. In *Proceedings of 4th ACM SIGCOMM workshop on Network and system support for games (NetGames '05)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/1103599.1103606

[119] Michael Yin and Robert Xiao. 2022. How Should I Respond to "Good Morning?": Understanding Choice in Narrative-Rich Games. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference* (Virtual Event, Australia) *(DIS '22)*. Association for Computing Machinery, New York, NY, USA, 726–744. https://doi.org/10.1145/3532106.3533459

[120] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story Writing With Large Language Models. In *27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 841–852. https://doi.org/10.1145/3490099.3511105

[121] Tricia Zakreski. 2012. Tell Me Lies: Lying, Storytelling, and the Romance Novel as Feminist Fiction. *Journal of Popular Romance Studies* 2, 2 (2012).

[122] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A Survey of Large Language Models. arXiv:2303.18223 [cs.CL]

[123] Pei Zhou, Andrew Zhu, Jennifer Hu, Jay Pujara, Xiang Ren, Chris Callison-Burch, Yejin Choi, and Prithviraj Ammanabrolu. 2023. I Cast Detect Thoughts: Learning to Converse and Guide with Intents and Theory-of-Mind in Dungeons and Dragons. arXiv:2212.10060 [cs.CL]