Artificial Intelligence, Games, and Education

Barbero, G.

Chapter 3

The Relevance of Games for Artificial Intelligence

The present chapter is centred on the relationship between games and AI. We start with a section that describes past successes and future challenges for AI in the field of video games. Through that, we will see how games represent milestones in the development of intelligent technologies. In the second section, we focus our attention on the use of generative AI. Generative AI has been applied in video games in order to create new and unique content through PCG [24, 25, 26]. We already mentioned in 1 how this thesis explores a different experimental application of generative AI as an agent. In this regard, the second section presents a practical example of how modern generative AI can already be embedded as an agent in video games. In particular, we will see how video games are effective contexts of interaction between humans and AI agents. The section also functions as an introduction to experimental work with generative AI in video game research. Finally, a third section introduces a new type of technological application to games, defining the field of *hybrid games*. While not directly related to video games, we consider this an extension of possible AI applications in playful contexts with related potentials. Also, in this case, each section is closed with a summary relating our findings to the research questions.

3.1 Challenges of open-world Games for AI: Insights from Human Gameplay

For decades, games have served as a reliable testing ground for new AI technologies. One of the first tests for AI, the Turing test, can arguably be considered a game [127]. In recent years, many technological leaps in the field of AI have been demonstrated by pitting humans against computers. We have famous examples such as the program AlphaGo [27] beating Go champion Lee Sedol in 2015, or AlphaStar reaching Grandmaster status in the strategy video game StarCraft II in 2019 [28]. Achievements of AI systems in the field of gaming have inspired relevant studies in various fields; besides developments in computer science (e.g., the study of evolutionary algorithms for AlphaStar [128]), they also sparked analyses in social sciences (e.g., in psychology [129]). In this paper, we focus our attention on the potential interaction between AI and open-world games. Open-world games are a genre of video games that offer players a freely explorable virtual environment, usually without strict linear gameplay [130]. This type of game intrinsically presents characteristics that make them notably more challenging to be tackled by AI; they tend to be heavily focused on curiosity-driven exploration, they allow numerous combinations of actions, they are less related to optimisation problems and more to adapting to a variable context [131]. Being able to proficiently *live* and play an open-world game would be a relevant development in the field of AI. It would demonstrate adaptability and flexibility while tackling very diverse challenges with fuzzy goals. It would also require autonomous reasoning derived from previous experience and current knowledge of the game context. Arguably, these challenges make playing open-world games a closer representation of real-world interactions. In the following sections, we present an overview of the current state of AI in open-world games. We examine the limitations associated with existing approaches and highlight the importance of an approach inspired by human play. Subsequently, we outline our methodology for investigating human gameplay in open-world games. We detail our experimental procedures, analyse the findings, and present three distinct perspectives, ranging from general to specific, describing human interactions within open-world digital environments. Finally, we use these perspectives to delineate areas in which AI advancement is necessary to address challenges posed by open-world games.

State of the Art

AI agents engaging with open-world games have gathered significant interest over time. A vast array of previous research exists, albeit mostly focused on very specific aspects of open-world AI, such as behaviour trees. Typically, the aim is to enhance player experiences, for example, improving player modelling using long-short-term memory neural networks [132]. Recent studies have reported promising advancements in players' goal recognition through multimodal deep learning and players' self-reflection [133]. Also, the potential of AI to tackle planning has been explored in open-world games like Minecraft, through LLMs [134]. However, without the planning ability of LLMs, challenges that focus on playing Minecraft through AI usually set a predetermined goal (e.g., obtaining the "diamond" resource) in order to have a quantifiable measure of success. Another focal point of research lies in the interaction with NPCs, which are central elements populating most open-world games and guiding the players' experience. Deep neural networks and other AI systems have been tested to imbue NPCs with human-like behavior [135]. Additionally, LLMs have been recently considered to generate context-aware background chatter, even though concerns remain over the lack of control over the output [136]. Overall, AI has primarily been used to tackle specific mechanics of open-world games; it often acts in auxiliary roles rather than as the main player. Alternatively, several efforts are being made to use AI in debugging phases. However, they are still in nascent stages, typically producing conceptual frameworks [137, 138]. Developing an AI capable of meaningfully engaging with open-world games remains a challenge. In this regard, the ambiguity of the term "meaningful" in this context is representative of the complexity of developing AI for open-world games. Considering the aforementioned definition of open-world [130] (but also [139] or [140]), it becomes evident that one of the salient characteristics of the genre is the tendency to diverge from linear game-play structures. Open-world games typically encourage the player to interact with the environment freely (e.g. adding side quests when interacting with certain NPCs, rewarding a visit to a previously unexplored area with experience points for the player) without predefined victory conditions or foreseeable objectives.

3.1.1 Problem Definition

In order to design an AI system to meaningfully play in open-world settings, we need to define what "meaningful play" entails. In particular, we explore it through the lens of three concentric activities, from broad to specific (see Figure 3.1):

- **Planning** - Entails formulating a functional concept of game completion. Without strict, predefined goals, how are humans able to select objectives and to find the theoretical steps to reach those objectives?

- **Decision making** - Involves identifying the types of knowledge involved in decision-making in open-world games. Given that the planning step mentioned above entails several decisions, what types of prior knowledge inform humans to make these decisions effectively?

- **Interacting** - Describes the interactions necessary in order to practically alter the game state. Once a plan is defined and theoretical decisions are made, it is necessary to be able to interact proficiently with the game environment. What are the game design elements we typically interact with to enact our plans?

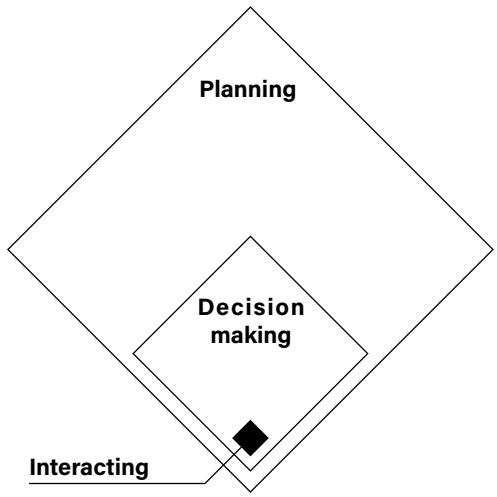

Figure 3.1: The structure of actions involved in meaningfully playing an open-world game

Given the definition itself of these questions, it is evident that human play needs to be central in our methodology. Therefore, our data collection design is rooted in traditional human-computer interaction methods. However, our data processing and analysis suite will be composed of both conventional techniques and approaches using modern technologies.

3.1.2 Methodology

We gather our data using a standard **talk-aloud protocol** to generate a play transcript. This technique has already been used successfully in game contexts to study players' behaviour [141]. For the experiment, we selected the game The Outer Worlds [142] because it includes most design patterns typical of open-world games (e.g., combat, freedom of choice, character development, etc.) and it aids replicability due to its availability on multiple platforms. However, it is important to point out that the experiment has been exclusively performed using a keyboard and mouse input setup. Moreover, the game has been prepared in advance, setting it up immediately after the tutorial. This enhances the comparability of transcripts while still providing comprehensive information about the game environment. As for the data analysis process, our methodology changes depending on the problem we tackle, as listed in 3.1.1. We explore how humans determine objectives and plan through **observation** of their play session. Subsequently, we define the types of knowledge involved in decision-making, highlighting respective moments in the talk-aloud protocol transcript and **manually clustering** them in categories of knowledge involved. Finally, we make use of a **GPT-3.5** LLM to extract the game elements used by players to interact with the environment from the transcript. Due to the risk of hallucinations associated with LLMs [143], we validate the list of game elements produced by finding related sentences in the transcript. Moreover, we compare the game elements with a list of game design patterns typical of open-world games [144].

Gathering data: Talk aloud protocol

The procedure starts with informed consent and a short gaming habits survey. The participants are asked three questions:

- How many hours per week do you play video games?
- Have you ever played an open-world game?
- If yes, which one(s)?

The participants are then introduced to the game and the input setup via the menu page, which lists all the key bindings. They are instructed to simply try to play the game while explicitly explaining their reasoning in the process. The researcher does not interact, except in case the participant needs to be reminded to talk aloud. The play-through is then transcribed automatically and double-checked using a separate

voice recording. Informed consent and survey take in total five minutes, and the talk aloud protocol takes twenty minutes.

Processing data: Observation

Throughout each session, we take notes on how the participants direct themselves in the game. We take notice of how they orient the character at the beginning and what affordances or game elements they seek in order to formulate objectives. From these observations, we deduce typical game-play goals and we formulate a functional framework that supports them. The framework aims to represent a standard system of steps that can arguably lead to completing the game.

Processing data: Manual Clustering

In the transcript obtained from the talk-aloud protocol, we highlight situations in which the player needs to make decisions. Subsequently, we extract what type of knowledge is used in order to make the decision. The types of knowledge are then clustered to find recurring categories. We predict that these categories have a certain degree of overlap with each other.

Processing data: GPT-3.5 Clustering

Finally, we feed all the transcripts to a GPT-3.5 LLM. We then ask the model to find common problem-solving techniques. The output is a list of design patterns that players utilise to practically interact with open-world environments. We validate the generated design patterns by comparing them with a preexisting list [144] and the scripts from the talk-aloud protocol themselves. We then discuss the results critically, speculate about how current AI would perform in similar contexts and highlight strategies to guide the development of new game AI systems.

3.1.3 Results

We recruit participants (N = 5) from the university students and staff of the faculty of science. The participants are between 21 and 45 years old, with an average age of 28.2. All participants reported playing video games at least one hour per week, with one playing more than ten hours. No participant reported playing less than one hour. All participants played open-world games before.
In the rest of this section, we report our results, structuring them in three subsections.

In the first, we describe the findings derived from the observation of the participants. We focus our attention on the ways players define their goals. Additionally, we compare the most common strategies with the players' gaming habits from the briefing survey. As a result, we define and describe a functional framework that schematises the participants' goals in relation to the game structure. In the second part, we report emblematic participants' statements related to how information is gathered and processed in the game. From these, we lead to three partially overlapping categories that indicate from what context the participants drew knowledge throughout their gameplay. In the last part, we report the results of the GPT-clustering as a list of game elements. These are typical design patterns of open-world games that the participants commonly use to interact with the game proficiently. Each is accompanied by a brief description.

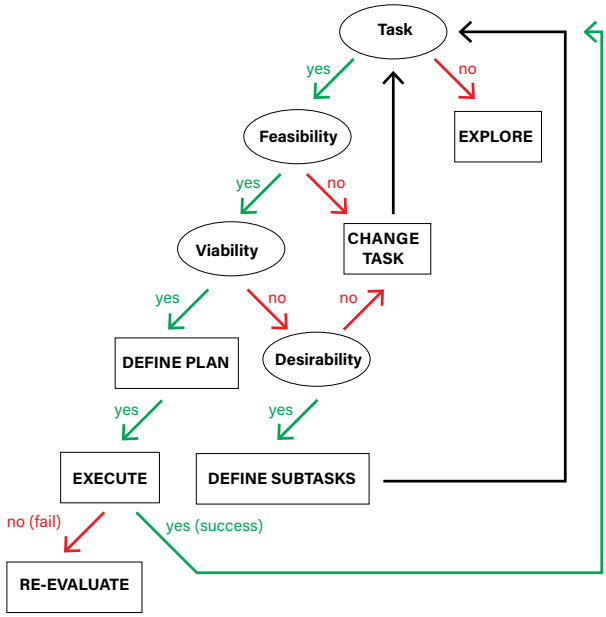

Figure 3.2: The Player Decisions Framework: in ellipses, the evaluation steps, in rectangles, action steps

3.1. Challenges of open-world Games for AI: Insights from Human Gameplay

Framework

In this experiment, we observe how players attempt to define an objective in the first instance of gameplay. We identified two main methods employed by the participants to analyse the game state and formulate goals:

- players with solid previous experience in open-world games tend to immediately look for affordances indicating directions or points of interest in the game interface (i.e., a quest mark on the built-in compass or a mission journal).
- players with less experience in this type of game tend to rely more on the environment, exploring the world and trying to identify relevant marks in the camera view (e.g., highlights on objects of interest or general direction marks).

The first method, which demonstrates previous knowledge of open-world games, is more direct and task-focused. It tends to find what objectives (or quests in this case) are currently available and how to complete them. The second approach tends to let the world reveal itself to the player by focusing on exploration. Both, however, aim to identify the current quest as the goal. Once the quest is correctly identified, the approaches to complete it can be diverse; while some participants use a direct trial-and-error approach, others check whether they are prepared to take on the challenge by estimating its difficulty. We generalise all the different types of approaches in a basic framework, inclusive of all this information (see Figure 3.2). It is structured as a series of steps that would theoretically exemplify a basic game-solving paradigm.

- **task**: the player evaluates whether they have a specific task in mind or not. In case the player has a specific task they want to complete, they move to the *feasibility* evaluation. If they do not, they need to *explore*.
- **explore**: we define exploring as the action required to find the next task. This can include the exploration of the virtual world or the exploration of the game system (e.g., the journal in order to find directions, the map, etc.). Exploring in open-world games is very dependent on the game itself and on its features.
- **feasibility**: the player needs to evaluate whether the task they selected is feasible, or possible for the actual rules of the game. Does the game allow that type of interaction? Does the task exist within the game? An example of an unfeasible task would be to try to hit a specific game agent (e.g., a giant bird) while the player does not recognise hitting that specific agent as possible (e.g., the player's arrow simply passes through the bird without any effect). If the

evaluation fails, the player has no choice but to *change task*. If the evaluation succeeds and the task is indeed feasible, the player can proceed to the *viability* evaluation.

- **change task**: the player needs to go back to find a new task that can pass the feasibility and viability evaluations.
- **viability**: the player here needs to evaluate whether the task selected can be completed with the current game state. This can involve current character levels, the acquisition of specific skills, or, more generally, meeting certain requirements. Within our framework, the player already knows at this point that the task is feasible. They have to define whether they can complete it in their current condition. Not all the games have completely defined states in this case: some would involve a viability evaluation that relates to how difficult the task might be (e.g., the task is viable but, at the current character level, the player will find it extremely challenging). Regardless, the options would still be two: in case the task is evaluated as viable, the player proceeds to define a plan (perhaps influenced by the challenge level). On the other hand, if the task is evaluated as not viable, the player moves towards the *desirability* evaluation.
- **desirability**: involves how desirable the task is, therefore, if it is really worth it. The reasons for finding a task desirable can vary from a personal preference to explicit game requirements (e.g., it is necessary to complete the task in order to advance in the main storyline).
- **define subtask**: the same framework can recursively be used to meet the requirements arising from the viability evaluation.
- **define plan**: the player, after having determined the task as feasible and viable, develops a plan to complete it. In most games, the typical design pattern of save-load cycles makes it possible to test the plan multiple times through execution.
- **execute plan**: the player executes the previously defined plan.
- **fail - re-evaluate**: the plan execution failed. The player needs to reevaluate the feasibility and viability first in order to determine whether the failure was caused by a misinterpretation of the game system and state. If both evaluations are deemed correct again, the player needs to define and test a new plan.
- **success - new task**: the plan execution succeeded. The player can proceed to find the next task.

Categories of Information

While the framework helps to define an approach to complete quests, it does not necessarily help to identify how and what information is used in the process (e.g., where can the player find the information to evaluate the quest's feasibility?). Using the transcript from the talk-aloud protocol, we focus our attention on the moments in which the participant evaluates the game state to make decisions. We then cluster the participants' statements depending on what type of knowledge they are using:

- **previous experience from real world**: players make decisions based on knowledge acquired in the real world. Some exemplary statements in this category are *"I don't know how to get there, maybe if I follow the road"* or *"The person I am looking for is a doctor, so probably I will head to the medical bay"*. In the first case, players identify a road as a landmark that connects different points. In the second case, they associate a medical facility with the presence of doctors.

- **previous experience in video games**: players make decisions based on knowledge acquired in other video games, as evidenced by statements such as *"Oh this dialogue has a long text so it is probably important"* or *"I will just look for the quest marker"*. In the first example, the player is used to the fact that primary dialogues are usually more extensive than secondary ones. In the second example, players are used to the presence of markers to identify quest destinations.

- **in-game information**: this category includes all the information that is introduced by the game itself, such as tutorials, pop-ups or advisory dialogues. In this case, exemplary statements are *"It highlights the person red so he must be an enemy"* or *"Because I stole something, I am now wanted"*. In both cases, the game provided information more or less explicitly.

These categories are not strictly separated, but they present a certain degree of overlap (see Figure 3.3). For example, the fact that the colour red usually identifies enemies is also something we can extract from previous experience with video games or even from the real world (where it identifies danger). Similarly, the fact that a road leads to an interesting point is knowledge that could be derived from other video games, but, at least in the case study of The Outer Worlds, is not explicitly reported by the game itself.

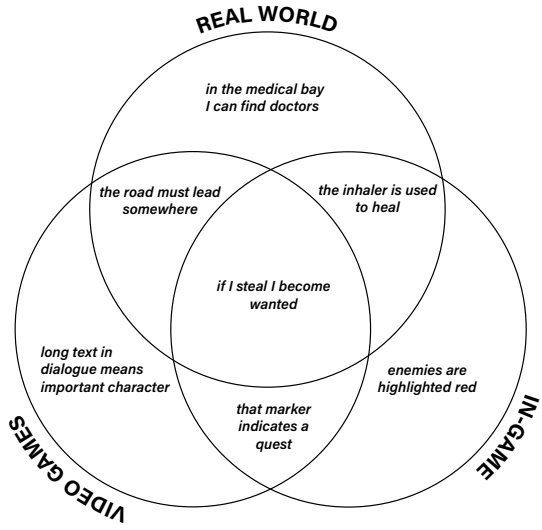

Figure 3.3: The three categories of information utilised to evaluate the game state and make decisions

Open-world Game Design Patterns

Following the two subsections above, we are able to identify the steps to select an objective and the information necessary to complete those steps. This last result refers to the actual game design patterns that need to be interacted with in order to practically perform the necessary actions. Feeding the transcripts to a GPT-3.5 model, we ask it to cluster recurring problem-solving strategies using the game elements on which they make use. The result is a list of typical open-world game mechanics that the participants interact with throughout their gameplay:

- **Interacting with NPCs**: The player interacts with NPCs to gather information, receive quests or trade for items.
- **Trial and Error**: The player tries different actions to progress in the game and understand its mechanics.
- **Inventory Management**: The player manages their inventory, including buying and selling items, equipping weapons and using consumables like health items.

- **Reading Text and Instructions**: The player reads in-game text, instructions and quest logs to understand objectives and game play mechanics.
- **Combat Strategies**: The player employs different combat strategies, such as using ranged weapons, melee attacks, or stealth approaches, depending on the situation and available resources.
- **Levelling Up and Skill Allocation**: The player invests points into skills and upgrades their character's abilities to improve combat effectiveness or unlock new features.
- **Problem-Solving through Dialogue**: The player engages in dialogue with NPCs to gather information, negotiate outcomes or progress through quests.
- **Navigation and Wayfinding**: The player navigates the game world, including using maps, quest markers and environmental cues to find their way to objectives.
- **Observation and Awareness**: The player pays attention to visual and auditory cues in the environment, such as enemy movements, objective markers and quest-related items.

3.1.4 Discussion

Once we gather and analyse the information derived from the experiment, we critically discuss the results from the perspective of game AI development. We cluster the main challenges we encounter and speculate whether current technology can tackle them or not. Then, we discuss how an AI can acquire the information necessary to evaluate and make decisions. We go further and check which game design patterns constitute a challenge for AI. Finally, we report on limitations that we can encounter in our methodology. While the framework generalises the steps and evaluations necessary to complete a game objective (and find a new one) to a high degree, it is functional in covering the main activities involved in meaningful play for open-world games.

Main Challenges

From the present research, we deduce three main challenges that AI developers face in the field of open-world games. The first is world **exploration** and the complexity of the processes involved [145]. This includes game actions that, albeit solvable, require quite a lot of computational power, such as *navigation and wayfinding*, *curiosity-driven exploration* or *trial and error*. The second one is the strong reliance of open-world play

on **generalisation**. It includes all those processes that involve understanding and planning beyond current or past game sessions. Examples are *inventory management*, *levelling up and skill allocation* or *viability and desirability* evaluations. Finally, an overarching and subsequent component of the challenge is the need for **coordination** of all the different solutions under a consistent one. We explore these categories in detail in the next chapters.

Exploration

Complexity can arise from several game activities. However, the component that can potentially be the most challenging for AI is **exploration**, intended as the link between different goals (or quests in our test case). Exploration is an important activity in open-world games. It is required to understand the game environment (both the world and the interface) and to gather the necessary information for subsequent evaluations. Exploration involves, first of all, *navigation and wayfinding* to navigate the virtual world. Even though this first challenge is not completely solved by AI systems, research in the field is proceeding with optimism. We can already cite practical attempts [146] [147] and more theoretical studies of human movement in games [148]. It also requires an understanding of the user interface. In this case, AI research has been sparse. Without the ability to understand the game interface system (or a hard-coded knowledge of it), an AI might not have access to certain information (e.g., journals or maps). Exploration also requires more than just roaming around the game world. AI agents need to be able to identify possible points of interest. In this regard, research is quickly ramping up in recent years with AI systems being able to perform curiosity-driven exploration [149] [150]. However, exploration in general is a component with which an AI would probably struggle.

Generalisation

Other challenges can arise from evaluations that require players to confront themselves with the environment in a comprehensive way. In this case, the two main ones would be **viability** and **desirability**. When we are talking about viability evaluation, we are not talking about an insurmountable obstacle, but a difficult-to-generalise one. Evaluating viability requires understanding the position of the player in relation to the game. This entails a good understanding of the game state and analysing previous experiences. An AI can arguably be able to perform this evaluation, for example,

by analysing the difficulty level of the environment around the selected task location. This is not a new challenge for AI, it is being extensively used to proficiently manage difficulty levels [151] [152] [153]. However, it covers only a part (albeit an important one) of what a viability evaluation entails. For example, tasks with specific prerequisites might be challenging to tackle for AI agents that would need to learn how to retrieve this type of information. This would make most solutions perhaps effective, but hardly generalisable. Desirability evaluations involve similar processes to viability evaluation in that they need an understanding of the game story's requirements. In this evaluation, the player needs to consider a) whether completing the task at hand is going to be necessary to proceed in the game and b) whether the task provides benefits that, although not necessary, make the outcome of future viability evaluations more likely positive. NPCs able to evaluate a task's benefits/cost ratio are already quite common in video games. Most strategy games already include this feature to make diplomatic decisions [154] [155]. These behaviours tend to be mostly scripted and led by specific numerical comparisons (e.g., the relative power balance between the NPC and other players/NPCs). Nevertheless, it is important to note that strategy games' decision-making processes were considered (too) challenging for AI years ago [156]. Then, AlphaStar managed to proficiently play Starcraft II [28] much earlier than expected. It is possible that similar techniques could be ultimately adapted to open-world games' risk-reward analyses.

Coordination

Coordination is a subsequent challenge to the aforementioned ones; it entails the merging of potential solutions to those problems into one. On the one hand, this can be intended as proficiently using all the gathered information and evaluations for decision-making.

On the other hand, this has more complex ramifications related to the ability to coordinate different decisions based on a consistent persona. This is highly important for desirability evaluations. For human players, desirability is also related to the player's personal goal (e.g., considering whether a task makes sense for their character's role). Adherence to the character's persona and story might seem exclusively relevant for human players. However, many open-world games strongly link character development (in terms of skills or levels) to their choices in the story [157]. Developing an AI that can consistently mimic the psychology of a consistent persona is a significant challenge. In this case, we are involving the character's believability, which, in the

case of AI agents, could be evaluated with its ability to pass Turing's test [158]. Of course, the validity of Turing's test has been often debated [159], but it is undeniable that the problem it describes is still valid nowadays.

Categories of Information and AI

We already mentioned that humans utilise different categories of information for decision-making in open-world games. However, not all humans have the same access to this information. For example, few of our participants have little experience with video games in general and almost none with open-world games. This translates into a lack of knowledge in the second category of information (experience from video games). Alternatively, other players miss certain tutorials that provide in-game types of information. However, regardless of these possible shortcomings, all players show the capacity to interact with the environment, basing themselves on real-world interactions. Therefore, while they might need some moments to understand that only highlighted objects foster interactions, they all can follow a road for pathfinding or instinctively know what type of professions they can encounter in a medical bay. The situation is reversed in the case of AI. While in-game information can be introduced easily, information derived from previous experience (both with video games and the real world) is much more difficult to attain. Currently, all the knowledge intelligent systems use in video games is usually obtained by training within specific contexts. While we can program an AI with specific "hard-coded" knowledge, we would hardly be able to cover all the necessary information to flexibly adapt to any open-world game situation.

Open-world Game Design Patterns and AI

In this final section of the discussion, we analyse the details of the typical design patterns that are usually involved in open-world games. In this regard, the matter is not one of feasibility but one of complexity. While some of the patterns reported can by themselves constitute moderate challenges for AI (e.g., *navigation and wayfinding*), the real difficulty is to coordinate all the necessary expertise under one intelligent agent. Certain elements, such as *interacting with NPCs* or *problem-solving through dialogue*, could be partially tackled proficiently by modern LLMs. However, these patterns are interrelated and heavily influenced by each other. This is the case as well for *problem-solving through dialogue* and *levelling up and skill allocation*. The latter also requires formulating an overall strategy about how to interact with the game. This is the case

for *inventory management* as well. In general, the capacity to plan and estimate the future relevance of skills and items is one of the hurdles for intelligent systems.

Limitations

Our study comes with a set of limitations that should be kept in consideration when evaluating the results. The first limitation is the use of strictly qualitative methods. This choice is taken, diverging from usual research in the field to provide a human-based perspective. As a result, the subjectivity of the methodology represents a strength but also a weakness. The sample size is also relatively small, even though the talk-aloud protocol does provide a large amount of information for each player. Finally, the use of a GPT-3.5 model as a clustering tool opens up issues of replicability. Like other stochastic tools, this is an ongoing problem in research involving the use of AI systems.

3.1.5 Future Research

Arguably, we can envision progress related to the issues of complexity and coordination. Examples come from the field of distributed machine learning [160] and artificial general intelligence (AGI) [161]. The challenge of generalisation, however, remains. Our research argues that this is currently the major gap between AI and human players, and it becomes evident in contexts where objectives are fuzzy, such as open-world games. Moreover, our research suggests that future developments in the field should pay close attention to human resort to previous experiences and knowledge in play. The ability to understand and artificially reproduce these processes is vital to further research in the field.

3.1.6 Section Summary

In the section above, we tackle RQ4;. First, we show how video games have historically presented themselves as challenges for AI technologies and how they guided and marked their development. Additionally, we speculate on future work in the field, proposing a new challenge and a new way to analyse it in terms of human gameplay. In this regard, we argue that the way humans play open-world games can be of great help to further improve AI capabilities and flexibility in tackling complex problems. The coming section will deal exactly with the relationship between humans and AI. With AI technology advancing at a rapid pace, researchers must study how humans relate to AI and vice versa. In our thesis, we argue that video games can be meeting

points between the two. The next section exemplifies how we can study the potential impact of AI within video game contexts and its relevance to the field.

3.2 The Effect of LLM-Based NPC Emotional States on Player Emotions: An Analysis of Interactive Game Play

The introduction of LLMs brought new possibilities in the field of video games, including the possibility of developing highly adaptable NPCs that can collaborate and interact with players in ways that were not possible before [162, 163]. The challenge lies in leveraging LLM-enhanced NPCs to enrich player experiences without sacrificing flexibility.

While these techniques allow for new ways to create immersive and engaging experiences for players, their inherent flexibility and novelty make it unclear how interacting with them may affect players on an emotional level. The main aim of this study is to elucidate how the interaction with LLM-enhanced NPCs can impact the emotions of the player. Emotionally adaptive NPCs represent a shift from traditional scripted interactions. Unlike static NPCs, LLM-driven characters can respond with dynamic emotions, making interactions feel more lifelike, creating, for example, the illusion of organic social interactions. Specifically, this study explores the impact of LLM NPCs that have been instructed to behave according to a predefined emotional state on players.

In order to elicit an observable effect, we design an experimental context that fosters players' emotional engagement. To do so, we stimulate interactions with NPCs in the form of dialogues, and we measure their effects through an information game called 'Black Stories'. This game engages players to solve a mystery through discussion with NPCs. We use assets to create LLM NPCs capable of conversation. These assets allow us to design AI agents with custom behaviour, emotional state, and knowledge. For data analysis, the emotional evaluation of player conversation is done through a pre-trained RoBERTa [164] model trained on the GoEmotions dataset [165]. Automatic prompt engineering techniques for small LLMs are also evaluated for this purpose. This approach allows us to continuously detect the emotional state of players at different phases of the game without disrupting the game loop.

This section presents the following contributions:

- We implement a version of the 'black stories' game where players interact with emotionally conditioned LLM NPCs to solve a mystery. This game constitutes a framework to observe and measure how the interaction with LLM agents impacts player emotions.
- We use 'black stories' to show how LLM NPCs initialised with specific emotions are capable of affecting the player. We analyse the effect of NPCs with different emotional conditionings on players' emotional states through measurements and comparisons.
- We compare two methods to measure player emotions from the text; One based on a RoBERTa model and one on prompting a 7B-parameter LLM.

The structure of the paper is as follows. In Section 3.2.1, we review recent work on LLMs and emotion recognition. The subsequent section (3.2.2) describes the game development process. Section 3.2.4 outlines our data collection approach and the models used for emotional measurement. Our findings are presented in Section 3.2.5, accompanied by emotion distribution plots. In the discussion section, we reflect on insights gained and address limitations. Finally, we conclude with an overview and potential future contributions.

3.2.1 Related work

LLMs: Recent advancements in natural language processing (NLP) have shown the potential of LLMs for games[166, 167] with applications including procedural content generation [168], game design [169], and game user research. Multimodal extensions to LLMs have also been considered to create NPCs capable of acting in complex game environments by leveraging data from other modalities besides text, such as visuals and sound[170].

Our work extends on the mixed-initiative gameplay literature, focusing on the usage of LLMs for gameplay where both the player and the LLM agents interact with one another. Multiple studies have explored the possibilities for such interactions in text-based games.[171, 172] use LLM agents to assist in story creation games, with the latter analysing how the cooperativeness and creativity of LLMs impact creativity in children. LLMs have also been used as dungeon master assistants for tabletop role-playing games [173].

While the technology has shown promise in different avenues, it remains unclear how design choices in the creation of LLM agents affect players. This work focuses

on measuring the impact of the emotional state of LLM agents on players in a mixed-initiative setting.

Emotion Recognition: Research in emotion recognition within human conversation has explored the impact of video games on emotional experiences, applying *affect theory* to game studies [174]. Video game exposure has shown positive effects on prosocial behaviours and thoughts, motivating investigations into NPCs' social effects on players through communication [175]. Growing interest has been shown in using LLMs to produce more human-like NPCs in video games since LLMs first appeared [167]. Models like RoBERTa, based on BERT, have been effective for understanding player emotions during NPC interactions [164], often leveraging datasets like GoEmotions, which provides labelled emotions for English Reddit comments [165]. LLMs have also been used for sentiment analysis for game design assistance, with OPT-175B being highly effective in sentiment analysis [176].

Understanding emotions in language is crucial since text serves as the primary communication channel between humans and computers [177]. Analysing emotions from text rather than other tests, such as questionnaires, provides a fine-grained signal of emotional state without disrupting play. The method is also safe from self-reporting biases, which can affect the validity of results [178]. Several works have focused on the detection of emotions in text, suggesting different approaches and demonstrating high accuracy [179, 180]. However, improving these systems' precision and robustness continues to be challenging [180]. Nonetheless, sentiment analysis has shown lots of promise, highlighting its usefulness and possible growth in the video game industry [181].

3.2.2 Black stories

'Black stories' is a puzzle game, inspired by the work of author Holger Bösch, that challenges players to solve complex and suspenseful scenarios or mysteries. These scenarios typically involve mysterious events, crimes, or puzzles that players must unravel through critical thinking and decision-making.

In each game, there are at least two participants: one Game Master and one or more players. The Game Master initiates the game by choosing a story from a set of mysteries. Then the Game Master presents a part of that story to the players. The players must uncover how this fraction of the story evolved through yes-no questions. The game concludes once the players successfully uncover some key details of the story. For example, the Game Master may explain that a character got stranded in a

desert, and the player would have to guess what course of action may have led to that situation.

For our purposes, we made a few changes to the traditional rules of the game. The player has a limit of 10 questions to ask the Game Master. If the player is able to solve the mystery within the 10 questions, they win; otherwise, they lose.

3.2.3 Video Game Development

In this subsection, we briefly explain and motivate the decisions taken during the development of the game. The video game is developed in Unity using imported assets for rooms and furniture. 3D models and animations from Mixamo, developed by Corazza et al., are used to make the environment more immersive.

We use the Inworld AI tool, created by Gelfenbeyn, to implement LLM-enhanced NPCs. This tool allows the designing of NPCs, capable of conversing with the player and adapting their behaviours through this process. These NPCs can be customised in terms of their knowledge, identity, and personality, making them the ideal choice for our study.

In our implementation of the 'black stories' game, two NPCs engage with the player in solving the mystery: an assistant called Amy and the Game Master. Personality traits and moods were set for each character using emotional and personality sliders, as well as text prompts. The sliders indicate the tendency of a character to express a given sentiment or personality trait during conversations.

Figure 3.4: The player interacting with Amy

These were curated to create seven different NPC profiles. One NPC profile is created for the Game Master and six for the assistant NPC, Amy. Her purpose is to assist by answering an unlimited number of questions about the story and replying with full answers based on both the player's theories and built-in intelligence. While all other profiles are neutral in emotions and personality, three of Amy's profiles have

been designed to exhibit a specific polarised behaviour. We refer to those profiles as Happy Amy, Sad Amy and Angry Amy. Neutral profiles will be explained explicitly in the next subsection. All of them are used to study how different sentiments by another party affect the player. The slider values set for these characters are reported in Table 3.1 for reproducibility purposes. The range of values is 0 to 8.

Table 3.1: Personality Sliders for Different Types of Teammates (Amy)

Personality Sliders (values: 0 to 8)				
Teammate NPCs	Neutral	Happy	Sad	Angry
Sadness to Joy	4	8	1	4
Negative to Positive	4	8	1	4
Anger to Fear	4	4	8	1
Discuss to Trust	4	7	1	1
Insecure to Confident	4	7	1	4
Aggressive to Peaceful	4	8	4	1
Cautious to Open slider	4	4	4	1

Game flow

Upon launching the game, the players are greeted by a short introduction that informs them about the rules of the game. The game is divided into three different sections. At the beginning of each section, an NPC starts the conversation and invites the player to interact with them. Amy is given different emotional states in each section. This approach is crucial for the objectives of our study. Figure 3.5 presents some details about these three phases and Amy's corresponding profile.

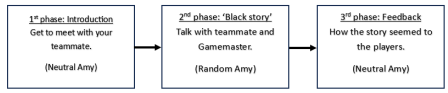

Figure 3.5: The phases of the game.

Phase 1: Introduction. This phase invites the player to engage with Amy and create an initial bond. During this phase, players have an opportunity to talk to Amy by discussing their hobbies and interests. This phase serves two main goals: to help the player become familiar with navigating through the game environment and to assess the player's initial emotions. For these reasons, Amy is set to be Neutral. This section concludes after receiving five responses from Amy.

Phase 2: Solving the mystery. In the second phase, players immerse themselves

in the co-op game 'black stories', where they cooperate with Amy to solve the mystery given by the Game Master. Amy is randomly assigned a different emotional profile out of the following list: neutral, angry, happy, or sad. The player can also interact with the Game Master in this section. The Game Master first introduces the mystery that the player and Amy have to solve. After doing so, the player gets the choice to interact either with Amy or the Game Master. The player interacts with Amy for theory crafting and with the Game Master for asking direct, close-ended questions about the mystery. This section ends after the player has asked 10 questions to the Game Master.

Phase 3: Feedback on the mystery game. The third and final stage focuses on obtaining feedback by having the player reflect on the mystery and game experience. In this phase, the player discusses the story with Amy and shares their thoughts about the story and the adventure in general. The emotional trait set for Amy in this section is Neutral. The purpose is to measure the sentiments of the player after the game in order to compare them with those from the previous phases of the game. This phase is concluded once Amy has responded to the player five times.

After completing Phase 3, the player will be taken to the game's final screen. This screen displays the solution to the mystery and thanks the players for playing the game.

3.2.4 Experimental analysis

This section discusses the gathering and analysis of conversation data, followed by a comparison between the use of a pre-trained RoBERTa model and prompting LLMs for emotion recognition. The RoBERTa model is then applied to extract emotions from conversation data.

Data collection and preprocessing

We record all conversation text between NPCs and players, together with Amy's corresponding behaviour settings. This allows us to inspect the interactions between the player and the NPCs, and it provides us with a solid foundation on which we can do further analysis.

The game is shared for public use on the platform itch.io. The passcode 'blue' is needed to enter the game. 19 people played out the game in full, with ages ranging between 22 and 39. The data from players who do not play all phases of the game are considered incomplete and hence are discarded.

To evaluate the player's emotional response during conversations, we analyse the sentences they write. For tracking changes in emotional state over time within a phase, we adopt a method that normalises data across players with varying text lengths: each player's concatenated sentences are split into the same number of sections. Each section is used as input to the language models to obtain emotional scores. This yields a sequence of emotional scores for each player, which is then used to compare emotional changes within a given phase.

Model on emotion recognition

To determine the best method for extracting sentiments from conversation data, we compare two approaches: a pre-trained RoBERTa model and prompting LLMs. In our comparison, we compare the performance of these methods on the GoEmotions dataset. This step helps us identify which model is more effective at emotion extraction. Once we establish the superior method, we proceed to use it for our emotion analysis in conversation data.

RoBERTa model vs LLM prompting Briefly, the RoBERTa model that we use is a pre-trained model found in the HuggingFace library. The model is a BERT-based model [164] that is fine-tuned on the GoEmotions dataset [165]. We also test an LLM prompting-based method, called Llama 2, which prompts an LLM to output which emotions are detected in a given text. We then compare this prediction to the ground truth to compute performance metrics. To discover optimal prompts, we employ an iterative approach inspired by Automatic Prompt Engineer (APE) [182]. We work with a set of 10 human-generated prompts and evaluate their performance on the GoEmotions dataset.

Emotion extraction from conversation data To extract emotions from conversation data, we use the optimal method found in Section 3.2.4. For the first and third phases of the game, we use the conversation data as it is to extract affections. This yields an aggregated affect score for each phase. For the second phase of the game, we use the conversation data to extract emotions throughout gameplay. Given that conversation data varies in length among players, we standardise the extraction process by selecting a fixed number of messages per player, proportional to their total number of messages. This allows us to segment the conversation data uniformly for each player and extract emotion scores from each segment. Consequently, we obtain

a sequence of emotion scores for each player, facilitating the analysis of emotional evolution throughout gameplay.

3.2.5 Results

In this section, we report the main insights extracted from our experiment, dividing this into four subsections: Optimal LLM prompts, the RoBERTa model vs LLM prompting, Emotion extraction for different phases of the game, and the impact observed of NPC emotions.

Optimal LLM prompts

We find the optimal LLM prompts using the iterative process described in Section 3.2.4. The optimal prompts that were generated achieved an F1 score of 0.117 and a recall score of 0.538, respectively.

RoBERTa model vs LLM prompting

We evaluate the RoBERTa model and LLM prompting-based methods on the GoEmotions test split. We observe that the LLM prompting-based methods perform worse than the RoBERTa models, being unable to achieve high scores across all evaluated metrics with either method. The RoBERTa model with optimised thresholds performs best, achieving the highest F1 scores and Matthews correlation coefficient (MCC). Thus, the pre-trained RoBERTa model with optimised thresholds is the best method for extracting sentiments from text and is used in the following sections.

Emotion extraction for different phases of the game

We extract emotions from the conversation data for the different phases of the game and compare the aggregate results. Figure 3.6 shows the results of this comparison on the most affected emotions. Results show that emotions differ between game phases. All phases have high scores for Neutral emotions. Gratitude, joy, and approval are present in the first phase, but not as much in the later phases. Curiosity is high in the first two phases, and confusion increases steadily across all phases. The right figure shows how emotions change during the middle phase of the game over time. We observe that the most common emotions stay mostly constant throughout time when looking at their scores averaged across all gaming sessions.

	Phase 1	Phase 2	Phase 3
Admiration	5.424	3.682	5.090
Anger	0.180	0.388	0.382
Approval	15.446	3.544	10.086
Confusion	5.196	20.739	14.301
Curiosity	17.068	32.030	7.457
Excitement	4.400	0.406	0.640
Gratitude	2.761	2.067	2.724
Joy	4.010	0.228	1.594
Neutral	37.391	47.829	41.681
Sadness	0.174	1.086	3.188

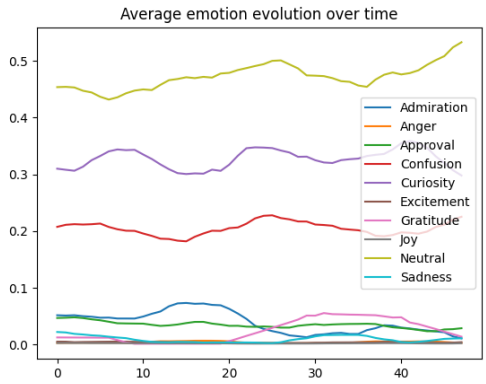

Figure 3.6: Left: Heatmap of the average emotion scores for the different phases of the game. Right: Average emotion scores during phase 2 of the game over time. Scores for a given timestep are computed on segments of the conversation as described in section 3.2.4.

To evaluate the game's impact on player emotional responses, we analyse the results presented in Figure 3.7. This assessment reveals how different phases of the game affect specific emotions. Some emotions show consistent scores across phases, while others exhibit changes. The overall impact is determined by comparing emotional scores between the first and last phases of the game. Players generally experienced increased confusion, disapproval, disappointment, and sadness, alongside reduced curiosity, love, approval, and excitement post-game. By separating the overall impact into gameplay and outro phases, we can identify which sections influenced specific emotions. Gameplay notably contributed to increased confusion and reductions in approval, excitement, joy, and love.

Impact of NPC emotions

In this experiment, the impact of different personalities of Amy is evaluated. Figure 3.8 shows the average emotion scores for each profile during gameplay on the most affected emotions. The scores are computed on messages from the player using the adapted windowing previously described.

The average scores show the previous main emotions: confusion, curiosity, and neutrality. The scores that were normalised with respect to the average emotions show that the different agents induced different emotions in the player's conversations. Neutral Amy reduced curiosity levels. Sad Amy increased neutrality and curiosity and

	Total impact	Phase 2 impact	Phase 3 impact
Admiration	-0.335	-1.742	1.408
Amusement	-0.003	-0.298	0.295
Anger	0.202	0.208	-0.007
Annoyance	2.527	0.724	1.803
Approval	-5.360	-11.901	6.541
Caring	-0.522	-0.686	0.164
Confusion	9.104	15.543	-6.438
Curiosity	-9.611	14.963	-24.574
Desire	1.852	-0.287	2.139
Disappointment	3.960	0.665	3.295
Disapproval	8.057	0.594	7.463
Disgust	0.101	0.043	0.059
Embarrassment	0.100	0.055	0.046
Excitement	-3.761	-3.995	0.234

	Total impact	Phase 2 impact	Phase 3 impact
Fear	-0.101	-0.020	-0.082
Gratitude	-0.037	-0.694	0.657
Grief	0.104	0.030	0.074
Joy	-2.416	-3.782	1.366
Love	-11.166	-13.245	2.078
Nervousness	0.251	0.029	0.222
Neutral	4.291	10.438	-6.148
Optimism	0.104	-0.428	0.532
Pride	-0.028	-0.073	0.046
Realization	-0.083	-1.188	1.104
Relief	0.009	-0.160	0.169
Remorse	0.363	0.854	-0.491
Sadness	3.014	0.911	2.103
Surprise	0.204	0.455	-0.251

Figure 3.7: Heatmap of the emotional impact of different game phases for all measured emotions. *Total impact* indicates the score difference between the last and first phases. *Phase 2 impact* indicates the score difference between the second and first phases, and *Phase 3 impact* indicates the score difference between the last phase and the second phase.

reduced gratitude. Angry Amy increased gratitude and reduced neutrality. Happy Amy increased admiration and curiosity, and reduced confusion and gratitude.

The impact of the four states of Amy is also explored in greater detail in Figure 3.9. The plots show the average impact of the different behaviours on the most affected emotions over time during gameplay. While Happy Amy had higher-than-average admiration scores, we observe that these scores are only high at the beginning of the conversation and decrease to below-average scores by the end of gameplay. Angry Amy also shows a peak of gratitude scores in the ending sections of gameplay. Sad Amy induces higher than average curiosity scores, and neutral Amy induces average responses, with the exception of a neutrality peak by the end of gameplay.

3.2.6 Discussion

This study explored how a game and its NPCs, characterised by distinct emotional states using LLMs, influence player sentiments. Our findings reveal emotional impact across game phases, particularly in increasing curiosity and confusion during gameplay, notably emphasised in Phase 2's mystery element. Our results show that NPCs and their emotional state play a pivotal role in shaping player emotions, with differ-

	Angry	Neutral	Happy	Sad
Admiration	2.670	2.163	6.181	1.900
Anger	0.702	0.248	0.413	0.298
Approval	4.287	4.652	3.181	1.434
Confusion	21.350	22.084	19.402	20.557
Curiosity	30.972	26.839	34.210	38.384
Excitement	0.474	0.285	0.502	0.355
Gratitude	9.828	0.143	1.229	0.111
Joy	0.276	0.193	0.260	0.175
Neutral	42.481	49.341	47.633	50.613
Sadness	1.406	1.740	0.499	0.826

	Angry	Neutral	Happy	Sad
Admiration	-0.559	-1.065	2.953	-1.329
Anger	0.287	-0.168	-0.002	-0.118
Approval	0.899	1.264	-0.208	-1.955
Confusion	0.502	1.236	-1.447	-0.291
Curiosity	-1.629	-5.762	1.609	5.782
Excitement	0.070	-0.119	0.098	-0.049
Gratitude	7.000	-2.685	-1.599	-2.717
Joy	0.050	-0.033	0.034	-0.051
Neutral	-5.036	1.824	0.116	3.096
Sadness	0.288	0.622	-0.619	-0.291

Figure 3.8: Player emotion scores over different Amy emotional states. The left figure shows average scores, and the right figure shows how the scores differ from the average.

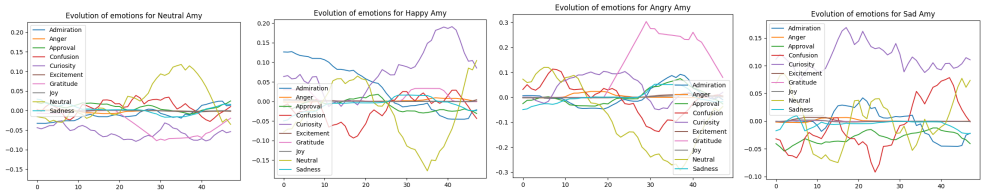

Figure 3.9: Computed scores of player emotions over different Amy profiles during gameplay. Each plot shows the normalised scores on the most affected emotions over time. Scores for a given timestep are computed on segments of the conversation as described in section 3.2.4 and normalised to show how they deviate from the average scores across players and emotional states of Amy.

ent characters eliciting different feelings such as gratitude or curiosity. For instance, 'Happy Amy' initially inspires admiration, which diminishes over time. This not only shows that the emotional state of NPCs affects the player but also illustrates how their influence evolves during gameplay. Through our analysis, we identify how specific game sections affect players' emotions and uncover lingering post-game emotional responses induced by gameplay.

The experiments reveal that some emotional states induce reactions in the player which can be unexpected. An example is the case of 'Angry Amy', which often evoked responses of gratitude from players. While this response may seem contradictory at first, literature in psychology linking anger and gratitude exists and could explain our observations [183]. Moreover, the increase in detected gratitude may be a sign of pacifying behaviour from the players, as an attempt to foster cooperation when faced with a more aggressive NPC assistant. It is also possible that the increased

difficulty in cooperating with an angry assistant would result in higher satisfaction when progressing through the game.

The results regarding 'Sad Amy' show that players display increased empathy and curiosity with her. These results seem to indicate that the agent and its emotional state are capable of inducing a desire to help the players, which might motivate them to immerse themselves more deeply in the game. The evolving emotional responses to different NPCs underscore the dynamic nature of player-NPC interactions. By normalising emotional responses with respect to the average, we demonstrate that different NPC personalities distinctly impact player emotions. This shows that the impact on the players' emotions is not only caused by the game, but specifically by the emotional state of the NPC.

In summary, our study investigates the complex and changing emotional landscape that players navigate in a game, shaped by both the game's design and its NPCs. These insights provide valuable information for designing games that better engage and resonate with players emotionally.

However, the game comes with its limitations. Firstly, it offers only a single mystery to solve, which may have constrained player interactions and responses by limiting variety. Secondly, implementing progress tracking would provide an interesting signal to consider when evaluating the impact of the emotional state of NPCs. While the research primarily focused on AI, the service used to provide player interactions with Agents also posed its own challenges. The AI, operating separately from the game, exhibited unexpected behaviours, leading to varied experiences for players. This unpredictability sometimes extends to the AI incorrectly presenting game rules, such as providing inaccurate answers about the mystery when interacting with the Game Master. Since characters could not interact with each other, they often improvised information due to their limited awareness. These quirks could potentially influence player experiences and interactions, impacting the research outcomes. Despite these challenges, the results remain compelling, highlighting how the unpredictable nature of LLM usage affects player experience, a critical factor for assessment.

3.2.7 Conclusion and Limitations

Through this project, we developed a game to measure the impact of the emotional state of LLM-based NPCs on player emotions. We designed a collaborative game called 'black stories' where players interact with LLM-based NPCs to solve a mystery. We evaluated various methods for emotion extraction from text and utilised a pre-

trained RoBERTa model for optimal emotion analysis. Our study reveals that player emotions vary across game phases and are influenced differently by the emotional profile of the LLM-based NPCs they interact with. We show that specific phases are capable of increasing the curiosity and confusion of players, while others show lower emotional engagement. We demonstrate that different emotional states of the NPCs induce different responses in players, and show that this effect can change throughout gameplay.

For future research, including a wider range of stories would allow us to validate our findings. Incorporating different story genres will help determine whether story type influences player emotions alongside dialogue. Additionally, while we uncovered the impact of specific emotional states, they do not fully represent the range of possible emotional states that an NPC can be designed to have. Expanding the research on other emotional settings could further the understanding of how LLM NPCs can impact a player. For example, this could be done by implementing emotion recognition techniques. Involving more participants and applying further statistical testing would also be recommended in order to uncover possible correlations. Finally, it would be valuable to explore practical implications, focusing on the ethical repercussions of potential implementations of this research. As AI becomes more integrated into gaming, ethical concerns arise that need to be carefully studied; AI-generated narratives can be inherently biased, while player data collection raises privacy concerns. Moreover, the impact of potential AI alignment on user experience requires further research.

3.2.8 Section Summary

The section above aims to provide an exemplary answer for RQ5;. We first explore the new developments in the field of AI and how they relate to video games. Subsequently, our experiment targets directly players, trying to elicit and measure emotional responses arising from their interaction with AI. The results show potential for emotional engagement but also emotional manipulation, which depicts a very interesting context. This indicates that video games are indeed a point of interaction between humans and AI and are important tools to further research this relationship. Techniques similar to the one we used will probably become more and more relevant with future technological advancements in the field. In the next section, we go beyond the digital boundaries of computers and see how games can be used to bring AI technologies into the physical world. In doing so, we show the potential for games to allow for human-computer interactions even outside digital contexts. An additional goal for the next

section is to propose a taxonomy as a tool to reflect upon the role of computational resources (including AI) in games. Specifically, we argue that the taxonomic system we propose can help designers in the development of hybrid games with AI components.

3.3 Towards a Taxonomy of AI in Hybrid Board Games

Over the past years, board games have been rising in popularity [184] in parallel to video games. Rather than standing in competition to one another, video games and board games offer different kinds of experiences that are both in demand. Naturally, this also creates more interest in game systems that borrow from both modalities. The overlap between video games, a term that we use synonymously with 'computer games' and 'digital games' in this paper, and physical games is often referred to as 'hybrid board games'. Hybrid board games can be understood as part of a wider range of 'hybrid games' that generally involve multiple and different types of media without necessarily being defined by the involvement of analogue and digital game elements [185]. At the same time, in the area of video games, the importance of AI is steadily rising, as the necessary technology becomes increasingly more capable of sophisticated decision-making and interpreting complex game states. This, in turn, allows for the creation of novel gameplay elements, as well as the development of systems that aid in the design and evaluation of video games [186, 24]. This trend is less pronounced in hybrid board games, where the use of AI appears to remain more rudimentary.

In this work, we present the first steps towards a taxonomy of AI in the area of hybrid board games with the purpose of aiding the research and development of AI that can support such games. We see the creation of a taxonomy as a catalyst for generating new ideas by structuring existing knowledge and, perhaps even more importantly, emphasising areas that lack either practical or theoretical knowledge. Finding such design spaces can highlight interesting opportunities for future work that would otherwise remain unexplored. Our efforts should therefore be understood as a call for action to strengthen the presented structure through further critical discourse and empirical investigations.

The following section provides our working definitions of hybrid board games, as well as what can be considered 'AI' in the context of such games. Section 3.3.2 outlines

a proposed taxonomy through different possible dimensions from which to attempt a differentiation of AI in hybrid board games. We conclude the paper with a discussion of the presented dimensions through illustrative examples.

3.3.1 Working Definitions

Before attempting to map out a taxonomy of AI in hybrid board games, we need to establish a definitional basis for the involved aspects. What forms of AI should be considered, and what do we mean when we talk about 'hybrid board games'? The focus here is less on arriving at indisputable demarcations (requiring considerably more argumentative writing space) than on outlining working definitions that provide a structure for further discussion.

In the context of this paper, we understand **hybrid board games as games that combine intentionally designed digital and physical modalities to create a game experience for players within the boundaries of a defined physical space** [185, 187]. The underlying games may be created for entertainment purposes, or fulfil additional purposes, such as to train players in a given task (often referred to as 'serious' games [188]). Under this definition, we exclude 'gamification', which is the use of individual game mechanics or aesthetics in otherwise non-gaming circumstances [9]. Our working definition further excludes games that lack physical or analogue artefacts that are explicitly designed for the purpose of facilitating a game session. This distinction is inherent in the term 'board' within 'hybrid board games'. Augmented reality games such as *Pokémon GO* [189] may indeed involve the physical environment, but do not define specific game spaces and do not contain physical artefacts that are intentionally designed. The digital domain of the game adapts to the physical domain, while the reverse does not occur.

Augmented reality games or mixed reality games can certainly be described as hybrid games, and the involvement of other domains might create hybrid games that are not defined by the use and interaction of both physical and digital components [185]. Likewise, any efforts to build a taxonomy may yield valuable insights for hybrid games of all sorts. However, we do see value in focusing on a specific sub-field, i.e. hybrid 'board' games, as it is also likely that some taxonomic dimensions that we will discuss do in fact not map to all hybrid games.

On the other hand, we consider the word 'board' a linguistic anchor that hints more at the involvement of physical artefacts, defined space and gameplay traditions than at the existence of a board in a strict sense. Card games or dice games, for

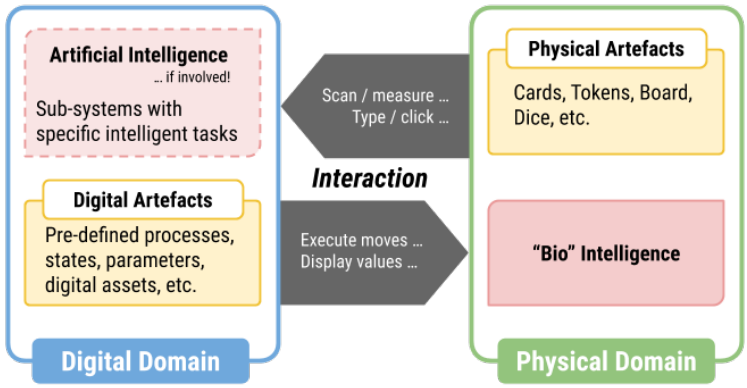

Figure 3.10: Illustration of high-level components involved in a hybrid board game. Both domains involve intentionally designed artefacts and interaction with the other domain. Typically, hybrid board games involve human, or 'bio' intelligence, but may also involve several AI sub-systems (which tend to be specific to the task)

example, may lack a physical board, but do involve intentionally designed physical artefacts and spaces. It would therefore perhaps be more accurate to talk about hybrid 'tabletop' games, as most of these games are traditionally played on a shared table. However, it should be noted that 'hybrid board games' is already an established and somewhat widely-used term that indeed appears to include physical games that lack a board. This is also where an excessive fragmentation of implementations is perhaps less useful in mapping out a potential design space.

In terms of what forms of AI should be considered for the taxonomic structure, we build on recent work in the field of game AI, which is focused on the use of AI for game purposes [186]. As a rough working definition, we are interested in mapping **any involvement of a computational system into a decision-making process that is part of a hybrid board game.** This, just like for game AI, includes decision-making processes before or after the game, as well as decisions that are more artistic than part of a game mechanic.

Figure 3.10 illustrates the conceptual components of a hybrid board game, as we understand it through the outlined working definitions. Components are separated between the digital and physical domains, both of which include artefacts that are intentionally designed to be part of the gameplay. Both domains further involve some degree of interaction with the other domain. The physical domain necessarily involves one or more intelligent entities[1] that usually take the form of human players (although

[1] ignoring the more philosophical musings on the concept of zero-player games [190]

the involvement of animal players is a possibility that fits perfectly well in this model). When AI is involved during the game session, the digital domain also involves one (or more) intelligent 'entities'. In contrast to human players, AI entities may not necessarily be featured as individualised agents, but can instead be constructed as compartmentalised sub-systems. Human players are generally capable of carrying out a wide range of decision-making tasks that are very different from one another. AI systems are more likely to be designed to fulfil specific tasks, thus leading to a number of systems that can be at play in parallel, even if they together only take control over a single game entity (if they represent an embodied agent at all).

We acknowledge that the presented working definitions likely leave questions open. For working towards a taxonomic structure, we consider this both a practical necessity and an opportunity for encouraging a broader discussion in an effort to better map out potential uses of AI in hybrid board games.

3.3.2 Taxonomic Lenses

In this section, we outline dimensions on which examples of AI in hybrid board games either already exist or could potentially exist. Each of these dimensions represents a 'lens' or perspective through which the involvement of AI in hybrid board games can be viewed and understood (see Table 3.2). The metaphor of different lenses follows a similar approach in efforts to outline the wide range of interrelated dimensions in the practice of game design [191]. It is important to note that we choose this metaphor in part because it reflects the fact that individual dimensions are not necessarily separated as definitively as is the case in other taxonomic models, such as the 'phylogenetic tree' or 'Linnaean taxonomy' in biology.

Throughout the following sub-sections, we use chess as a case study to illustrate how it can be (and has been) modified to act as a hybrid board game with AI involvement. The point here is, of course, not that chess is the most suitable game for such efforts. However, it provides a widely known game example that is useful for illustration purposes.

Embodiment

AI in games is perhaps most prominently represented by the involvement of intelligent agents that are embodied in some form. In video games, this embodiment happens in the digital domain, through 'bots' that compete with human players, or NPCs that give players the opportunity for diegetic interaction. Such agents also exist in hybrid

Taxonomic Lenses	→ Constituent Sub-dimensions
Embodiment	→ **Relationship** between agent(s) and players → **Believability** of interaction → **Amount** of agents
Physical Domain	→ **Awareness** of physical domain → **Interactivity** with physical domain
Temporal Domain	→ Temporal **involvement** within/outside a game session → Temporal **resolution**
Gameplay	→ **Centrality** to gameplay
Role	→ **Actor-Director** spectrum

Table 3.2: The left column lists the individual taxonomic lenses that are discussed in this paper. Each lens should be understood as an independent perspective on AI in hybrid board games. Each lens can be further deconstructed into constituent sub-dimensions. The table is not exhaustive and should be understood as a structural foundation.

board games, either as physical entities or as virtual entities with varying degrees of defined embodiment. Early chess computers might require players to carry out turns for a computational agent on their behalf, but they still act as a virtually embodied entity (i.e. attributing any game interactions to an 'enemy' or opponent, rather than responding to unattributed changes in a game environment).

One dimension that falls under agent embodiment is **the relationship between AI agents and players**. AI agents may act fully collaborative, fully competitive, or somewhere in between. This can extend to the expression of personalities through the way in which an AI agent plays. Competitive actions by a human player might trigger AI agents to respond in kind for the rest of a game session, thus giving the appearance of a resentful AI player. While the possibility for such behaviour depends in part on the underlying game, even fully competitive games can provide opportunities to display 'emotions', such as in the way that an agent responds to a loss (e.g., congratulating or antagonising). Many games do not necessarily feature a single, clearly superior strategy for competitive play, thus providing venues to express an agent's personality (e.g., through aggressive, risk-taking play). It is also worth noting that competitive play can originate fully from the rules of a game, without involving a model of competition in AI agents themselves.

The display of such 'emotions' ties into another sub-dimension: the **believability**

of any interaction with an AI agent. Believability of agents is closely connected to what kind of embodiment is given to them by the design of the game. If they are given similar gameplay possibilities as human players, an agent AI will likely face a higher degree of scrutiny by players as to what is or is not believable. Here it is important to highlight that in the context of hybrid board games, the high end of the believability spectrum is less about the perfect simulation of human behavior[2], and more about maintaining a player's suspension of disbelief.

Another sub-dimension is **the amount of embodied agents**. A game might involve multiple AI agents with very rudimentary decision-making that present an obstacle to other players simply by their existence. Such AI agents can be thought to have no relationship to the player at all, instead carrying out tasks without any consideration for other agents (human or otherwise).

Agents can also be classified according to their relative power compared to the player's. For example, we can have AI agents acting as opponents, limited by the same rules and driven by the same opportunities as the players, but it is also possible to involve agents with different levels of advantages or limitations in their gameplay. This can also be moderated by the game settings, making the match more or less challenging for the human players.

The possibility of multiple AI agents brings up another dimension that is part of embodied AI involvement: **the number of agents that are controlled by an AI**. An AI system might be embodied as a single entity (whether fully virtual or with a physical representation), or consist of multiple, potentially infinite, embodied agents. Mapping an AI on this spectrum is not necessarily straightforward. In the example of chess, one could argue that only two agents are involved, as it is played by two players moving pawns. On the other hand, the embodiment of each player within the game space can also be thought of as 16 agents that act through a hive mind. The question of how many agents are in a (hybrid board) game is thus dependent on whether the focus is on the actual embodiment or on the intelligence that controls these embodiments. A hybrid version of chess could indeed be realised with multiple AI 'minds' that share the control of their 16 embodied agents, such as by developing competing strategies internally before settling on an externalised action. This form of hidden multi-agent setup is indeed used to treat the most difficult game-playing AI tasks, such as beating professional human players in *StarCraft II* [192, 28].

[2]Although, clearly, any progress towards solving 'AI-complete' problems are likely beneficial for the task of creating believable agents.

Physical Domain

Given that we define AI as the involvement of computational systems with decision-making capabilities, we can expect any AI to have easy access to any digital data that is kept as part of a hybrid board game. Such data might originate in the digital domain, but still require physical modalities to inform human players. The most straightforward method is the involvement of additional devices such as smartphones or tablets to facilitate the communication between the AI and the physical environment. On the other hand, to register actions in the physical world and interact with it, a degree of physical awareness is required. The **dimension of awareness in the physical domain** thus describes to what extent a physical input or signal is digitalised. In addition to physical awareness, an AI can differ in the degree to which it is capable of acting in the physical domain. **This dimension can be considered the interactivity of an AI in the physical domain.** Much of the existing academic work on hybrid board games focuses on how this translation between physical and digital states can be implemented [193].

However, for the purpose of building a taxonomy, the question of how awareness and interactivity with the physical domain is achieved might not be as important as to which it is involved at all. It is difficult to imagine examples in which an AI requires no degree of physical awareness, nor any form of interactivity with the physical domain. Early chess computers would require human players to provide information about the physical world (i.e., pawn movement) and to carry out AI movements correctly. While it may seem that full automation of such actions is always beneficial for human players, there is also some evidence that leaving some 'house-keeping' tasks to human players may be desirable [194].

Temporal Domain

Another lens to look at AI in hybrid board games is to consider the temporal domain: when is AI involved in the larger context of a game session, and at what temporal resolution does it operate?

The dimension of temporal involvement, or when AI is involved, seems less suitable for framing as a continuous spectrum than as distinctive ordinal categories, involving AI either: (1) before a game session, (2) during a game session, or (3) after a game session. It is conceivable that an AI is involved in some or all of these temporal categories, but it is more likely that this would involve different AI systems that target specific tasks within such a category.

The involvement of AI during a game session is perhaps the most apparent implementation and is exemplified by any AI agent that plays 'with' or 'against' players in a game. However, a taxonomy of AI in hybrid board games should also account for the use of AI in the preparation of a game session or even in the (co-)creation of the overall game [195, 196]. AI agents can, for example, be created not to act as opponents during a game session, but to serve as test 'participants' as part of the game development process [197]. Given that game development is often an iterative process, information about a play session will frequently be fed back into the design of a game. As such, post-play involvement may transition somewhat seamlessly into pre-play involvement. For the purpose of establishing a taxonomy, we may argue that the interpretation of gameplay data is more closely related to post-play involvement, while acting on that interpretation to improve a game is closer to pre-play involvement. As with (partly) automated play testing, a feedback loop encompassing in-play AI as replacement for the player, post-play game analysis and a pre-play game design angle happens in (partly) automated game balancing [198].

Another dimension related to temporal events is the resolution at which time is 'experienced' or processed. On one end of the spectrum, actions can be expressed or perceived continuously in real-time. On the other side, actions and events may be regulated in discrete steps. This is not necessarily connected to the gameplay rules of a game. Taking chess as an example again, any moves take place in turns and can thus be said to happen in a discrete manner. However, an AI system could monitor the game state in real-time, using the idle time to consider possible moves, and immediately react to moves by the opponent as they occur. On the discrete side of this example, the same AI system could instead not have a concept of real-time and instead only evaluate game states after a specific event (e.g. when the opponent indicates that they have made their turn).

Gameplay

Understanding the involvement of AI through the lens of gameplay means to **establish how central an AI system or agent is to the game itself.** On one side of the spectrum, AI systems might be involved for convenience or aesthetic purposes, without having an impact on the way a game unfolds. This does not necessarily make the involvement less valuable for players, and might involve AI systems that are just as complex or even more so than those that are more central to the gameplay. An example can be found in computational systems that take care of board game 'chores', such as keeping track of game states [199].

On the other end of the spectrum, the involvement of AI might fundamentally shape the gameplay. This end of the spectrum is arguably harder to find among hybrid board games, as they often involve only incremental change over non-digital board games. However, returning once again to the example of chess, the involvement of an AI agent as an opponent can make it central to the gameplay. While early implementations of artificial chess opponents may have only provided a trivial challenge, they have long since become real training partners that can inspire novel strategies.

While creating AI agents that can substitute for human players presents interesting research and development challenges, there is largely untapped potential in hybrid board games that are built around the involvement of AI. Such games could extend the design space with implementations that go beyond substitution.

Role

The final lens we propose is the role AI has within a hybrid board game. The actor-director dimension positions an AI on a spectrum between carrying out very narrowly defined actions on the one side and directing all aspects of a game on the other.

This dimension is almost inseparably linked with how much information a computational system is given (or can access) about the state of the game, as well as the extent to which it is permitted to modify it. Systems that generate aesthetic assets can, for example, function fully independently from the state of a game, and thus carry little information, but have a large effect on how the game progresses. The opposite would also be conceivable, e.g. by means of an AI-driven assistant that analyses the complex state of the game in order to display it in simplified form to human players.

If an AI is given wide access to game state information as well as designed to actively modify such states, it can be compared more closely to the role of a 'game master' in pen-and-paper role-playing games. In this role, the system might be designed to find an optimum between challenge, relaxation, and diversity in order to provide a game experience that suits the idiosyncratic preferences of any participating player. Such balancing can be as simple as reducing the difficulty of a challenge by modifying hidden parameters, or as complex as changing the game narrative based on interpreted player preferences.

One of the upcoming topics in game AI is 'human/computer collaboration', which may be seen as one side of 'team AI'. In our context, this could entail all possible roles from allowing competitive gameplay, replacing missing human players with AI agents, to just providing more interesting interactions for human players, such that they do not feel lost. AI agents may have a 'digital life of their own' in an otherwise mostly

Figure 3.11: Photograph of *Anki Overdrive*, a physical miniature racing hybrid board game that can be played against AI opponents.

physical game, such that they neither have full access to the state of the game, nor do they have a large effect on the course of the game.

3.3.3 Discussion and Conclusion

In the previous sections, we have outlined different taxonomic lenses through which AI in hybrid board games can be discussed and explored. Given that hybrid board games are a relatively 'young' medium, there is a limited number of widely-known examples. Before concluding this paper, we look at game examples that can be described along the aforementioned dimensions with the aim of providing a better understanding of the individual dimensions.

One example that can be helpful in expanding the view on how hybrid board games can look like is the racing game *Anki Overdrive* [200] (see Figure 3.11). In the game, players take control of physical miniature cars and race against opponents. Cars can combat each other with virtual weapons that create a simulated physical impact and feature simulated differences in terms of car characteristics (e.g. speed and defence). Looking at the game through the lens of 'Embodiment as Agent', it can be described

as a game with a variable number of AI agents, including the possibility of letting AI agents race against each other by themselves. The relationship to the player is primarily competitive, with some game modes focusing on sabotaging other players, while others are more concerned with competing through flawless performance.

In terms of the 'Physical Domain', *Anki Overdrive* involves AI that has only limited awareness of the physical world. Cars in the game can only drive on specialised tracks, and obstacles that may be present cannot be detected, with the exception of other cars. Interactivity with the physical world is fairly high, as all racing manoeuvres are physical actions. While weapons cannot be seen directly, they can be perceived through the simulation of their impact on other cars.

In regard to the 'Temporal Domain', AI is primarily involved 'in-play', i.e., during the game session. Given that the game involves a companion application for mobile devices, the game could potentially involve AI for pre-play purposes. Here, the application could automatically generate patterns as suggestions for the player while including pre-computed parameters such as the expected difficulty. The temporal resolution in which the AI operates within the game is necessarily in real-time, given that any input by human players is carried out (almost) immediately. As such, any response needs to be processed and acted upon close to the reaction time of human players.

As long as players in the game lack another human player, the involvement of AI in the game is absolutely central to the gameplay. While players can race alone, the design of the game is built around competition, and thus, AI opponents in lieu of other human players.

Looking at the actor-director spectrum, i.e. the 'Role' AI plays within the game, we find that the game is closer to the midpoint than it might seem. While the individual AI-based opponents act as individual actors, their performance is actually in part dependent on how well human players perform. The developer at least claims that "the better you play, the better they become". Adjustments to the difficulty of a game, also known as 'rubber banding', fall closer to the 'director' side of the spectrum, as it suggests that weaker performance of a player also results in a less aggressive opponent. As such, the AI in the game is likely not only concerned with providing the best possible performance, but also considers what performance level results in the best player experience.

A complementary example case of a similar hybrid board game, both in its subject matter and its ability to inspire a broader view towards the medium, is *Room Racers* [201]. The game was developed as a research project and allows players to race

Figure 3.12: Photograph of *Room Racers*, a spatial augmented reality racing hybrid board game that involves AI as part of the racing track generation.

with cars that are 'projection-mapped' onto an arbitrary surface. Instead of involving physical cars, it involves the physical environment and asks players to create a racing track out of a variety of objects. While *Room Racers* exists at the fringes of the definitions of a hybrid board game, it involves intentionally designed physical artefacts, even if they are provided ad-hoc by players. In this example, AI is involved primarily through computer vision, as the outline of the track is processed in real-time from the physical environment. In contrast to *Pokémon GO*, the physical domain involves physical artefacts, even if they are designed by players instead of the game designer.

These two examples are intentionally chosen as use cases that test the boundaries of our working definitions. A game such as *XCOM: the Board Game* [202] is perhaps more easily identifiable as a hybrid board game with AI involvement, as it features a physical board and a companion application that includes some degree of scenario generation. While such games undoubtedly provide entertainment to their players, there is value in exploring other implementations that push the boundaries of what game AI can potentially contribute.

A potentially contentious edge case that we have not discussed up to this point could be found in games that are fully digital but use the digital domain to simulate physical board game elements. Such simulations can be as simple as using virtual cards and tokens (e.g. in *Tabletop Simulator* [203]) or involve a wider range of modal-

ities to invoke the feeling of a physical board game. At this point, we will leave the classification of such games up to future discussions that we hope this paper will encourage.

Finally, we have not addressed all possible attributes that may describe an AI in this first presentation of the taxonomy. For example, we did not include the power or skill of AI systems, despite the reality that much of the work in game AI focuses on balancing such attributes. An argument could be made that a skilled AI system likely needs to operate at a lower skill level in order to give human players a fair chance. Within the taxonomy that is presented, we could consider this a factor that is represented in part by the believability of an agent (i.e. 'to what extent does the AI play as a human would?') and the role of an AI on the actor-director spectrum (i.e. 'to what extent does the AI facilitate an enjoyable game session?'). However, this challenge of classifying the quality of an AI system in hybrid games emphasises that more conceptual and argumentative work is required to strengthen the currently presented foundation.

Overall, we have presented the conceptual foundation for developing a taxonomy of AI in hybrid video games. Part of this effort has been the establishment of working definitions that focus the exploration of the design space. We believe that future work, both applied and academic, can build on these efforts. This will ultimately allow for the development of novel game mechanics and support systems that contribute to the enrichment of the medium of hybrid board games.

3.3.4 Section Summary

In this last section, we tackled RQ6; by building upon the previous sections and developing a taxonomy for hybrid games. We indicate that digital elements in physical game contexts can assume different roles. In particular, AI agents represent the most intriguing element of hybrid games and can be points of potential development thanks to new AI technologies. In general, while older intelligent systems are already shaping the landscape of play beyond purely digital settings, more recent developments will open up to more natural human-AI interactions. Moreover, hybrid games can represent a rare instance of meeting points that happen in the real world rather than in the digital one; this makes them very promising tools for further research. In this chapter, we highlight the intimate relationship between games and AI. We start from historical and current perspectives and continue illustrating the potential for new interactive modes and contexts, all within the field of games. We aim to convey the final impression of

a medium that is both versatile and intuitive.