Universiteit Leiden
The Netherlands

Artificial Intelligence, Games, and Education

Barbero, G.

Citation
Barbero, G. (2025, September 16). *Artificial Intelligence, Games, and Education*. Retrieved from https://hdl.handle.net/1887/4260512

Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/4260512

Note: To cite this publication please use the final published version (if applicable).

Artificial Intelligence, Games, and Education

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof.dr.ir. H. Bijl,
volgens besluit van het College voor Promoties
te verdedigen op dinsdag 16 september 2025
klokke 11.30 uur

door Giulio Barbero

geboren te Acqui Terme, Italië

in 1992

Promotores:	Prof.dr. M.M. Bonsangue	
	Prof.dr.ir. F.F.J. Hermans	Vrije Universiteit Amsterdam
Co-promotor:	Dr. M. Preuss	
Promotiecomissie:	Dr. C. Anutariya	Asian Institute of Technology
	Dr.ir. E. Aivaloglu	Technische Universiteit Delft
	Prof.dr. S. Lammes	
	Prof.dr. M. Spruit	
	Prof.dr. S. Verberne	

Copyright © *2025 Giulio Barbero.*

Cover design *Giulio Barbero.*

Contents

1 Introduction **1**

1.1 Video games 2

1.2 Serious games 3

1.3 Games and Learning 4

1.4 Games and Artificial Intelligence 6

1.5 Artificial Intelligence and Learning 7

1.6 Research Questions 7

1.6.1 Sub Research Questions 8

1.7 Structure of the Thesis 9

1.8 Contributions of this Thesis 9

1.9 Other Work by the Author 10

2 The Use of Video Games in Scientific Education **11**

2.1 Introduction 11

2.2 How to Evaluate Games in Education: A Literature Review 12

2.2.1 Research Space Definition 14

2.2.2 Method 15

2.2.3 Results 17

2.2.4 The Dataset 17

2.2.5 Discussion 20

2.2.6 Conclusions 22

2.2.7 Section Summary 23

2.3 Computational Thinking Through Design Patterns in Video Games . . 24

2.3.1 Related Work 25

2.3.2 Concepts of Computational Thinking in Video Games 26

2.3.3 Discussion 32

2.3.4 Section Summary 32

3 The Relevance of Games for Artificial Intelligence 35

3.1 Challenges of open-world Games for AI: Insights from Human Gameplay 36

3.1.1 Problem Definition 37

3.1.2 Methodology 39

3.1.3 Results 40

3.1.4 Discussion 46

3.1.5 Future Research 50

3.1.6 Section Summary 50

3.2 The Effect of LLM-Based NPC Emotional States on Player Emotions: An Analysis of Interactive Game Play 51

3.2.1 Related work 52

3.2.2 Black stories 53

3.2.3 Video Game Development 54

3.2.4 Experimental analysis 56

3.2.5 Results 58

3.2.6 Discussion 60

3.2.7 Conclusion and Limitations 62

3.2.8 Section Summary 63

3.3 Towards a Taxonomy of AI in Hybrid Board Games 64

3.3.1 Working Definitions 65

3.3.2 Taxonomic Lenses 67

3.3.3 Discussion and Conclusion 73

3.3.4 Section Summary 76

4 Scientific Education, Video Games, and Artificial Intelligence 79

4.1 Generative AI and Programming Education: Considerations from Current Studies 80

4.1.1 Literature Review 80

4.1.2 Discussion 82

4.1.3 Future Work and Final Considerations 84

4.1.4 Section Summary 85

4.2 Video Games as Mediators of Generative Artificial Intelligence 86

4.2.1 Background 87

4.2.2 Research Question 88

4.2.3 Methodology 89

4.2.4 Results 90
4.2.5 Discussion and Conclusion 93
4.2.6 Section Summary 96

5 Discussion and Conclusions **97**
5.1 Discussion 97
5.1.1 RQ1; How effective are video games in the field of higher scientific education? 97
5.1.2 RQ2; How is research in the field currently carried on? 98
5.1.3 RQ3; What common affordances connect video games and computer science education? 99
5.1.4 RQ4; How do games present challenges for artificial intelligence development and study? 99
5.1.5 RQ5; How does artificial intelligence impact the development of hybrid games? 100
5.1.6 RQ6; How does AI impact programming education? 100
5.1.7 RQ7; How does the implementation of AI in video games perform in educational settings? 100
5.2 Conclusions 101

Bibliography **105**

Summary **125**

Samenvatting **127**

Acknowledgements **129**

Curriculum Vitae **131**

Chapter 1

Introduction

Play is one of the most important learning activities for human cognitive development [1, 2]. Throughout history, we encoded play in reusable and flexible systems, creating games as environments based on rules [3]. Since then, many games intrinsically assume a relevance beyond pure entertainment, punctuating human history as tools for escapism, community building, activism and others. Subsequently, with the development of digital technologies, the rise of video games has been observed, mixing characteristics of traditional games and interactive arts. Many aspects contribute to the popularisation of video games. One of these is the increasing processing power available throughout the second half of the 20th century. However, what fascinates many is the potential of video games to transport players to previously unattainable situations. In this regard, it is relevant to think that the first popular video games were mostly set in space [4, 5] or in ancient history settings [6]. Moreover, video games, as a product of play, inherited characteristics typical of role-playing, allowing players to impersonate characters with different potentials and problems from their own. Similar lines of reasoning can be carried on for experiencing different rules, abilities, and interactions. Therefore, video games stand out as a medium to create different realities and envelop players. This envelopment is typically conveyed by various aspects of engagement, which consists of occupying the willing player's attention. The willingness of attention is very relevant since it is one of the aspects that most of all allows video games to reach uses beyond entertainment; in layman's terms, games can make activities that are not inherently fun enjoyable. This aspect is studied and used in innumerable contexts, including education, citizenship, and research. Nowadays, video game research is an established, interdisciplinary, and popular field of academia. It is

able to confront itself with different subjects and bring necessary knowledge and skills to perform experiments that would have hardly achieved ecological validity otherwise [7]. Yet, the history of games has not been characterised exclusively by enthusiasm and acceptance. Throughout the eighties and nineties, we witnessed a raging debate about the impact of video games on youth, with the aforementioned opportunities eclipsed by concerns related to changes in entertainment habits and, in some cases, poorly substantiated research [8]. Such is the nature of playful interactions, as their pleasurable side is often associated with a lack of productivity and futility. In turn, video games as media were, and in part still are, met with scepticism and questions about their relevance. The present dissertation aims to show possible interactions between programming education and artificial intelligence (AI) through the lens of video games. The goal is to provide a perspective that highlights how these media can be tools to moderate and empower the interaction between the two. To do so, we first explore affordances in video games and programming education. Then, we dive into opportunities arising from video games and artificial intelligence. Finally, we analyse the impact of generative AI on programming education and introduce the challenges and opportunities for games to intervene in this context.

1.1 Video games

The first step, albeit banal, is to define the subject medium of this thesis. Defining video games in one sentence is quite complex, as the immediate answer intrinsically requires more clarification: they are digital and interactive games, often with a playful design [9]. Defining games, however, is far from straightforward. Most definitions focus on salient aspects of games, defining characteristics and boundaries. However, these boundaries have been repeatedly pushed and argued throughout the development of video games. For example, Salen and Zimmerman [10] define games as *systems in which players engage in an artificial conflict, defined by rules, that results in a quantifiable outcome*. Their definition fits a vast number of games, and it includes essential elements such as the necessity of rules. On the other hand, certain elements can be viewed as restrictive if confronted with modern games; consider city-management video games (e.g., *Cities Skyline* [11]), which often have blurred end conditions, making outcome evaluation arguable. A different approach was taken by Frasca [3]. As we cited above, he defined games as *simulations of rule-based systems*. This definition has a broader scope than the previous one and arguably covers all games. Moreover, it does mention the simulated relation between games and reality, which is a highly

fascinating aspect of the medium. However, some would consider the definition too broad, apt to cover also non-playful systems (e.g., climate models used for weather forecasts). In this debate, we take a pragmatic approach; while we are aware of the importance and limitations of these definitions, we believe that they are all relevant for the (video) games involved in our research.

1.2 Serious games

At this point, we can enter the field of games outside entertainment purposes. In this case, we are not simply talking about games designed with "serious" objectives, but also games as objects of study beyond entertainment. The practice of attaching additional values and goals to games began a long time ago. For example, board games like Go or Chess were designed to mimic military strategy. Throughout the 19th century, this evolved further with the game genre *Kriegsspiel* developed by the Prussian army to train officers [12]. Furthermore, games represent playful media for serious political changes. Examples include *Pank-a-Squith* [13], a fundraising tool for the British Women's Social and Political Union. The game depicts the conflict between suffragette leader Emmeline Pankhurst and British Prime Minister H. H. Asquith (hence the name). The game uses similar rules as *Chutes and Ladders*, associating tiles with the several challenges suffragettes needed to overcome to bring their petition to the Houses of Parliament (see Figure 1.1). We also have similar examples designed to raise awareness; *Womanopoly* [14], follows again a *Chutes and Ladders* structure with the rule variation that tiles have different effects depending on the gender of players. The game aims to communicate the challenges faced by women in modern society by expressively pushing men to play *the woman's part* (see Figure 1.2). In academia, games are well-recognised tools in psychological research. One of the most well-known types of games used in the field's experimental research is, in fact, the cooperation game. A type of cooperation game is the *game of trust*. In its most generic version, this game revolves around a first player deciding whether to cooperate or not; if the former is selected, the second player can decide whether to exploit the other player or share a reward [15]. This type of game has been studied extensively with numerous variations, and it is just one example of many experimental games used to study human behaviour. In psychology, games are appreciated for their inexpensiveness, complexity and, especially, for their ecological validity [16].

Figure 1.1: Pank-a-Squith, from the People's History Museum of Manchester

1.3 Games and Learning

As a subcategory of serious games, serious **video** games go beyond strictly digitalising existing opportunities and applications. With the development of the field of interaction design, game patterns have become common elements in the digital landscape. Moreover, thanks to the Internet, video games are appreciated in a serious context for their ability to easily reach people from different parts of the world. In general, games in digital environments evolve along two trajectories: **gamification** and **game-based learning** (or serious video games). The former represents the use of game elements outside of entertainment contexts [9]. The latter consists of fully fledged video games in which entertainment is not the main goal. Today, gamification is widely integrated into many aspects of daily life. The rise of digitalisation has made it easy to implement game elements in applications; it is enough to think about all the services that provide badges as rewards for specific behaviours. Other examples would be the ubiquity of leaderboards and various playful tools to stimulate competition [17]. Gamification has spread due to its well-documented effectiveness as a persuasive technique, particularly in enhancing user motivation [18], while remaining relatively inexpensive and simple to implement. However, this also opens to legitimate criticism of gamification; in fact, it is at times defined as *exploitationware* for its potential to tap into users' behavioural biases, persuading them to behave in a way they normally would not [19]. Within the

Figure 1.2: Womanopoly, by Stella Dadzie, from the Black Cultural Archive collection

context of this thesis, gamification is a direct example of the effect of games, or game elements more precisely, on motivation. Serious video games are arguably a more complex medium than most gamified environments; they require game design expertise as well as other topical knowledge. They are inherently interdisciplinary systems. They are fully fledged and developed games, designed with more or less classical basic requirements (engaging, interactive, etc.) [20]. However, their goals or motivations go beyond simple entertainment. Serious games are adaptable and are applied to several other fields. An excellent example is *EndeavorRx* [21], the first FDA-approved video game treatment for attention deficit hyperactivity disorder. The game is considered an effective medium for cognitive training therapy, even though it has initially been met with scepticism by some [22]. Moreover, there is a strong connection between serious games and the transmission of knowledge and training: this is the broad field defined as *game-based learning* [23]. This thesis has as its object of study this specific field,

intended not only as the domain of video games designed for learning but also learning conveyed by video games beyond their design purpose. In Chapter 2, we first present a literature review to describe the state of the art regarding the use of video games for higher scientific education experiments. Subsequently, we study the inherent connection between video game design patterns and computational thinking, arguing that video games can transmit important digital skills even when not expressly designed for it.

1.4 Games and Artificial Intelligence

Games and artificial intelligence mutually influence each other's development. On the one hand, games have accompanied the development of artificial intelligence, representing ideal environments for challenging, training and demonstrating new algorithms' effectiveness. On the other hand, artificial intelligence has taken a more and more important role in the development of new opportunities for play, most notably in video games. A well-known example is Procedural Content Generation (PCG). In this case, AI algorithms can generate new unique content based on the player's interactions with the video game [24]. This is such an established technology that has been implemented in numerous commercial video games [25, 26]. In academic contexts, we can define two main roles that games can take when it comes to intelligent systems: challenges and modeling. Games act as benchmarks for testing artificial intelligence's ability to solve complex problems traditionally handled by humans. In this regard, notable mentions are the game of *Go* and the corresponding algorithm AlphaGo [27] or *StarCraft II* and AlphaStar [28]. Other games are used or expressly created as models to train artificial intelligence. A notable example is OpenAI's work with reinforcement learning agents playing hide-and-seek, revealing emergent strategies [29]. In this thesis, we take an intermediate approach, studying specifically generative AI's potential as an interactive agent in video games. The interaction between the player and the AI can be interpreted as a challenge. At the same time, video game environments are also models, intended as models for interaction with humans. In Chapter 3, we present a study about open-world games as future challenges for intelligent systems and propose a framework to tackle them. As mentioned above, artificial intelligence can also have an impact on games. Video games present the most intuitive affordances, in particular when it comes to non-player characters (NPCs). Whether we refer to rivals or allies, artificial intelligence can empower gameplay by providing depth and proficiency to video games. However, it is often disregarded how digital technologies can also

be introduced to traditionally non-digital games. In this case, the roles of artificial intelligence vary. Again, in Chapter 3, we present a study about the potential effects of intelligent agents as NPCs. We will also present a taxonomy describing the use of digital technologies in what we define as *hybrid* games.

1.5 Artificial Intelligence and Learning

The latest developments in generative artificial intelligence have quickly and deeply impacted education. Recent studies show that the use of large language models (LLM) in computer science education hinders students' retention [30]. Moreover, education techniques are lagging behind the disruption caused by these new technologies [31]. Other perspectives look at the artificial intelligence proficiency of end users (i.e., what critical elements are necessary to use generative artificial intelligence effectively). In this regard, tying new technologies with computational thinking education is fundamental [32]. This thesis takes a critical approach to the use of generative artificial intelligence in programming education. Chapter 4 focuses on bringing together artificial intelligence and video games for educational purposes. This is a new field, seldom explored and speculative in nature; in the chapter, we use existing literature to analyse the state of the art and discuss the impact of generative AI specifically on programming education. We then embark on a simulated design process to develop prescriptive suggestions for future experiments making use of games in education. At the same time, we explore the role of video games as limiters for AI in educational contexts and highlight opportunities and challenges.

1.6 Research Questions

RQ; What is the role of video games in programming education in the era of artificial intelligence?

The question requires us to investigate three interconnected directions. First, we look at past and current use of video games for scientific and programming education. Then, we explore research in the field of generative AI used in educational contexts. Finally, we analyse video games as potential mediators between learners and education.

1.6.1 Sub Research Questions

RQ1; **(Chapter 2) How effective are video games in the field of higher scientific education?**
We study and compare existing research to clarify the impact of video games on education. In particular, we focus on the effects on students' performance and motivation.

RQ2; **(Chapter 2) How is research in the field currently carried on?**
We discuss the diversity of methodologies between studies in the field and its impact on comparability. We also discuss the lack but necessity of common practices to improve reliability.

RQ3; **(Chapter 2) What common affordances connect video games and computer science education?**
The goal is to identify common thinking patterns between digital gaming and programming. These commonalities compose the framework that supports the use of video games for programming education.

RQ4; **(Chapter 3) How do games present challenges for artificial intelligence development and study?**
We explore and demonstrate the intimate connection between artificial intelligence and video games. We estimate future challenges and build support for the use of video games as a common connection between artificial intelligence and programming education.

RQ5; **(Chapter 3) How can games be ideal meeting points for humans and artificial intelligence?**
We investigate how humans interact with intelligent agents in video games and how artificial intelligence technology can conversely impact players.

RQ6; **(Chapter 3) How does artificial intelligence impact the development of hybrid games?**
We analyse the impact that artificial intelligence can have beyond purely digital games. Conversely, we see how games can be a flexible medium to allow human-AI interactivity outside computer screens.

RQ7; **(Chapter 4) How does AI impact programming education?**
By developing a position on the effect of generative AI on education, we can identify weaknesses and opportunities. The goal is to see how video games can

fit into these systems, overcoming some of the weak points and exploiting the opportunities.

RQ8; **(Chapter 4) How does the implementation of AI in video games perform in educational settings?**
The final goal is to coordinate all the answers to the previous questions into one general reasoning about the role of video games in the future of AI and education.

1.7 Structure of the Thesis

Besides the present introduction, this thesis is structured around four more chapters. The first three present articles connected by different topics. Chapter 2 revolves around the role of video games in scientific higher education and programming skills development [33, 34]. In Chapter 3, we focus on interactions between AI and video games [35, 36, 37]. Chapter 4 analyses the intersection of generative AI, (programming) education, and video games. Finally, in Chapter 5, we discuss the research questions mentioned above and conclude by talking about the potential affordances of video games as moderating media to use generative AI in programming education.

1.8 Contributions of this Thesis

[34] G. Barbero, M. A. Gómez-Maureira, and F. F. J. Hermans, "Computational thinking through design patterns in video games," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, 9 2020.

[36] A. Marincioni, M. Miltiadous, K. Zacharia, R. Heemskerk, G. Doukeris, M. Preuss, and G. Barbero, "The effect of llm-based npc emotional states on player emotions: An analysis of interactive game play," in *2024 IEEE Conference on Games (CoG)*, pp. 1–6, 2024.

[33] G. Barbero, M. M. Bonsangue, and F. F. J. Hermans, "How to evaluate games in education: A literature review," in *Smart Learning for A Sustainable Society* (C. Anutariya, D. Liu, Kinshuk, A. Tlili, J. Yang, and M. Chang, eds.), (Singapore), pp. 32–41, Springer Nature Singapore, 2023.

[37] M. A. Gómez-Maureira, G. Barbero, M. Freese, and M. Preuss, "Towards a taxonomy of ai in hybrid board games," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, 9 2020.

[35] G. Barbero, M. Müller-Brockhausen, and M. Preuss, *Challenges of Open World Games for AI: Insights from Human Gameplay*, p. 127–141. Springer Nature Singapore, Nov. 2024.

1.9 Other Work by the Author

[38] G. Barbero, R. Albrecht, C. Daske, and M. van Noordenne, *Emotion Recognition: Benefits and Human Rights in VR Environments*, p. 17–32. Springer Nature Switzerland, Oct. 2024.

[39] M. A. Gómez-Maureira, I. Kniestedt, G. Barbero, H. Yu, and M. Preuss, "An explorer's journal for machines: Exploring the case of cyberpunk 2077," *Journal of Gaming & Virtual Worlds*, vol. 14, p. 111–135, Apr. 2022.

Chapter 2

The Use of Video Games in Scientific Education

2.1 Introduction

This chapter explores aspects of the relationship between video games and scientific education. It is composed of two sections [33, 34]; in the first, we introduce the topic through a literature review. The review is focused on the experimental uses of video games in higher scientific education. Its goal is to highlight the variety and development of the field while also shedding light on criticalities that arise from the diversity of applications. The second article studies the relationship between video game design patterns and computational thinking skills. It examines inherent affordances between the two with practical examples. We position our work within the framework of computational thinking as defined by Wing [40] and further refined by Brennan et al. [41]. Specifically, we investigate whether game design patterns facilitate computational thinking practices, extending previous research in game-based learning with focus on engagement and motivation (e.g., [42] or [43]) as well as in constructivist learning theory [44], which suggests that learners construct knowledge through active engagement with their environment. In fact, games provide a structured yet flexible space where players experiment, make mistakes, and iterate on their learning strategies, a process closely related to the theory of constructionism [45]. We argue that video games create an environment that facilitates the activation of mental patterns as computational thinking through active experimentation. We conclude each section

with a short summary of its relation to the overall chapter and the research questions.

2.2 How to Evaluate Games in Education: A Literature Review

The digitalisation of education is a common trend in many different contexts. Just to cite a few examples, the number of courses available online is countless, most books also make a digital version available, and a massive amount of lectures are streamed and available online every day. Digitalisation opens the way to new methods to convey information and its applications in education, creating spaces for new teaching tools and techniques. A very promising one is extracting elements from video games and their application to enhance education (*gamification*). Moreover, video games can also be used in their entirety to convey scholastic knowledge (*game-based learning*).

The current section collects and analyses studies involving controlled experiments on gamified and game-based learning. The focus of this review is experimental studies of gamified or game-based learning techniques applied to scientific secondary or higher education. When gamification is applied in this context, experimentation often translates as the application of playful elements in education. On the other hand, game-based learning usually involves the design of video games, fully focusing on conveying educational content. The goals of these experimental tools can vary from enacting behavioural changes [46, 47] to improving knowledge acquisition. In their literature review, Hainey et al. [48] analyse more than 100 studies in the field and report knowledge acquisition as the predominant learning outcome of 64 studies out of 105.

Regardless of the field of application, the cognitive effects of games or game elements vary as well. The most commonly reported effect is an increase in students' motivation and engagement, in line with the historical trend to utilise games in education as a medium to make experiences more pleasurable [9, 49]. On the other hand, the link between game elements and actual education effectiveness is not as clear. Even though many studies report a positive relation between the two [50] and are backed by empirical game research [51], many experiments in the field of games present opposite results [52] with the respective theoretical research to support them [53].

Such diversity of studies (and results) in the field justifies the relatively high number of meta-studies. These tend to focus on specific aspects of the gamified/game-based

learning experiments, but also include more general information about the contexts of application (where possible). For example, the above-mentioned review by Hainey [48] focuses on documenting the diversity in the field. This is reported in terms of learning outcomes, topic of application, and quality of the study. The review selects empirical studies from 2000 to 2013 that apply game-based learning in primary education curricular subjects. Laine and Lindberg [54] focus on game motivators and how different studies in education used them and reported different effects. While the selection criteria are less strict compared to Hainey et al., the study also includes more recent papers (from 2000 to 2019). Finally, Hamari et al. [55] perform a broader review, focusing on how the studies were carried on. Through this perspective, they draw several conclusions about the quality of the experiments. In particular, they define multiple issues arising from the lack of clarity in the reporting style of many studies. They also report a sharp increase in studies involving gamification, which more than quadrupled in one year (2011-2012).

Problem Definition and Research Questions

The lack of clarity in the reporting style of many experiments makes comparative approaches challenging. In particular, the impact of different contexts of application related to the implemented game components is considered. The relevance of the context can become especially evident in controlled experiments. In these, the effect of experimental conditions (i.e. the game elements applied to education) can usually be better analysed with a clear understanding of the control conditions (i.e. the standard education method for that specific context). However, this is often challenging or impossible due to a lack of clarity in the description of the control conditions [55]. In practice, the lack of this type of contextual information makes it difficult to answer important questions to evaluate the experimental approach; how is the subject taught in the control group? What type of material is used? How are the experimental and control groups evaluated?

The present review arises from the need to study this lack of clarity and to define what elements of control groups (or, more generally, the no-game approach) are most frequently disregarded. Therefore, our first research question *Q1* is "(In controlled studies involving the use of game elements in education published between 2013 and 2020,) how is the information about the control group reported?" with a specific focus on the quality (and clarity) of this information regarding teaching method, teaching material/media, and experimental evaluation method.

Another element that can potentially influence the clarity of reported information

is the inherent differences between the fields in which gamification and game-based learning are applied [48]. This leads us to our second research question (Q2): "How does the subject of application influence information clarity?".

The present review aims to build on existing knowledge and aid experimental research by improving replicability, enhancing comparability between studies, and highlighting the use of games and game elements in different fields.

2.2.1 Research Space Definition

In this section, we explore relevant characteristics of the field further as we motivate the initial selection criteria and analysis tools. As mentioned above, our research question is derived and influenced by two main characteristics of the field: diversity in the type of studies and diversity in the context of its application.

Characteristic 1: Type of studies

Many empirical studies that utilise games involve the use of an experimental and a control group. Although information about control (or no-game) conditions can be helpful to evaluate findings, many studies in the field omit it to different degrees. This is a known issue that can hinder analytical approaches focused on the influence of experimental conditions [56]. Other studies do not use control groups and only rely on qualitative analyses of information gathered over the entire population. Categorising these studies is even more complex, and comparison with different studies is challenging. Based on this initial distinction, we create the following initial criterion to select papers relevant to our research question:

- Presence of a control group: defining starting conditions as the standard academic path, we add this criterion in order to ensure the relevance of pre-intervention context descriptions (also when provided by the same course results in previous years).

We also determine common elements that are necessary to replicate a controlled experiment involving the use of gamification and game-based in education:

- Elements of starting conditions: we categorise each paper by reporting how much information we can find about starting conditions, usually represented by the control group. In this regard, we define three elements as relevant: type of teaching material (the tools used to transmit course content), teaching method

(how the course is taught), and evaluation method (how the effectiveness of experimental and control methods is evaluated).

Characteristic 2: Context of application

Games have been studied and used in education throughout different academic curricula, from scientific to humanistic subjects, in academic and technical education. In the study of languages, for example, game elements have been used and appreciated in both academic and "more commercial" settings [57, 58]. Also, the field of mathematics experimented with adding game components to different grades of education [59]. Moreover, games are used and studied for both practical (training) and theoretical knowledge acquisition. Finally, another big part of the studies involves the use of game elements in behavioural change projects aimed at educational environments, for example, to promote safer sexual practices [60]. Such diversity, which naturally arises from the shared interest of many disciplines in the use of educational games, makes comparisons between studies very challenging. Therefore, we narrow our search by using strict criteria to select studies to include in this review:

- Scientific subjects: we focus on studies about how games influence the absorption of scientific notions. Since we focus on how different fields produce different studies, we include both natural and social sciences in order to preserve some variation.

- Knowledge acquisition: we include studies that aim at the acquisition of theoretical or practical knowledge. The rationale behind this is to include studies in the field of medicine and nursing, which often mix the two. On the other hand, we exclude studies whose goal is to develop behavioural change.

- Secondary or higher education: the studies we select involve students currently enrolled in secondary education. This includes high school and university-level courses. It excludes doctoral and specialisation studies in which participants are often professionals.

2.2.2 Method

Inclusion criteria

We summarise and motivate further criteria used to filter the studies, including those identified in the previous chapters:

- Date: 2013-2020. We have chosen to focus on where Hainey et al. [48] left off. As mentioned in Hamari et al. [55], the number of game-based experiments in education is increasing sharply every year. Focusing on recent years allows us to also investigate recent developments in the field.
- Database: Leiden University Library [61].
- Type: video games and hybrid games. We added this criterion in light of the fact that the majority of the studies collected involve digital components. In this way, we want to eliminate the few outliers which could prove difficult to compare.
- Includes: game elements and games. The experiments involve the use of game elements or full games. This excludes pure simulations in which game systems, or more generally, "pleasurable" components, are not implemented.
- Search terms: ("serious game" OR "game-based" OR "gamification" OR "game elements") AND ("experiment" OR "evaluation" OR "impacts" OR "outcomes" OR "effects" OR "education" OR "learning")
- Language: English

Selection

After using the filters available on the database website ("Date" and "Language") to limit the date and the language of the studies, we proceed to read the abstract of each of the first 100 papers (sorted by Relevance and using a combination of the "Search Terms"). We then determine whether it respects the rest of the aforementioned selection criteria (double-checking the year of publication). In case it does not respect one or more of them, the study is excluded from the present review. At the end of this selection process, we collect 89 studies. We then proceed to read the full paper and make a final selection. The final number of included studies is 43.

Categorisation

With the goal of documenting diversity and clarity of starting conditions in mind, we summarise the categories through which the studies are classified:

- Field of application: medical sciences (medicine and nursing), natural sciences (biology, mathematics, physics, computer science, etc.), economics (economics,

business, management), and social sciences (sociology, psychology, anthropology).

- Type of education: theoretical, practical (training), or a mixture of both
- Clarity in the context of application: type of teaching material, teaching method, or evaluation method. For each:
  - Unclear - the experiment would not be replicable with the information presented
  - Clear - the experiment would be replicable with the information presented
- Grade of education: high school level or university level
- Results: The effect of the experimental condition on motivation and performance:
  - Negative (overall worse results in game condition)
  - Mixed/No-change (negative and positive results or no change in game condition)
  - Positive (overall better results in game condition)

2.2.3 Results

In this section, we report the quantitative results derived from the analysis of the included studies through the aforementioned categories.

2.2.4 The Dataset

We collected 43 controlled studies[1]. The studies are, in general, quite recent; almost half (N=20) were equally published in 2017 and 2019. Also, a good number (N=9) were published in 2020. The average year is 2018.

Classification by Field of Application

Medicine is the topic with the highest number of studies (N=8), followed by math (N=6) and computer science (N=6). All the other topics score equal to or smaller than 4. The number of studies which investigate the effect of game-based education,

[1] [62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104]

also (at least in part) on training, is 10, of which 8 are contextualised in courses involving a life science (medicine, nursing, or physiotherapy)(see Figure 2.1).

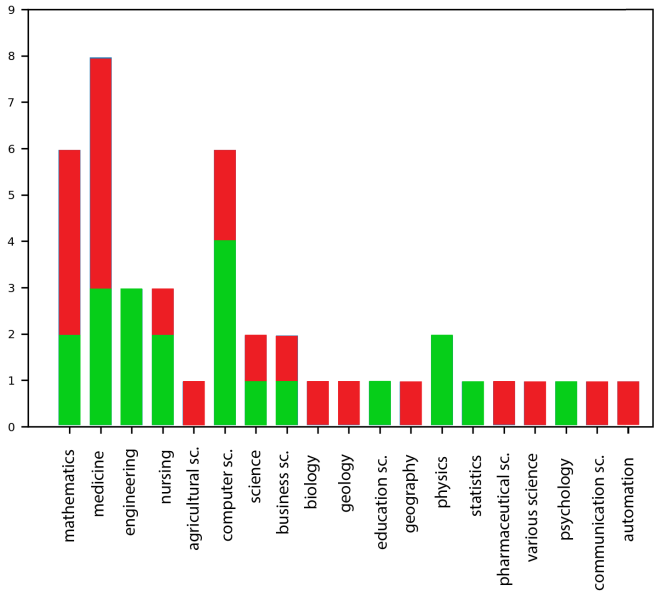

Figure 2.1: Number of studies per subject divided by success rate ("positive" in green, "mixed" and "negative" in red)

We cluster the fields reported in the studies into four categories: hard sciences, life sciences, social sciences, and engineering. For each category, we then calculate the success rate. The two categories with the highest success rate are hard sciences (N=10/18) and engineering (N=3/4), with computer science being the subject with the highest success rate (N=4/6). It is important to note that there is some possible overlap between these two categories (for example, subjects studied in computer science could also be studied in computer engineering). Life sciences present the lowest success rate (N=5/14). Almost the majority of the studies reported "positive" results (N=21/43) while 12 studies reported "mixed" results.

Classification by Clarity Scores

We analyse the frequency of the scores for the details of the educational context. For 22 studies, the details reporting the type of teaching material used in the control groups are deemed unclear (see Figure 2.2). In the case of the teaching method, the results show that in 30 studies, this information is evaluated as unclear(see Figure 2.2). On

the other hand, testing methods are often more clearly reported, with only 12 studies marked as unclear (see Figure 2.2). In general, only a few studies lack detail in all three categories (N=6). Some are unclear in one or two categories (for both cases N=15) while a few others are clear in all three (N=7).

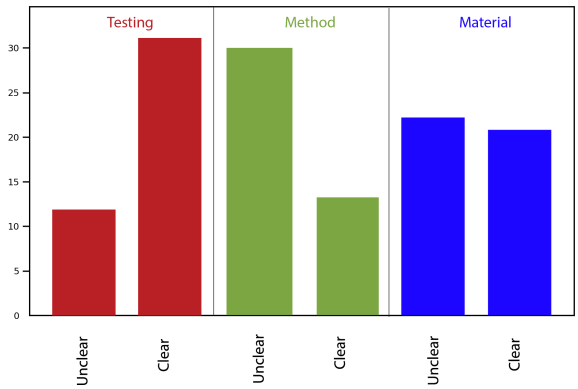

Figure 2.2: Number of studies per clarity score related to, from left to right: testing method, teaching method, and teaching material

Classification by Success Score

Almost half of the studies (N=21) report positive results across both performance and student attitudes (when this was relevant, see "5.1 Limitations"). 12 studies report mixed results, noticing only partial improvement in either performance or student attitudes. The rest (N=10) report worse results in the conditions involving games(see Figure 2.3).

Pearson correlation

We calculate the correlation coefficient between the success rate (mapping the scale 'negative' - 'mixed' - 'positive' over a scale from 0 to 2) and the average of the context clarity values (considering 'clear' labels as 1 and 'unclear' as 0) for each reviewed article. The result shows no meaningful correlation between the two (r=0.073). The correlation coefficient for the individual clarity of educational context values and the success scores also does not report a meaningful correlation (teaching material r=-0.078, teaching method r=0.229, testing method r=0.004). We additionally calculate the correlation coefficient between the values for the clarity of the educational context. Also in this case, the results are all well within the interval of -0.6 and 0.6; therefore, no

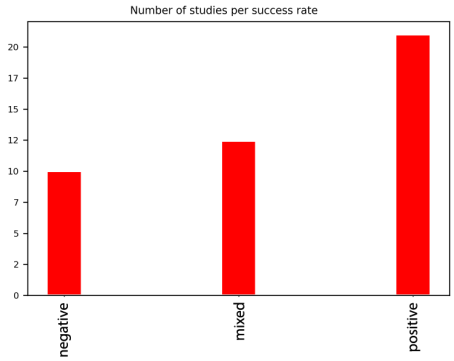

Figure 2.3: Number of studies per success score

meaningful correlation can be reported (material-teaching method r=0.369, material-testing r=0.089, testing-teaching method r=-0.042).

Clustering the Lack of Detailing

Within the studies we found lacking in terms of detailing scores, we delineate different groups based on how the lacking aspect is treated. This creates two categories valid for both the scores related to the clarity in reporting teaching material and method. One category includes studies that simply did not mention these two components [62, 64]. The other category does mention these elements of the control conditions but resolves them in generic terms (e.g., defining them as "standard", "classical" or "traditional") [67, 105].

The same categories cannot be applied to the scores of the evaluation method detailing. The main reason is that these studies are more consistent. Even when being "unclear", the evaluation is mentioned in the paper as an essential part of the experimental design. However, these cases refer to their results (e.g., the presence of an "improvement") without explaining in detail the type of tests and analyses that led to those conclusions.

2.2.5 Discussion

Q1: "How is the information about the control group reported?"

The present review focuses on how information in controlled studies involving games in education is presented. Our analysis shows that the most frequently unclear element

of the educational context is the teaching method. In this regard, many studies fail to mention the way the teaching material is presented. It is important to notice that, when it comes to this teaching material, the actual information provided is more complete than the information regarding the teaching method; many studies report at least the type of material used (specific books, PowerPoint slides, and online courses are the most common).

Following also the qualitative parts of our analysis, we cluster those studies which are unclear according to these first two scoring systems and postulate reasons for the lack of detailing. Some of these studies rely on the idea that the respective educational practices (teaching method and material in this case) are supposedly standardised in certain contexts. However, this type of standardisation often does not account for teaching methods and/or the variation of these courses throughout the years in terms of content and material used. Other studies, in particular regarding the educational method, might simply overlook the importance of these elements and do not touch the subject.

The description of the testing method is usually clearer and in line with expectations. Few studies score 0, usually focusing on additional qualitative evaluations after the actual performance tests. Overall, it is rarer to encounter studies reporting no information at all in this regard, probably due to the intrinsic and fundamental nature of evaluation methods in experimental designs.

Q2: "How does the subject of application influence the results?"

Looking at the success rate, results show that many studies (N = 21) reported benefits from using games for parts of their educational components. However, looking at the success score for each topic, games or game elements do not seem to be equally effective in every subject; for example, only three studies in the field of medicine report "positive" results (N = 3/8). This could indicate that game elements are more effective in some fields compared to others. Topics that often heavily involve the study of technology (engineering and computer science) score above average (respectively, N = 3/3 for engineering and N = 4/6 for computer science). This could indicate that students who are usually more involved, or at least interested, in technology are more susceptible to the effects of games or more motivated by such tools. However, other subjects that make heavy use of computational tools, such as mathematics, show a low success rate (N=2/6).

Finally, it is important to observe that the lack of correlation between the scores (in particular detailing scores and success scores) can in part attest to the research's

quality since the final results of the experiment do not seem to influence reporting decisions.

Limitations

Our inclusion criteria are effective in defining the area of interest of our research. However, they also present some intrinsic limitations:

- Limited field: we focus our research on subjects related to natural, social, and technical sciences. On one hand, this is functional to collect studies that follow a common scientific experimental method. On the other hand, it excludes a relevant amount of studies, in particular in the fields of language acquisition and history education.
- Limited types of courses: our current collection contains studies that involve both practical and theoretical knowledge acquisition. However, our previously mentioned field selection (sciences) automatically excludes studies in purely professional training settings. This often (but not always) affects experimental applications to training courses in vocational schools.
- Subjective clarity scoring system: the scoring system we use is strongly related to our personal interpretation of the various texts. This refers to the natural distance between the intention of the author and the interpretation of the reader.
- Relative success scoring system: the success scoring system (negative-mixed-positive results) is highly dependent on the scope of the individual study. Some focus and report exclusively on performance improvement, while others might include students' engagement and motivation. However, it is a valid parameter to determine how the results are analysed in each study and what perspective each paper takes on game-based education effectiveness.

2.2.6 Conclusions

The study successfully highlights the lack of clarity in describing the educational context as a common issue in the field of game-based education studies. This has an impact on the difficulty of interpreting exactly what educational elements the gamified/game-based experience goes to replace. Subsequently, this can strongly hinder comparative analyses. Moreover, it is very challenging to perform any in-depth categorisation according to the educational context within studies with incomplete

information. Future studies indicated below need to either adapt to this issue to minimise its impact. The study also hints at a relationship between the success of gamified/game-based education and students' intrinsic motivation. Although this link is usually accepted, it is often disregarded in terms of its intrinsic value for learning. We argue that, in part, this is related to the tendency to belittle the importance of play as an activity that fosters fun, limited to the definition of 'autotelic' as an activity that has a purpose in itself. If we accept the effect of video games on motivation in educational contexts, we then understand how the purpose of play as a pleasurable activity goes, in certain cases, beyond play itself.

Future Work

Basing ourselves also on the aforementioned limitations, there are several directions to extend this work:

- One can use the same method and apply it to studies in the field of humanities. This is a better option than incorporating these studies directly into the current selection, since the two groups present very different teaching goals and methods. Moreover, the way study results are reported may also vary.

- It is also relevant to apply the same method to studies in vocational education. Notice that the three clarity scores (material, teaching, and testing) might need to be adjusted to include categories more closely linked to practical education.

- As shown in Figure 2.2, the studies are very diverse, also when it comes to the completeness of information reported. Comparative studies based on this collection should take this diversity into account, developing a methodology that can either adapt to it or compare studies from a different perspective, which minimises the impact of starting conditions.

2.2.7 Section Summary

The study above relates to two of the research questions of our thesis, in particular RQ1; and RQ2;. First, it highlights the effectiveness of games in education when it comes to motivation. It also indicates a positive impact on performance, though further research is needed. Secondly, the section investigates how experimental research in the field is carried out. It reports critical points in terms of clarity and, therefore, comparability. In general, we identify the need for tools for further analysis of the

effect of video games in education. In the next section, we are introducing a potential solution to this problem. We highlight a correspondence between computational thinking skills and game design patterns. In doing so, we show how this tool can be effective in framing individual game elements and analysing them in terms of how they influence play.

2.3 Computational Thinking Through Design Patterns in Video Games

Learning how to program involves more than absorbing the syntax and semantics of a specific programming language; it also requires sensibility in combining and implementing these terms in an efficient and functional way. Programming makes use of procedural thinking, planning, data analysis, data re-elaboration and established practices such as testing and debugging [106]. All these components and skills find definition under the concept of "computational thinking". Training computational thinking skills and being able to use them proficiently is a common objective for programming education, and it is often one of the most challenging components for learners.

An important part of the research in the field is focused on finding new media and techniques to facilitate the development of these skills. Promising research has been conducted using computer games to train computational thinking components [107].

Video games present advantageous characteristics for this scope: they can support problem-based learning, require information retrieval to succeed, provide immediate feedback allowing testing, can easily embed assessments and often create a social environment or community [108]. They also motivate users with challenges and entertaining components [109]. Prior research at the intersection of video games and computational skills has often been carried out in two main directions: one tends to embed and test these components in environments that were created specifically for that purpose, such as is generally the case in educational games [110]. The other seeks to analyse the effect of general gaming experiences (i.e., not purpose-built for personal improvement) on a set of computational thinking skills [106].

While the first approach tends to deliver results for the study of the actual medium, the second takes a too general point of view that often faces noisy results due to the extreme diversity of elements in the game world. In this paper, we argue that focus-

ing on generalizable game features and design patterns that benefit the development of computational thinking skills offers a valuable middle ground between these two approaches. We present this approach by outlining examples of design patterns that are most promising in the context of supporting programming education. In the following sections, we present and describe the set of computational thinking skills we decided to use. We will then list and describe design patterns that we think can be positively connected to each of them, also providing practical examples of where they are applied. The diversity and specificity of the examples suggest that each skill is activated by different video game components. Recognising these, we can explore a new potential way to study the relation between gaming and computational thinking.

2.3.1 Related Work

Computational Thinking and Programming Education

The definition of computational thinking skills varies depending on the author, with different sets of overlapping components; often conceptually related to methods for data extraction and re-elaboration or logical and procedural reasoning. A commonly cited model comes from Kazimoglu et al. [111] and builds on the work of Wing 2006 [40], Wing 2008 [112], Ater-Kranov et al. [113] and Berland & Lee [114]. It lists five fundamental computational thinking skills. These are (1) conditional logic, (2) building algorithms, (3) debugging, (4) simulation and (5) distributed computation.

Conditional logic involves an understanding of true and false values and their use in control flow statements. This involves being able to evaluate the status of a system in a specific local statement and an understanding of how each operation manipulates it. **Building algorithms** is a form of step-by-step problem-solving that requires a more solution-driven view of the multiple conditional logic instances. It shares some overlap with the previous skill, but in this case, it is necessary to have an overview of all the single manipulations to understand how the system reaches a desired final status (i.e., the solution). **Debugging** describes the process of testing in order to spot and find solutions to problems in the code. **Simulation** refers to the creation of mental or physical models to define how to implement algorithms and which circumstances apply. Finally, **distributed computation** groups all the social aspects of programming, from project-oriented working to making use of and contributing to a community [111].

The main advantage of this set of computational thinking skills is its practicality; each skill is well-defined, easy to understand and covers a good part of the components

that are necessary in the process of programming. It goes to depict a picture of computational thinking as a reasoning process that goes from the detail (conditional logic) to a larger view of the relations between them (building algorithms). It further includes practical elements that are necessary throughout the whole programming process, such as debugging and simulating interactions between the components to reach the desired final state. Finally, programming is often an activity that heavily relies on the community behind it, and distributed computation can be used to describe all the skills necessary to access, use and contribute to this community.

Design Patterns in Video Games

In order to identify useful video game components, we follow the definition of "game design patterns" as described by Björk and Holopainen [115]. Generally, design patterns are reusable structures for finding solutions to common problems in a domain (such as architecture [116] or computer science [117]). Depending on the field, this can range from the application of narrowly defined instructions to more general recommendations for specific circumstances. In the field of game research, Björk and Holopainen define game design patterns as *"semi-formal interdependent descriptions of commonly reoccurring parts of the design of a game that concerns gameplay"*.

Game design patterns fit the purpose of our approach for multiple reasons. First, their definition is derived from a common term in the field of computer science. When studying computational thinking, this is a beneficial connection, especially in terms of communication and the structure of the knowledge for both the fields of computer science and game research. Second, they help to deconstruct video games into elements that can be studied and used more flexibly than focusing on the entirety of a game.

2.3.2 Concepts of Computational Thinking in Video Games

While previous research has examined computational thinking in educational games, little work has explored how video games (inadvertently or not) teach these skills. In this section, we propose a new perspective that extends existing models by incorporating game design patterns as a means for the activation and, perhaps, training of computational problem-solving.

Conditional logic: Some could argue that most decisions, especially in video games, are binary, therefore based on conditional logic. Even though this could be the case, there are some particularly connected components in video games that are worth mentioning. In [115] we find the pattern of **"Incompatible goals"** which refers to

those situations in which pursuing a certain objective automatically forbids trying to pursue others. Players need to be able to evaluate the conditional status of those game elements that are triggering this pattern in order to understand the logical development of the video game. Another pattern that requires the application of conditional logic is **"Varied gameplay"**. This pattern describes how certain choices and settings can provide the players with completely different game instances. Especially role-playing games (RPGs) serve as a fitting example, given that every decision players make opens up some paths while closing down others. A popular example can be found in the 'The Elder Scrolls' [118] series where different choices in the character creation and in the story itself lead to very different gameplay options and overall narrative experiences (see Figure 2.4). This is facilitated by a sequence of conditional choices that allow and disallow certain features within the game as players progress.

Figure 2.4: Example of an interaction with mutually exclusive choices with an NPC in Skyrim

Building algorithms: Building algorithms entails following a step-by-step plan to solve a problem. It requires the ability to individually evaluate those steps (using conditional logic) and to analyse the results of their sequential combination. Many of the game patterns that stimulate this skill make use of different aspects of this skill, requiring players to plan, concatenate and modulate the manipulations necessary to reach the desired status. A fitting game design pattern is the **"producer-consumer"**

pattern, which guides the use and importance of resources. In some games, it even determines the speed of the gameplay [119, 120]. In complex systems, this pattern generates a network of interrelated producers and consumers. Often, to reach a specific objective, there are multiple steps of resource gathering, production and manipulation (which usually includes consumption) to be developed. Being able to foresee and plan over multiple cycles of discovering, extracting, transporting, storing and consuming resources requires similar mental mechanisms as building a computational algorithm. We can compare the producer-consumers to different functions returning elements as outputs and requiring outputs from other functions as input. The sequence of these elements and their inputs and outputs must be planned carefully in order to reach a certain goal. A very important concept of building an algorithm is taking a step-by-step approach [111], and similarly, we can see a step-by-step approach when building a producer-consumer network in many '4X' games (a sub-genre of strategy games that involves eXploration, eXpansion, eXploitation, and eXtermination). This game design pattern is noted to conflict with the pattern "predictable consequences" which makes sense from a computational thinking point of view as well: complex algorithms with multiple steps and data manipulations are often more difficult to manage and usually require careful debugging.

Practical examples for this pattern are numerous, and it is arguably an essential component to most strategy games. For instance, in 'Stellaris' [119], developed by Paradox Development Studio, certain jobs (tasks in the game) produce 'minerals' that are then consumed to produce 'consumer goods'. These can then be consumed again by other jobs to produce 'research'. Since the ratio of consumption to production is hardly 1:1, this system requires players to carefully plan the construction of jobs, usually in order not to end up with a negative balance in any of the above-mentioned resources. Another typical example is the complex and ramified network of resources of 'Thea 2: The Shattering' [121], developed by MuHa Games and Eerie Forest Studio. In this case, we have four different tiers of resources, with the last one being 'crafted', consuming resources of the previous tier and requiring the acquisition of specific technologies. In this case, we can see a sequence of steps necessary in order to acquire technologies, gather resources and craft them into a higher level one. Similarly, in 'Civilization V' [120], the player funnels resources into 'science' production, which in turn unlocks the exploitation of new resources.

Debugging: In order to analyse debugging, we need to unpack the several elements that compose this skill. Debugging is understood as a process of trial and error that is developed through testing. A corresponding game design pattern is "experi-

Figure 2.5: Screenshot of 'Doodle God' showing combinations of basic elements to create complex elements.

mentation". It usually indicates that a part of the game mechanics requires a process of trial and error to be evident, understood or mastered. In the most extreme cases, the whole gameplay revolves around experimentation in the form of 'puzzles' to be solved. Experimentation is often realised through trial and error as well, with testing being a necessary component of it. Similarly, debugging usually involves being able to critically think about the current configuration and can necessitate multiple trials to determine where the errors are, as well as how to fix them efficiently. It is important to point out that experimentation is a quite broad pattern and its usefulness to debugging skills generally holds only under specific applications. If we want to better specify the context, we need to limit it to the intersection with the game design pattern referred to as "Puzzle-solving". This refers to game features that need to be solved through inductive or deductive reasoning. If applied together with experimentation, we are arguably defining an even closer reasoning to debugging; a mental process that, through deductive or inductive trials and errors, attempts to spot and solve problems in the current solution. An interesting game example can be found in the mobile game 'Doodle God' [122] by JoyBits (see Figure 2.5). In the game, the player needs to combine basic elements (such as fire or water) to create new, more complex ones (for example, life or energy) that can be used to create even more complex elements. The

whole game is based on a process of reasoning and, especially, trial and error. Players can think about potential element combinations and try them to see if they achieve new elements. The game also features helpful support for players that can partially (but rarely completely) provide hints to the creation of new components, highlighting one of the two elements that need to be combined.

Simulation: Similarly to 'conditional logic', simulation is a very broad category that refers to essential concepts of many video games built as representations of real or imaginary phenomena. It seems self-evident that video games include simulations of some sort. However, it can be beneficial to focus on patterns that allow or elicit simulation skills within games themselves rather than understanding games as a simulation of real-life processes. Here we encounter some overlap with "debugging" since the game design pattern "experimentation" can once again be useful in this context. Players can create a set of possible actions and evaluate them using simulation skills. Subsequently, these actions are validated by experimenting with them. In general, experimentation requires activating mental simulation mechanisms in order to narrow the set of possibilities to try. Other game patterns do not directly trigger simulations but might favour them. A typical example is the 'Save-load cycle' pattern, where players can save and reload games at specific points (or, in more flexible cases, from the main menu), allowing them to revert to a previous state and replay challenges or actions. This can elicit simulation skills similarly to experimentation; players can simulate and select certain solutions and then try them in multiple rounds, loading back the game at every iteration. In the 'Final Fantasy' game series (for instance 'Final Fantasy X' [123], see Figure 2.6), players can usually find saving spots right before the most challenging battles. In this way, the player can try certain settings and, in case of failure, analyse their own errors, improving on them after reloading the game from that last saving location.

Distributed Computation: We can identify several game design patterns that show useful traits in aspects of distributed computation. One example is "communication channels" which are present in many games that allow players to communicate with each other. "Cooperation" is another design pattern that can be connected to the idea of working together for a goal. However, we argue that it is even more compelling to notice how programming and gaming often behave similarly in their relationship with the respective communities [124]. We would argue that distributed computation is a skill that is necessary not only in computational thinking but also in many multiplayer games, or perhaps even certain single-player games with multi-agent aspects. Indeed, both involve massive online communities interacting, debating and

Figure 2.6: Example of a saving point in Final Fantasy X

sharing information to complete tasks that would be hard or impossible to achieve alone. Moreover, even though both tend to also congregate on different platforms focusing on the context (e.g. GitHub for programmers and Steam for gamers), sometimes these communities even interact on the same online networks such as Twitch or Reddit. The similarities do not stop here; as already mentioned, both communities developed mainly online and became important resources for the fields (at least in many modern video games). In this sense, both communities based themselves on a "Remix culture" encouraging the sharing and re-elaboration of information, blurring the line between final consumer and contributor [125, 126]. The overall argument in this case is that learning to make use and contribute to a gaming community probably involves similar skills to be able to do the same in a programming community because of these underlying similarities. A great example of an online community not directly connected to a video game (not referring to online gaming necessarily) is the community that formed around the 'Elder Scrolls' series. Many online resources surrounding that series share information with both new and more seasoned players about how to develop their characters and how to customise their game experience. Another more general example is the massive amount of online videos of 'playthroughs' (often referred to as "Let's Play" content) in which players record their game sessions while commenting and explaining their actions to show other players how to achieve a certain goal in a video game. Curiously, we can find similar videos about programming, with (more or less) expert programmers coding and illustrating how to use certain languages, libraries or functions.

2.3.3 Discussion

When highlighting video games as educative media for computational thinking purposes, we often tend to neglect the great variety of genres, mechanics and, more in general, game experiences. However, as we argue in this thesis, the individual constituent game design patterns are perhaps a more fitting lens to assess the potential of a video game to improve computational thinking skills. Such individual elements can trigger very different thinking strategies and stimulate users in different ways. What we proposed is a different way of studying the connection between gaming and the development of programming skills, starting from the design elements that make up a video game rather than from the medium or specific game titles in general. This approach can also be instrumental to the creation of educational video games that make use of design concepts developed specifically for the medium rather than adapting them to the scope. In this paper, we presented examples of design patterns and games that could be investigated regarding their ability to improve computational thinking skills. Further research, both empirical and theoretical in nature, should focus on developing a design process to create or modify video games for that purpose. This would strengthen the case for the use of video games to improve computational thinking skills in general, and deepen our understanding of how to target such efforts towards individual skills.

2.3.4 Section Summary

In the work above, we describe the connections between game design patterns and computational thinking skills. In doing so, we answer one of our research questions RQ3;. We see how specific mental mechanisms that are activated by certain game elements present similarities with computational thinking skills. Moreover, we practically show how game design patterns are a recognised tool to analyse video games and can be applied to the field of games for education. On the other hand, it is important to note how individual games can be played in different ways, and the thinking mechanisms involved can present some degree of variation. This could potentially impact the correspondence between the individual design pattern and computational thinking skill. However, the methodology we present can still be a useful tool in comparative work among research in the field.

In this chapter, we have noticed how game-based learning is particularly susceptible to natural differences arising from the context of application; different subjects,

students, or design choices yield very different results. However, recent advancements in AI-driven procedural content generation (PCG) and adaptive learning environments offer new opportunities to personalise programming education. Building on previous work on AI-assisted game design, it could be interesting to explore how generative AI can be used to create dynamic learning experiences that adapt to different skill levels and learning styles. Another interesting question is whether the increasing integration of generative AI tools in programming education (e.g., code completion with GPT-based models) interacts with traditional learning methods, for example, by investigating if generative AI enhances or diminishes the role of games in developing computational thinking and problem-solving skills. In the coming chapter, we will first discuss aspects of the relationship between AI and video games. We will also investigate to what extent this relationship can bridge the digital gap in the real world in *hybrid* games.

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?

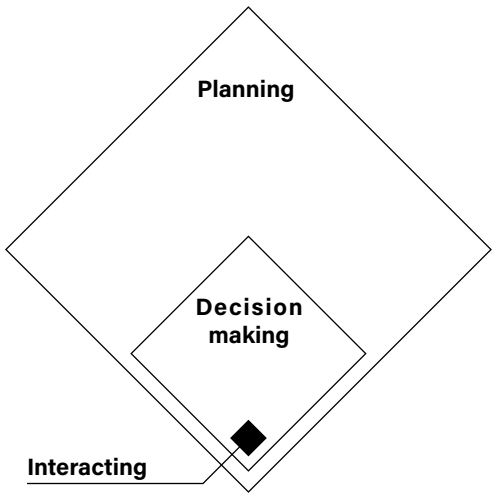

Figure 3.1: The structure of actions involved in meaningfully playing an open-world game

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.

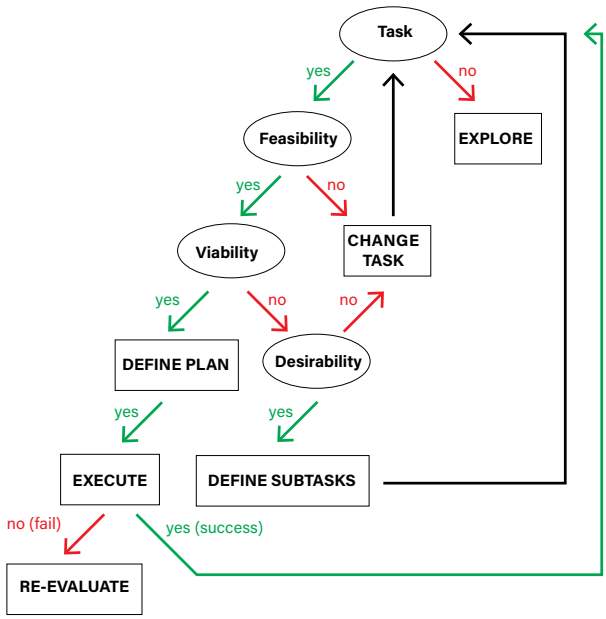

Figure 3.2: The Player Decisions Framework: in ellipses, the evaluation steps, in rectangles, action steps

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.

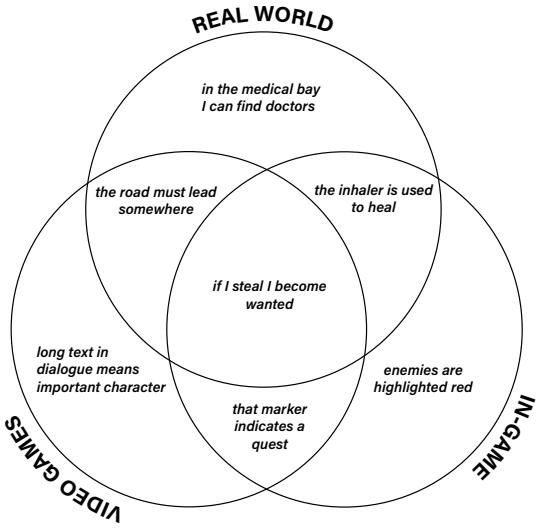

Figure 3.3: The three categories of information utilised to evaluate the game state and make decisions

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.

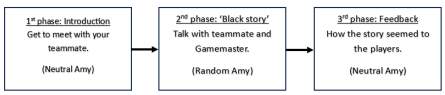

Figure 3.5: The phases of the game.

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

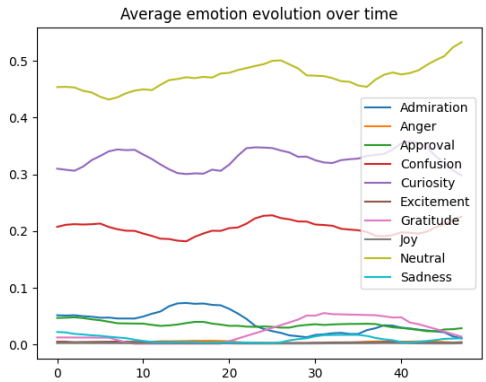

Figure 3.6: Left: Heatmap of the average emotion scores for the different phases of the game. Right: Average emotion scores during phase 2 of the game over time. Scores for a given timestep are computed on segments of the conversation as described in section 3.2.4.

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.

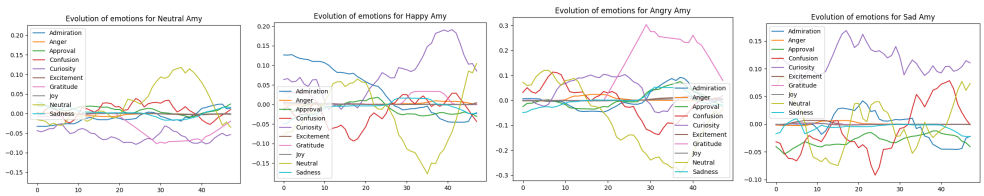

Figure 3.9: Computed scores of player emotions over different Amy profiles during gameplay. Each plot shows the normalised scores on the most affected emotions over time. Scores for a given timestep are computed on segments of the conversation as described in section 3.2.4 and normalised to show how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

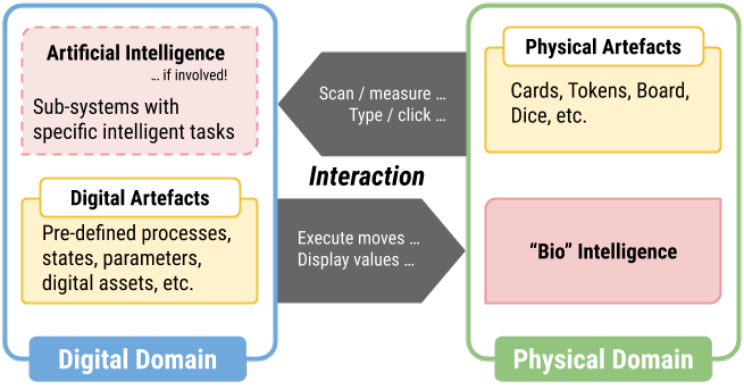

Figure 3.10: Illustration of high-level components involved in a hybrid board game. Both domains involve intentionally designed artefacts and interaction with the other domain. Typically, hybrid board games involve human, or 'bio' intelligence, but may also involve several AI sub-systems (which tend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

Chapter 4

Scientific Education, Video Games, and Artificial Intelligence

In this chapter, we bring together previous information and define the current developments in:

- Generative AI in programming education
- Video Games for Generative AI in education

The goal of this chapter is to discuss and speculate on the impact of generative AI in education by exploring current literature. In the first section, we derive pitfalls but also opportunities for the future. With the perspective that generative AI is here to stay, we propose a direction to limit its negative impact on education while highlighting its potential to motivate and aid students' learning. Spoiler alert: it's games. In the second section, we embark on a simulated design process that has as a fictional goal the one of creating an LLM-powered NPC to teach Roman history in game settings. In the process, we reflect on our design considerations and generalise them for future developers. We also evaluate the final product and draw conclusions on the feasibility of incorporating generative AI (in particular, LLMs) in game-based learning. Each section is connected to the rest of the thesis by a short summary of the findings.

4.1 Generative AI and Programming Education: Considerations from Current Studies

Since the release of GPT in 2022, generative AI has quickly become ubiquitous, being utilised in different human activities. As is often the case, the advancement of disruptive technologies fosters new discussions within old contexts; think about the legal considerations around AI-generated content [204] or the studies about potential applications of generative AI in education [205]. With regard to the latter, many studies highlight the risks of using generative AI in computer science education. For example, with the assistance of GPT models, students are able to complete assignments more quickly, but they also retain less information compared to their peers who worked without AI help [30, 206]. However, other studies in similar settings revealed that students' computational thinking skills improved using generative AI [207]. This section aims to compare selected research in the field in order to clarify the impact of generative AI in programming education. We define the following research questions:

- How is the impact of AI-powered interventions measured?
- What are the results of these interventions?
- How can we reconcile apparently opposite results in the field?
- What can we learn for future empirical research involving generative AI in computer science education?

4.1.1 Literature Review

Our understanding of generative AI's impact on programming education remains limited. A recent literature review about empirical research in the field includes only thirty-seven studies [208]. Of these, only two evaluate students' computational thinking and programming skills. In fact, the majority of the studies are limited to how LLMs perform in terms of programming. These include skills such as debugging [209], pair programming [210] or program generation [211]. However, these skills do not necessarily reflect the impact of generative AI on learning; they depict it as a useful tool for programmers. This seems to be a common misconception in current research in the field, where generative AI performance in teaching environments is evaluated based on its programming performance [208]. Going back to the aforementioned literature review [208], the two studies focusing on students' computational thinking

and programming skills development present very different methodologies. The first one highlights the positive effects on both metrics for students using generative AI. However, the measurement of computational thinking is quite complex and heavily dependent on the model used. In the case of this experiment, computational thinking skills are measured using a scale based on a more abstract model [212]. This includes skills that are not necessarily exclusive or focused on programming: creativity, algorithmic thinking, problem solving, critical thinking, cooperativity and communication skills. However, the most critical flaw of the study is the testing methodology for programming skills development, measured using a self-efficacy scale focused on students' confidence in tackling abstract programming problems [213](see figure 4.1). We argue that, with these tools, conclusively evaluating the actual impact of generative AI on students' programming learning is impossible.

Item No.	Item Description (alpha reliability estimates for factors in parentheses)
	Factor 1: Independence and persistence (alpha = .94)
23	Complete a programming project if I had a lot of time to complete the program.
22	Complete a programming project once someone else helped me get started.
21	Complete a programming project if I could call someone for help if I got stuck.
24	Complete a programming project if I had just the built-in help facility for assistance.
20	Complete a programming project if I had only the language reference manual for help.
19	Complete a programming project if someone showed me how to solve the problem first.
25	Find ways of overcoming the problem if I got stuck at a point while working on a programming project.
17	Debug (correct all the errors) a long and complex program that I had written, and make it work.
	Factor 2: Complex programming tasks (alpha = .94)
13	Understand the object-oriented paradigm.
16	Make use of a class that is already defined, given a clearly labeled declaration of the class.
14	Identify the objects in the problem domain and declare, define, and use them.
8	Build my own C++ libraries.
18	Comprehend a long, complex multi-file program.
12	Organize and design my program in a modular manner.
32	Write a program that someone else could comprehend and add features to at a later date.
15	Make use of a pre-written function, given a clearly labeled declaration of the function.
11	Write a long and complex C++ program to solve any given problem as long as the specifications are clearly defined.
28	Mentally trace through the execution of a long, complex, multi-file program given to me.
29	Rewrite lengthy confusing portions of code to be more readable and clear.

Figure 4.1: Example of statements used for students' self-efficacy evaluation in [213].

The second study has a completely different approach. In this case, generative AI has been implemented in a gamified interface [214]. The article presents the experiment as ongoing and, therefore, actual data about students' performance and improvement

is not available and necessitates further studies. The author only reports positive effects on motivation and acceptance that are typical of a gamified environment (as discussed in chapter 3). However, the interesting element in the study is the innate limitations to generative AI that a gamified environment can present. Both studies claim to address computational thinking and programming skills, but their measurements lack focus on these topics. Studies measuring actual student performance have only recently emerged. For example, a recent experiment in a programming class of Fortran indicates that retention is superior for students who do not use generative AI [206]. In the experimental setup, the experimental groups are allowed a quite free use of various modern generative models. The control group is allowed to use Google exclusively. However, other studies take a different perspective, evaluating generative AI in its ability to tutor students. Current research highlights the potential for LLMs to perform almost as well as human tutors [215]. Obviously, evaluation of tutoring is quite complex, and it often relies on experts. Moreover, in order to make results comparable, many studies limit what students can ask, for example, using preselected prompts [215]. In general, most studies report positive results on students' acceptance and motivation using generative AI [216]. However, others highlight a negative correlation between perceived ease of use and perceived usefulness [217].

4.1.2 Discussion

In this chapter, we use the considerations from the above literature review in order to provide answers to the research questions. It is important to note that our work is far from systematic. It is, in fact, based on a selective exploration of salient studies in the field. Arguably, the low number of empirical studies specifically focusing on programming students' retention and skills development makes systematic approaches inadequate.

How is the impact of AI-powered interventions measured?

Current research primarily measures the acceptance of generative AI among students and teachers. This is often performed through the typical technology acceptance model, a well-studied and validated tool for measurement. On the other hand, other aspects of the impact of generative AI on programming education are more difficult to measure. We argue that metrics and research goals are not always well aligned. In the case of computational thinking skills development, different models can be used as references. However, each model has a different perspective on these skills and select-

ing the best-suited one is a necessary evaluation. As for programming skills, teachers already have many tools to test students' learning. Self-assessment tools have their own reason and space; however, they tend to be more strongly related to the respondent's confidence than their actual development. Confidence is particularly reinforced with the ability to perform a certain job, something that, especially in introductory programming curricula, generative AI can certainly provide. However, we argue that this is not necessarily related to students' proficiency in programming but with their perceived capacity to pass the assignments provided. Other studies strongly focus on human comparison. In these cases, measurements are usually performed by humans who evaluate generative AI's performance compared to other human experts. It is the case of experiments centred around the effect of AI-powered tutoring. Another specific characteristic of this format is that it often relies, by necessity, on predetermined prompts or other forms of limitations that make human and AI tutoring comparable.

What are the results of these interventions?

Empirical research shows a clear improvement in students' motivation. This is definitely a relevant effect, probably related to the ease of use of generative AI and the enthusiasm emerging from a new, and quite frankly impressive, technology. Results emerging from technology acceptance studies (using variations of the technology acceptance model [218]) are also extremely encouraging, especially for younger participants. As mentioned above, many studies also highlight a positive effect on students' confidence and self-assessment. However, these effects do not automatically translate into students' final retention and learning. In this regard, the ability to have at one's disposal immediate solutions may hinder deep learning, as students bypass the work required to internalise concepts. In fact, cognitive load theory suggests that excessive assistance reduces mental effort, preventing students from actively engaging in knowledge construction, no matter if the assistance is by humans or AI. In studies about AI tutoring, human tutors evaluate generated answers positively, almost at the level of human tutoring. However, the performance of AI tutoring without constraints compared to its human counterpart is still unknown.

How can we reconcile apparently opposite results in the field?

The diversity of results in the field can be justified by two main elements:

- The field is still in its infancy; as shown in [208], a related literature review on empirical studies only reports thirty-seven studies. It is natural that this leads

to a great variation of results as the novelty gives great space for exploration.

- Mainly, the results are not necessarily conflicting; effects on motivation and self-efficacy reports do not necessarily translate to performance or retention (as also seen in 3). Students can feel more motivated and more confident in engaging with their tasks, but at the same time, not fully absorb the necessary information.

What can we learn for future empirical research involving generative AI in computer science education?

We have a few learning outcomes from our literature review:

- On a prescriptive note, alignment between research questions and testing methods is fundamental, especially for educational research in younger fields. When focusing on computational thinking skills development, it is valuable to select models that adhere closely to the subject practice (i.e., programming). For example, some models are more focused on programming practice (such as [111]), and they could arguably be better tools to test students' improvement. As for programming skills, we suggest that participants could be tested on their performance in completing assignments without the help of generative AI or, alternatively, with aimed questions, testing specific concepts covered by the curriculum.

- Defining opportunities for future studies, we notice potential in the application of generative AI as tutoring. With the term "tutoring", we refer to contexts in which students are provided with limited AI tools. They are able to use it for hints and directions, but their freedom of interaction is previously regulated by human teachers. We argue that the result is that teachers are able to provide a high number of students with necessary support while retaining control over the learning process. As a corollary of this outcome, current literature seems to suggest that free, unrestricted access to generative AI in programming learning environments can be deleterious.

4.1.3 Future Work and Final Considerations

Future work in the field should focus on two main directions. First, it is important to continue to develop empirical literature to paint a clearer picture of the impact of generative AI on programming education. In particular, additional research focused on the performance and retention of programming knowledge and skills is needed. While

performing evaluations in an actual teaching context is extremely valuable, we have to consider that the ubiquity of generative AI could influence the results in unpredictable ways. Therefore, it could be valuable to start in smaller and more controlled settings. The second direction involves the study of methods to restrict AI and direct its potential towards specific uses. If, as argued above, controlled prompting has a positive impact on education, then we need to investigate how to design frameworks that can act as mediators. In this regard, interesting studies can be developed in the field of games and gamification (as in the case of [214]). These are well-established media for integrating both students and intelligent algorithms.
In this section, we reviewed examples of current empirical research in the field of generative AI for programming education. We have seen some recurring pitfalls and characteristics of related experiments. In particular, we noticed that, in some cases, testing methods should be carefully reviewed to better match research questions. On the other hand, experiments also highlight opportunities for future developments. In this regard, the effect on students' motivation and acceptance is noticeable. Moreover, experiments that include some form of restrictions or control on student-AI interaction yield promising results. Interestingly, in these types of experiments, the AI seems to take more of a tutoring role than a simple problem-solving tool. Generative AI has a definitely disruptive impact on programming education. However, as with many other digital tools, related research can teach us to control this impact and direct it towards having a positive effect. Of course, this requires time to explore different possibilities and free ourselves from potential preconceptions. Our contributions aim to provide a different perspective and, hopefully, some guidance for future research in the field.

4.1.4 Section Summary

The previous section introduces the impact of generative AI, in particular LLMs, in programming education. Many of these considerations can be transposed outside the specific field of application to different subjects. We see that, while a lot of research stemmed from the enthusiasm for these new technologies, it often skipped very important elements related to the analysis of students' performance. As such, many studies cannot conclusively find beneficial effects of LLMs in education. On the other hand, when proper metrics are applied, we see a strong negative effect related to the tendency to plainly take the generated output without a full understanding of it. There is, however, hope within certain specific contexts. For example, research has shown the potential to use LLMs as tutors, using parameters to limit their tendency to

take over students' problem-solving. Others have seen potential in the application of AI agents in gamified environments. In the following section, we follow an explorative process to investigate the potential for LLMs to be used in game-based learning. We follow a game design approach with the goal of creating an LLM-powered NPC in an educational video game. While potentially providing information to win the game, the NPC aims to transmit knowledge in an engaging and personal way.

4.2 Video Games as Mediators of Generative Artificial Intelligence

When it comes to learning, we have seen that games and gamification have strong effects on student motivation and engagement [33]. In fact, this is one of the salient aspects of play in general. However, the impact on final performance and retention is more complex to evaluate and often depends heavily on the subject and game design. In general, it is undeniable that video games are a signature medium of our time, and most people nowadays are able to interact with them intuitively. Their studied effects and accessibility make video games one of the most appealing media for educational research.

Another important digital development of the last 20 years is AI. When it comes to education, AI has a much more complex history. Although we strive to introduce components of AI into many school curricula, its actual impact on education is quite disruptive. In particular, generative AI has been noted as problematic in terms of retention and learning[206, 30]. However, many studies highlight the potential for generative AI to be used as a tutor for students, scaffolding actual learning by focusing more on teaching than direct problem-solving [215]. We talk, in this case, about *restricted* generative AI, introduced within a framework that:

- engages and challenges students, providing both intrinsic and extrinsic motivation [219]
- guides problem-solving without fully engaging with it

The purpose of this section is to highlight critical design decisions and existing challenges involved when introducing generative AI in education video games. In order to do so, we engage in a simulated design process, aiming to create an information game for the purpose of history education. While designing a minimum viable product

for testing, we journal through important decisions to consider about the integration of LLMs in video games. Finally, we evaluate the resulting game in terms of effectiveness in meeting the requirements and its viability within the current technological framework.

4.2.1 Background

Applications of generative AI in game-based learning and gamification (and vice versa) are still quite recent; a recently updated review on studies involving LLMs and games does not report even one involving game-based learning [220]. However, research is starting to develop in different directions. A very prominent one is the use of game elements for AI education [221]. In the same category, more constructivist approaches have been suggested to transmit AI interaction skills *through* play [222]. In this regard, AI plays the role of a topic in these game research studies. Other studies involved LLMs as players' evaluators in the context of education [223]. In this case, the AI is playing the role of analyst. Finally, other studies use generative AI to design narrative-based games for education [224], hence using AI as a designer. In this section, we focus on generative AI as a tool to provide educational information to players within video game contexts. In particular, we involve three fundamental parameters: role-playing, accuracy and viability. When we talk about **role-playing**, we refer to the ability of an LLM to interpret a character situated in the digital context of a specific video game. Current research on the topic attempts to create more reliable personas through the use of detailed profiles [225]. Additionally, role-playing prompting can improve output accuracy [226]. However, evaluating the level of role-playing efficacy is complex, especially when large amounts of conversational data are involved [227]. Some studies propose LLM-powered evaluators, but there are still limitations related to the stochasticity and unreliability of these models [225, 228]. Other studies focus on the use of both demographic information and opinion training in order to achieve the best alignment between LLMs and humans belonging to the same demographics [229]. This opens the creation of intelligent agents able to interact more realistically in their belief network. However, current research remarks that not all roles are played the same. For example, LLMs perform better when they interpret a doctor than a family member or an animal. Additionally, they seem to rely heavily on cultural stereotypes and biases to build their character [230]. Accuracy is another key concern for generative AI and relates to the actual correctness of the information the LLM provides. This is particularly important in the context of our research due

to the educational perspective. We have already mentioned the risk of hallucinations and incorrect pieces of information that are sometimes very difficult to spot [143]. In educational settings, this issue has a deep and troubling impact on the potential applications of generative AI [231]. Moreover, it is exacerbated by the aforementioned risks associated with excessive reliance on LLM-powered tools and the tendency not to critically analyse their output. Finally, viability relates to the actual capacity for the LLM to be embedded in video game systems without negatively impacting their functioning and performance. Previous research investigates the processing power required to run an LLM; it reports encouraging considerations related to the possibility of running smaller models in most AAA games [136]. However, these studies focused on specific contexts that are not necessarily translatable to educational applications, which often rely on video games of a smaller size. At the same time, the requirements for LLMs in educational contexts are quite high, and this can conflict with the use of smaller models.

4.2.2 Research Question

While existing research touches upon several elements related to the use of LLMs in education or video games, our goal is to take a holistic approach to educational game design and LLMs. The central topic of this section is to investigate how LLMs can be implemented in game-based learning. In this regard, we develop a series of subresearch questions:

- **what design considerations influence the development of a game incorporating LLMs with educational purposes?**
  We focus on the design requirements that arise specifically from the incorporation of text-based generative AI.

- **how do characteristics of models and prompting influence the quality of the final product?**
  We test different models with different sizes. We elaborate on how feasible their implementation is. We also experiment with prompting, investigating how well LLMs can follow relevant instructions for educational contexts.

- **how effective is an LLM in balancing role-play and educational content?**
  In the previous questions, we focus on requirements for LLMs as tutors. However, (educational) video games have natural requirements related to consistency

in terms of narrative engagement. An AI-powered NPC needs to be able to maintain a persona in order not to disrupt the player's experience.

- **how viable is this incorporation with current technology for the general public?**
  We summarise the findings from the previous questions, evaluating the actual viability of this type of game-based learning.

4.2.3 Methodology

We simulate a design process for an education app incorporating LLM characters. We aim to develop a minimum viable product to test different parameters and evaluate the final experience in terms of role-playing, accuracy and viability. We report the considerations that arise from the process and the analysis. We decided to contextualise the game project on the subject of Latin history. The choice has been made by considering the opportunities arising from the simulation of interactions with historical civilisations and the relative presence of facts that can be objectively evaluated in the subject. The product takes the form of an information game in which the player is thrown back to the year 1 BC and needs to figure out the location and construction context of the Ara Pacis (Roman monument built in 9 BC). As an educational goal, the player is supposed to acquire knowledge related to everyday life in the Roman Empire by interacting with LLM-powered characters. It is important to understand that the game characteristics can influence the design process and the result of the successive evaluation. For example, certain language models might be inherently more effective in portraying NPCs from other civilisations or eras [230]. Therefore, the results of this study should be considered in the presented context. We evaluate the process and the result with regard to the following parameters:

- Model characteristics: we use different LLMs with different characteristics. We will explore bigger models and smaller ones. We finally evaluate the results in terms of viability and accuracy and identify possible relations with the model characteristics.

- Contextual information: we provide the LLM with contextual information related to its persona and the historical context. We also provide information about the style of the output that is supposed to be generated and its scope. We edit and experiment with prompting to identify possible strategies to improve role-playing and accuracy.

- Integration: considerations related to the integration of the model in the game environment are relatively more straightforward. We evaluate the options of using a local or online LLM and critically discuss the implications with regard to viability.

We finalise our investigation by summarising our findings. We also report on aspects of these systems that are still challenging for current technologies and that necessitate further research.

4.2.4 Results

Design Considerations

When designing educational games involving LLM technologies, specific considerations arise throughout the process. The first element, in the case of generative AI being used to power NPCs, would be the interaction type. Text-based role-playing games (RPG) with natural language as an input component are not unusual; probably, the most famous example in this case is the video game Zork [232]. However, up to today, these games rely on a hard-coded interpretation of the user's textual input. Instead, LLMs allow for actual interpretation and reaction to users' interactions. This impacts design in two directions: the input and output interfaces. In the case of the input interface, the design might need to control the maximum length of users' messages. This can be a challenge with current models (see 4.2.4). Moreover, it is important to filter the input in order to avoid undesirable outputs as a reaction and restrict the possibilities for users to intentionally tamper with the model's alignment. Similarly, the output message might need to be constrained, and this can present a challenge for current technology (see 4.2.4). Finally, especially in educational contexts, the output of an LLM is considered dangerously unpredictable. However, in our experimentation, we find that most models are able to avoid producing inappropriate output. In this regard, the design consideration is to pay attention to experimenting with the selected model. Due to the inherent stochasticity of AI, certain filters for inappropriate responses should still be put in place. The narrative design can also be influenced by the introduction of LLMs. In this regard, the AI can generate content that inadvertently conflicts with the designer's plan. A good practice is to design for flexibility, creating systems of narrative and game objectives that take into account the lack of control over the conversational aspects of the video game. Additionally, in the context of education, it is helpful to have more general learning objectives (such as the one we illustrated for our game in 3.2.4). Conversely, designing with LLMs in

order to elicit specific knowledge transmission can expose to the risk of inaccurate or misdirected information.

Model Characteristics

First, we want to define how the model's size impacts accuracy. Then, we will evaluate the viability of using different models in educational games. When it comes to accuracy, the relation between model size and performance is not necessarily linear [233]. However, with simpler goals, such as in our case, the information quality definitely improves with bigger models. In our case, we experiment with various models presenting three main size groups: between 6 and 8 billion parameters, between 13 and 15 billion, and 70 billion. The choice of size groups has been dictated by sampling reasons. Moreover, bigger models have been excluded for technical availability reasons, which would, anyway, make them intuitively impossible to apply in reality. After experimentation, the models that could be run locally with acceptable speed and with commonly available hardware were only those between 6 and 8 billion parameters. Those between 13 and 15 showed a great improvement in terms of accuracy, but not specifically in terms of role-playing capabilities. Moreover, they demonstrated to be far slower. The models with 70 billion parameters did not show a great improvement in terms of accuracy compared to their 13-15 counterparts.

Contextual Information

Within the context of a video game, we need to build a prompting framework that reinforces specific desirable behaviour from the LLM. The prompt used has a fixed structure that always ends with the formula *"Behave and react to the text only in ways appropriate for your character"*. In our experiment, we manipulate specific aspects that come before the aforementioned formula:

- Character and historical context: we change the level of details we provide to define the character and how they relate to the historical context. The first experiment utilises fully natural language with fluent syntax and goes into relative details about the life and the context of the persona selected. The introductory prompt is *'You are a Latin noblewoman living in Rome in 1 BC during Emperor Augustus's reign named Lucillia.'*. In the second attempt, we play mainly with syntax, building the information in shorter sentences: *'You are a Latin noblewoman named Lucillia. You live in Rome in 1 BC during Emperor Augustus's reign. You were born in 20BC.'*. Finally, we use a less specific prompt

to test whether the result can improve: *'You are a Latin noblewoman in Rome during Emperor Augustus's reign.'*. We experimented with keeping the input "Hello, where is the Ara Pacis?" as the most direct and basic interaction possible. While the agents provide valuable knowledge, each prompting style has some drawbacks; with the first two, the agent does not stay in character. With the last, the output provided is extensive and goes way beyond the simple question. All three styles of prompting output partially incorrect information: the first two, when they break out of character, provide a wrong name for the museum holding the Ara Pacis today. The last one gives an incorrect original location of the monument. In all these incorrect cases, we notice the LLM resorts to more generic and stereotypical expressions, using locations such as the Roman Forum or the Museum of Roman Civilisation (which, by the way, is then incorrectly translated to Italian) in its mistakes. In terms of sheer educational quality, the third prompting style stimulated the most extensive responses and, with these, it provided the most complete information about Roman lifestyle and civilisation.

- Educational context: We provide details about the educational context in which the product is supposed to be applied. In this regard, we experiment with explicitly adding to our input *'Try to also provide educational information about life in the Roman Empire.'*. The results are quite evident; this input does provide a good context for application to the LLM, which is able to respond with much more extensive and detailed information about Roman civilisation. Moreover, it seems able to adapt to its context of application and has a more conversational style in providing answers. However, in this case, the final results often contain incorrect or imprecise information.

- Additional reinforcement: we change the level of explicit information we provide in order to keep the model on the character and avoid providing unrealistic information. The first experiment aims to improve role-playing by explicitly asking the LLM to avoid using information not available for their persona. Therefore, we add the sentence *'Do not use information not available in your historical context.'*. With this addition, the occurrence of incorrectly contextualised messages was greatly reduced. However, the accuracy of the rest of the information provided is not necessarily improved. Additionally, we experiment with a broader, yet simpler, addition: *'Never break character.'*. This edit seems to be the most impactful among those tried. In all cases mentioned above, adding this sentence results in content that is more accurate, contextualised and concise. On the

other hand, the information might be more generic. Finally, we attempt to improve conciseness by adding *'Keep the answer under 60 words.'*. The results are disappointing; the LLM completely disregards this edit in all cases.

Integration

The last component through which we want to analyse the minimum viable product we proposed as a case study is integration. In this regard, we take into account the aforementioned information and discuss its implications in terms of actual viability for future applications of LLMs in educational video games. From experimenting with model characteristics, we come to the conclusion that running models locally dramatically limits the integration of bigger (bigger than 13b) models. As long as we keep the model local, we always have to balance accuracy with the reaction speed of our NPCs. Smaller models that could easily run locally performed extremely poorly for the standards of an educational game. Besides failing at times to role-play and presenting engagement-breaking information, they often presented wrong facts. Bigger models, on the other hand, are not viably runnable on the average laptop and have excessive latency times between input and output. Therefore, using an online (via API) solution is the only way to move forward in terms of integration. In this case, we recommend testing first mid-size models (13 to 15 billion) since they often offer good accuracy without the computational power required by their bigger counterparts. Obviously, there are drawbacks to relying heavily on online models. First, they require a stable internet connection. Second, depending on where the servers are located, they might present some criticalities in terms of privacy. Generally, in terms of integration and its viability, LLMs in video games with educational purposes are required to be used in contexts with considerable resources and, as such, they lose some of the universal applicability that characterises educational games. However, it is definitely possible to design them and run them in some more privileged contexts.

4.2.5 Discussion and Conclusion

Our explorative investigation into the use of LLM for educational video games reveals certainly potential for these systems. However, there are relevant challenges that make the applications of these tools unfeasible at the current time. In terms of design considerations, because of the sensitivity of LLMs to specific forms of inputs, filters are necessary to avoid misalignment. These can even be hardcoded, and while we discover more about AI alignment, we will have a better grasp of what to target precisely. Our

investigation also shows that the risk of random inappropriate output is greatly reduced as long as the aforementioned misalignment filters are in place. However, LLMs tend to still perform poorly when it comes to following plot and persona descriptions. This translates into the necessity for more flexible and generic game narratives, which might be in conflict with the nature of the subject. A big drawback that we highlight both in the design considerations and the prompt experimentation is the difficulty of respecting certain limits to the length of output. This can impact the user interface design.

The characteristics of the model and of the contextual information provided obviously play an important role in the final product performance. Bigger models tend to improve accuracy, which is an important aspect for educational activities. However, when it comes to a locally run model, striking the balance between accuracy and computational performance can be extremely challenging. Smaller models (6-8 billion) tend to perform relatively poorly, at least in the context of this investigation. Bigger models (13-15 billion or above) are quite slow if we consider the average hardware available to students. On the other hand, online integration can, in part, solve these issues, even though some privacy concerns need to be considered. However, using local servers and stable internet connections, these drawbacks can be somewhat limited. As for the contextual information provided, we see how models tend to provide more educational information when we build a more generic persona. Also, specifying the educational settings has a positive impact on the output; it also slightly changes the registry of the agent towards a more "tutoring" tone. Finally, to improve role-playing in the settings of this investigation, we tried to confine the agent, asking it not to provide historical information unavailable to its persona. Although this had some impact, it did not fully avoid incorrect or unbelievable behaviour. Also, in this case, the more generic demand not to ever break character seems to be more effective. In general, we find that LLMs have limited role-playing power that improves when the persona and the boundaries are described more generically. Although this is to be expected, it also strongly limits the control that the designer or educator has over the tool and the narrative.

In general, with regard to our third research question *"how effective is an LLM in balancing role-play and educational content?"*, we can say that, especially in the case of smaller models, current tools perform poorly. While role-playing can be improved with more generic contexts, over many interactions (which would also happen in real educational settings), the LLM has been demonstrated to be unreliable, often providing incorrect information. Often, this information was generated to look correct,

basing itself on stereotypes and likely knowledge. This is definitely a significant challenge for the application of LLM technology in education. However, it also opens opportunities to incorporate output reviewing habits in education. This can be structured as a learning activity in itself with the additional function of mitigating LLMs' hallucinations and inaccuracies.
Finally, when it comes to integration and the related viability, although some of the challenges we mentioned above persist, bigger and more reliable models could be integrated in a bigger system involving online implementations and local servers to provide the necessary computational power. These solutions are, however, extremely expensive and complex to realise. On the other hand, the field is advancing at a rapid pace, especially in terms of making models more lightweight and efficient [234]. In general, considering the unreliability of the information provided, while integration is indeed possible, at the current stage, we argue is not worth the investment. Moreover, one of the founding aspects of game-based education is the possibility for it to be available for many users, transmitting knowledge in engaging settings, even with relatively limited tools. With the current technology, this would not hold true anymore in the case of LLMs for education; as mentioned above, the infrastructure necessary to make it viable is quite extensive and, arguably, available only in privileged contexts. Considering how impactful the digital divide already is and all the challenges yet to face in order to make the environment effective, we argue that this type of educational/game system is not worth the investment necessary.

Limitations

As mentioned in the introductory information, our exploration is extremely limited in many aspects. We did explore the application of LLMs in the specific context we described: the subject of Latin history and the form of an information game. Our considerations can vary widely, even in slightly different settings. However, most of the critical points we highlighted might endure. Moreover, the settings we selected have characteristics that are arguably favourable for LLMs (very defined goals, generic settings and surface knowledge of the topic required). While the change in settings surely has a big impact on our considerations, it is not likely that the overall performance would improve.

4.2.6 Section Summary

In this last section, we mainly tackled RQ8;. We argued in the previous section that AI can be quite performative in the role of tutor for students. While this might hold true, the implementation of the technology in video games presents several obstacles, especially for the educational nature of the settings. The stochasticity and tendency to use and produce generic information make LLM-generated messages unreliable for education. Moreover, the necessary precautions to make the game-LLM system even viable are relatively expensive and arrive anyway to achieve mediocre results. However, it is also important to note that we explore very open topics, in which the LLMs were mostly evaluated based on the quality of their guidance over non-specific information (in our case, Latin culture and civilisation). Other research that we presented in 4.1 showed potential in the use of the technology in contexts involving very specific knowledge. For example, it is likely that generative AI technologies can have viable and positive applications in guiding students in debugging exercises in programming education. While further research and new methods are necessary, and steps need to be taken towards the adaptation of teaching techniques to the existence of these new technologies, we recommend caution in the acritical application of AI in education research.

Chapter 5

Discussion and Conclusions

5.1 Discussion

We now reach the concluding chapter of our thesis. In the following pages, we go back to the research questions presented in chapter 1; using the articles we presented so far, we try to answer them. We then draw conclusions about:

- The importance of video games for programming education.
- The importance of video games for AI development.
- How video games are central meeting points between AI and programming education.

5.1.1 RQ1; How effective are video games in the field of higher scientific education?

Video games are an effective medium in the context of higher scientific education. This includes all the natural sciences, life sciences and engineering from secondary education to university-level courses. Overall, empirical studies in the field show a positive effect on students' motivation and their willingness to engage with the study material. However, the effectiveness on students' performance is still debated, and different studies yield different results. The effect on motivation is probably the most noticeable. As engaging media, video games make learning more enjoyable. Current research tends to minimise this aspect while focusing on its impact on performance. We argue that, although performance is an important indicator of success for empirical

studies in the field, the positive effects on students' motivation already demonstrate the effectiveness of video games in education. It is relevant to remark here that, even though performance is not necessarily improved using video games, it is also not negatively affected. Therefore, as educators, we need to question whether making classroom activities more pleasurable is not already an important achievement. As mentioned, the final performance is often unaffected. However, some studies do indicate positive effects, while a few report negative ones. Our research points out that performance is mostly impacted by game design per se; well-designed games tend to yield better results also in terms of students' performance. Alternatively, it could be relevant to focus on individual design patterns and their impact.
Finally, we should consider the main drawbacks of game-based learning. First, most research in the field is carried out in contexts in which participants are usually more or less familiar with the video game medium. While this is a fair assumption considering the availability of video games today, attention should be paid to completely generalising the starting conditions of participants. For example, video games would need to be introduced differently in contexts where the digital divide is more marked; in this case, precautions in development (e.g., reducing the processing power required, designing for mobile play) should be taken when designing educative video games. However, as we mentioned in 4, this issue represents an important challenge in safely incorporating LLMs in games.

5.1.2 RQ2; How is research in the field currently carried on?

Controlled experiments in the field of higher scientific education and games are usually carried out by experts in the respective educational fields. In turn, this causes a great variety of methods and considerations, especially when it comes to game design. However, considering the impact of game design on final students' performance, we argue that the involvement of experts from the field of games is necessary in order to make video games more effective. Serious video games are naturally interdisciplinary and, therefore, require a variety of skills and knowledge in order to ensure proper functioning. Another point of concern, which derives from similar causes, is the challenges in terms of comparability. Differences in methodology and design processes make studies difficult to compare. This is also caused by a lack of conventions and care with regard to game design. How can we compare two different games when we cannot identify their individual components? In this case, solutions and frameworks can vary. In this thesis, we suggest that game design patterns can be a good starting point to describe

and, subsequently, analyse video games. Similar considerations are valid for the description of the control groups. Because of the inherent interdisciplinarity of serious video games, studies in the field often omit detailed descriptions of how normal teaching in the specific context is carried on. In turn, this represents another challenge for comparability. In general, the field would greatly benefit from better knowledge about game design procedures and more detailed descriptions of standard education techniques.

5.1.3 RQ3; What common affordances connect video games and computer science education?

In chapter 2, we highlighted similar mental affordances between computational thinking and video games. In this regard, we argue that certain game design patterns can require similar mental work to computational thinking skills to be proficiently used. For example, video games often include iterative and recursive design structures. Obviously, there is also the foundational element of the nature of the medium; video games are inherently played on computers, fundamentally engaging users in interacting with them. In other words, if frequent utilisation of computers gets people accustomed to them, video games can be more pleasant than most other software to do so.

5.1.4 RQ4; How do games present challenges for artificial intelligence development and study?

As mentioned in chapter 1, games have historically been a fertile ground for AI experimentation due to their ease of implementation and evaluation. As AI technology developed, so did video games. Nowadays, new challenges arise from video games with non-linear structures. Open-world games, for example, do not have a linear narrative or prescribed objectives. These characteristics present great hurdles for AI systems. However, they also raise opportunities for development; in order to study how intelligent algorithms can tackle these challenges, we can look at human gameplay as an example of problem-solving in complex contexts. Codifying and generalising human strategies can provide valuable insights for AI development.

5.1.5 RQ5; How does artificial intelligence impact the development of hybrid games?

Hybrid games are games that build a bridge between the digital and the real world. Already today, intelligent agents are present in multiple roles, such as adversaries to allies. Current technological developments, in particular the rise of generative AI, will probably further extend the digital features of hybrid games. Generative AI can enhance agent behaviour and introduce innovative game mechanics. Hybrid games stand out as an interesting medium because they retain the AI-human interaction potentials of games while they transfer this interaction to a middle ground between the digital and the real.

5.1.6 RQ6; How does AI impact programming education?

Recent AI developments in the field of generative intelligence have a definitely disruptive effect on programming education. Current research in the field indicates that applications of generative AI in the classroom are usually met with relative enthusiasm, and they positively affect students' motivation. However, other empirical studies highlight a negative effect of the unrestricted use of generative tools on students' retention. Even though they feel more confident and engaged using AI tools, they seem to learn less compared to their peers who are not allowed to use these systems. Conversely, alternative research directions highlight potential applications. In particular, generative AI seems to be quite effective if used as a programming tutor, provided that prompting has been restricted to predetermined inputs. In this case, we argue that the intelligent system does not simply provide the answer but stimulates users to think about a solution while providing relevant hints. We conclude that unrestricted generative AI can have a negative impact on actual students' learning. However, restricted systems can provide relevant guidance while still allowing students to foster their own learning.

5.1.7 RQ7; How does the implementation of AI in video games perform in educational settings?

Implementing generative AI in educational games, whether it is game-based learning or gamification, is an interesting idea with a lot of potential. In particular, NPCs are ideal cases to apply AI technologies to communicate with humans, which makes them also ideal for tutoring. However, there are some criticalities that arise from the

design process to the actual viability of the resulting product. In the design process, particular care should be taken in filtering the type of input accepted in order to retain some control over the output. AI alignment with educational goals is, in particular, a very important aspect. Our findings indicate that LLMs struggle with role-playing in highly specific contexts. However, more generic contexts can conflict with the game narrative and generate off-topic material. Moreover, smaller models tend to output a lot of incorrect or imprecise information. In this regard, larger models are more reliable but require a bigger infrastructure to make them available to a bigger population of learners. Overall, the educational settings create much higher standards, which generative AI still struggles to reach without a big associated investment. Therefore, we argue that an actual implementation of AI in video games for educational purposes is not yet viable and still presents important challenges that need to be overcome.
On the other hand, AI development is advancing rapidly. New models are quickly becoming smaller and more powerful. This leads us to believe that, in the future, the implementation of AI in video games will become a reality. In order to overcome those challenges and shortcomings, we investigated in 3 and 4, AI models will need to improve their reasoning skills. For example, if we aim to allow AI some form of control over a game narrative, we will need it to be able to reflect and explain its narrative structure. In this sense, advancements in the field of explainable AI will be essential for the introduction of LLMs in game development. Another frequent characteristic of video games is that they can easily reach a large and diverse population. AI-game implementations will need to reckon with this and, especially in educational contexts, consider working towards better AI alignment in order to provide safe and culturally relevant information.

5.2 Conclusions

In this thesis, we explored the intersection between video games, education and AI. We highlighted the impact that video games have on human development, personally, but also historically. We also described games that defined important movements and moments in our history. Subsequently, we delved into the research about the use of games and game elements in education. We have seen the diversity in the field and how empirical investigations have flourished in the last decades. However, we found comparing these studies challenging because of the lack of common practices and vocabulary. Moreover, often, these studies are carried on by experts in the respective fields of application without involving colleagues with expertise in game research. We

then moved on to more specific topics, analysing how fundamental game elements have similarities with computational thinking skills. We speculated over the effect of video game design patterns on players and how they can stimulate similar thinking patterns as computational thinking.
In the third chapter, we analysed the relationship between AI and video games. Since video games always represent ideal challenges for AI algorithms, we propose the next one: open-world games. We also explore the characteristics that make this genre particularly fascinating for the study of AI and propose a different approach to study it. We argue that human player-based heuristics can provide valuable information to train the next generation of artificial players. In the second part of the chapter, we start analysing the implementation of generative AI, specifically LLMs, in video games and their impact on humans. We see how we are able to design intelligent agents to behave in very specific ways and, in turn, impact players' emotional states. Moreover, this highlights how video games can be important points of contact for the study of human-AI interaction. The final part of the chapter explores applications of AI in different play contexts, in particular, physical games, which then take the form of hybrid games. We propose a taxonomy with the goal of showcasing possible points of application of AI outside of strictly digital contexts. Often, the intelligent agent places itself as an embodied figure, portraying an opponent or an assistant. This opens to other considerations in terms of believability and role-playing.
Our fourth chapter attempts to analyse the relation between the three pillars we listed above. We start by analysing the impact of AI in education through existing literature. We see how research in the field often struggles to detach itself from the inherent enthusiasm for new AI technologies (in particular LLMs) and proposes metrics that are not representative of learners' performance. We also see how, when performance is actually analysed, the impact of unrestricted generative AI is disruptive and yields worse results compared to traditional methods. However, other researchers obtained more encouraging results applying LLMs in restricted settings and by limiting users' input or framing the interaction in gamified environments. In the second part of the chapter, we tested this approach by designing a simple educational game. We kept notes throughout the design process and reported our considerations. Additionally, we evaluated the final result in terms of its effectiveness in role-playing and accuracy. We then analysed the viability of these systems. We argued that applying LLMs in games for educational purposes presents several weaknesses. The first one is accuracy, intended as the tendency to output likely but incorrect information. There are also important drawbacks in terms of technical implementation and the necessity of invest-

ments that can hardly be justified considering the final performance.
The goal of the thesis was to explore the relations among education, video games and AI. We speculated over many exciting opportunities that arise from these connections. We have highlighted successful applications that are having a positive impact in actual teaching settings. We have also identified challenges in integrating generative AI technologies. We are still confident in the potential for LLMs to be implemented successfully in video games and, more importantly, in the potential of video games to frame LLMs and control their use in order to empower education instead of disrupting it. However, as researchers, it is important to be aware of all the limitations and to be able to see beyond the enthusiasm that new potential opportunities might elicit. We believe that a careful study of the role intelligent agents play, the development of proper design methodologies and practices for educational game research, and critical analyses of the results are necessary to develop effective applications of generative AI in games and education.
Moreover, we should pay attention to the ethical ramifications of these technologies. Video games have proved themselves to be flexible media which can bring education and knowledge to a large population. They can also connect players from very different socio-economic contexts in playful environments. However, LLMs are currently still expensive, especially with respect to safety and privacy measures. The development of educational games with AI components should take this into account and strive to preserve the ecumenical value of game-based learning.

Bibliography

[1] N. K. Lai, T. F. Ang, L. Y. Por, and C. S. Liew, "The impact of play on child development - a literature review," *European Early Childhood Education Research Journal*, vol. 26, pp. 625–643, 9 2018.

[2] S. Ahmad, M. Phil, and M. Malik, "Play and cognitive development: Formal operational perspective of piaget's theory," *Journal of Education and Practice*, vol. 7, 2016.

[3] G. Frasca, "Simulation versus narrative: Introduction to ludology," in *The video game theory reader*, pp. 221–235, Routledge, 2013.

[4] S. Russels, "Spacewar!," 1962.

[5] K. Thompson, "Space travel," 1969.

[6] IBM and Board of Cooperative Educational Services of Westchester County, New York, "The sumerian game," 1964.

[7] J. F. Kihlstrom, "Ecological validity and "ecological validity"," *Perspectives on Psychological Science*, vol. 16, pp. 466–471, 3 2021.

[8] A. M. Bean, R. K. Nielsen, A. J. van Rooij, and C. J. Ferguson, "Video game addiction: The push to pathologize video games," *Professional Psychology: Research and Practice*, vol. 48, pp. 378–389, 10 2017.

[9] S. Deterding, D. Dixon, R. Khaled, and L. Nacke, "From game design elements to gamefulness: Defining "Gamification"," in *Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments*, MindTrek '11, (New York, NY, USA), pp. 9–15, Association for Computing Machinery, 2011.

[10] K. S. Tekinbas and E. Zimmerman, *Rules of play: Game design fundamentals*. MIT press, 2003.

[11] Colossal Order, "Cities: Skyline," 2015.

[12] B. von Reisswitz, "Anleitung zur darstellung militairischer manöver mit dem apparat des kriegs-spieles.," 1824.

[13] Women's Social and Political Union, "Pank-a-squith," 1909.

[14] S. Dadzie, "Womanopoly," 1970.

[15] W. Güth, P. Ockenfels, and M. Wendel, "Cooperation based on trust. an experimental investigation," *Journal of Economic Psychology*, vol. 18, no. 1, pp. 15–43, 1997.

[16] D. A. Washburn, "The games psychologists play (and the data they provide)," *Behavior Research Methods, Instruments, & Computers*, vol. 35, no. 2, pp. 185–193, 2003.

[17] K. Robson, K. Plangger, J. H. Kietzmann, I. McCarthy, and L. Pitt, "Is it all a game? understanding the principles of gamification," *Business Horizons*, vol. 58, pp. 411–420, 7 2015.

[18] R. S. Alsawaier, *The Effect of Gamification on Motivation and Engagement in Three WSU College Courses*. PhD thesis, Washington State University, 2018.

[19] J. Hamari, K. Huotari, and J. Tolvanen, "Gamification and Economics," in *The Gameful World: Approaches, Issues, Applications*, The MIT Press, 01 2015.

[20] K. Sanford, L. J. Starr, L. Merkel, and S. B. Kurki, "Serious games: Video games for good?," *E-Learning and Digital Media*, vol. 12, pp. 90–106, 1 2015.

[21] Akili Interactive, "Endeavorrx," 2020.

[22] S. W. Evans, T. P. Beauchaine, A. Chronis-Tuscano, S. P. Becker, A. Chacko, R. Gallagher, C. M. Hartung, M. J. Kofler, B. K. Schultz, L. Tamm, *et al.*, "The efficacy of cognitive videogame training for adhd and what fda clearance means for clinicians," *Evidence-Based Practice in Child and Adolescent Mental Health*, vol. 6, no. 1, pp. 116–130, 2021.

[23] J. L. Plass, B. D. Homer, and C. K. Kinzer, "Foundations of game-based learning," *Educational Psychologist*, vol. 50, pp. 258–283, 10 2015.

[24] G. N. Yannakakis and J. Togelius, *Artificial Intelligence and Games*. Springer, 2018.

[25] Ludeon Studios, "Rimworld," 2018.

[26] Hello Games, "No man's sky," 2016.

[27] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. V. D. Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, pp. 484–489, 1 2016.

[28] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, *et al.*, "Grandmaster level in starcraft ii using multi-agent reinforcement learning," *Nature*, vol. 575, no. 7782, pp. 350–354, 2019.

[29] B. Baker, I. Kanitscheider, T. Markov, Y. Wu, G. Powell, B. McGrew, and I. Mordatch, "Emergent tool use from multi-agent autocurricula," *arXiv preprint arXiv:1909.07528*, 2019.

[30] E. Klopfer, J. Reich, H. Abelson, and C. Breazeal, "Generative AI and K-12 Education: An MIT Perspective," *An MIT Exploration of Generative AI*, mar 27 2024. https://mit-genai.pubpub.org/pub/4k9msp17.

[31] P. Ravi, A. Broski, G. Stump, H. Abelson, E. Klopfer, and C. Breazeal, "Understanding teacher perspectives and experiences after deployment of ai literacy curriculum in middle-school classrooms," *arXiv preprint arXiv:2312.04839*, 2023.

[32] P. Denny, J. Prather, B. A. Becker, J. Finnie-Ansley, A. Hellas, J. Leinonen, A. Luxton-Reilly, B. N. Reeves, E. A. Santos, and S. Sarsa, "Computing education in the era of generative ai," *Communications of the ACM*, vol. 67, no. 2, pp. 56–67, 2024.

[33] G. Barbero, M. M. Bonsangue, and F. F. J. Hermans, "How to evaluate games in education: A literature review," in *Smart Learning for A Sustainable Society* (C. Anutariya, D. Liu, Kinshuk, A. Tlili, J. Yang, and M. Chang, eds.), (Singapore), pp. 32–41, Springer Nature Singapore, 2023.

[34] G. Barbero, M. A. Gómez-Maureira, and F. F. J. Hermans, "Computational thinking through design patterns in video games," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, 9 2020.

[35] G. Barbero, M. Müller-Brockhausen, and M. Preuss, *Challenges of Open World Games for AI: Insights from Human Gameplay*, p. 127–141. Springer Nature Singapore, Nov. 2024.

[36] A. Marincioni, M. Miltiadous, K. Zacharia, R. Heemskerk, G. Doukeris, M. Preuss, and G. Barbero, "The effect of llm-based npc emotional states on player emotions: An analysis of interactive game play," in *2024 IEEE Conference on Games (CoG)*, pp. 1–6, 2024.

[37] M. A. Gómez-Maureira, G. Barbero, M. Freese, and M. Preuss, "Towards a taxonomy of ai in hybrid board games," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, 9 2020.

[38] G. Barbero, R. Albrecht, C. Daske, and M. van Noordenne, *Emotion Recognition: Benefits and Human Rights in VR Environments*, p. 17–32. Springer Nature Switzerland, Oct. 2024.

[39] M. A. Gómez-Maureira, I. Kniestedt, G. Barbero, H. Yu, and M. Preuss, "An explorer's journal for machines: Exploring the case of cyberpunk 2077," *Journal of Gaming & Virtual Worlds*, vol. 14, p. 111–135, Apr. 2022.

[40] J. M. Wing, "Computational thinking," *Communications of the ACM*, vol. 49, no. 3, pp. 33–35, 2006.

[41] K. Brennan and M. Resnick, "New frameworks for studying and assessing the development of computational thinking," in *Proceedings of the 2012 annual meeting of the American educational research association, Vancouver, Canada*, vol. 1, p. 25, 2012.

[42] E. L. Deci and R. M. Ryan, *Intrinsic motivation and self-determination in human behavior*. Springer Science & Business Media, 2013.

[43] D. J. Shernoff, M. Csikszentmihalyi, B. Schneider, and E. S. Shernoff, "Student Engagement in High School Classrooms from the Perspective of Flow Theory," in *Applications of Flow in Human Development and Education*, pp. 475–494, Dordrecht: Springer Netherlands, 2014.

[44] S. Papert, "Personal computing and its impact on education," *The computer in the school: Tutor, tool, tutee*, pp. 197–202, 1980.

[45] S. A. Papert, *Mindstorms: Children, computers, and powerful ideas*. Basic books, 2020.

[46] D. King, F. Greaves, C. Exeter, and A. Darzi, "'gamification': Influencing health behaviours with games," 3 2013.

[47] E. A. Edwards, J. Lumsden, C. Rivas, L. Steed, L. A. Edwards, A. Thiyagarajan, R. Sohanpal, H. Caton, C. J. Griffiths, M. R. Munafò, S. Taylor, and R. T. Walton, "Gamification for health promotion: systematic review of behaviour change techniques in smartphone apps," 10 2016.

[48] T. Hainey, T. M. Connolly, E. A. Boyle, A. Wilson, and A. Razak, "A systematic literature review of games-based learning empirical evidence in primary education," *Computers & Education*, vol. 102, pp. 202–223, 2016.

[49] M. Hassenzahl, M. Burmester, and F. Koller, "Attrakdiff: Ein fragebogen zur messung wahrgenommener hedonischer und pragmatischer qualität," *Mensch & Computer 2003: Interaktion in Bewegung*, pp. 187–196, 2003.

[50] M. Liu, L. Horton, J. Olmanson, and P. Toprac, "A study of learning and motivation in a new media enriched environment for middle school science," *Educational technology research and development*, vol. 59, pp. 249–265, 2011.

[51] B. Huang and K. F. Hew, "Do points, badges and leaderboard increase learning and activity: A quasi-experiment on the effects of gamification," in *Proceedings of the 23rd international conference on computers in education*, pp. 275–280, 2015.

[52] A. Domínguez, J. Saenz-de Navarrete, L. De-Marcos, L. Fernández-Sanz, C. Pagés, and J.-J. Martínez-Herráiz, "Gamifying learning experiences: Practical implications and outcomes," *Computers & education*, vol. 63, pp. 380–392, 2013.

[53] C. Melo, L. Madariaga, M. Nussbaum, R. Heller, S. Bennett, C.-C. Tsai, and J. van Braak, "Editorial: Educational technology and addictions," *Computers & Education*, vol. 145, p. 103730, 2020.

[54] T. H. Laine and R. S. Lindberg, "Designing engaging games for education: A systematic literature review on game motivators and design principles," *IEEE Transactions on Learning Technologies*, vol. 13, no. 4, pp. 804–821, 2020.

[55] J. Hamari, J. Koivisto, and H. Sarsa, "Does gamification work?–a literature review of empirical studies on gamification," in *2014 47th Hawaii international conference on system sciences*, pp. 3025–3034, Ieee, 2014.

[56] N. J. Falkner and K. E. Falkner, "" whither, badges?" or" wither, badges!" a metastudy of badges in computer science education to clarify effects, significance and influence," in *Proceedings of the 14th Koli Calling international conference on computing education research*, pp. 127–135, 2014.

[57] Y. Bouzid, M. A. Khenissi, F. Essalmi, and M. Jemni, "Using educational games for sign language learning-a signwriting learning game: Case study," *Journal of Educational Technology & Society*, vol. 19, no. 1, pp. 129–141, 2016.

[58] S. Vandercruysse, M. Vandewaetere, F. Cornillie, and G. Clarebout, "Competition and students' perceptions in a game-based language learning environment," *Educational Technology Research and Development*, vol. 61, pp. 927–950, 2013.

[59] M. Hartono, M. A. Candramata, K. N. Adhyatmoko, and B. Yulianto, "Math education game for primary school," in *2016 International Conference on Information Management and Technology (ICIMTech)*, pp. 93–96, IEEE, 2016.

[60] J. N. Wanyama, B. Castelnuovo, G. Robertson, K. Newell, J. B. Sempa, A. Kambugu, Y. C. Manabe, and R. Colebunders, "A randomized controlled trial to evaluate the effectiveness of a board game on patients' knowledge uptake of hiv and sexually transmitted diseases at the infectious diseases institute, kampala, uganda," *JAIDS Journal of Acquired Immune Deficiency Syndromes*, vol. 59, no. 3, pp. 253–258, 2012.

[61] "Leiden university libraries - leiden university."

[62] C. Barros, A. A. Carvalho, and A. Salgueiro, "The effect of the serious game tempoly on learning arithmetic polynomial operations," *Education and Information Technologies*, vol. 25, pp. 1497–1509, 2020.

[63] M. Dankbaar, "Serious games and blended learning; effects on performance and motivation in medical education," *Perspectives on medical education*, vol. 6, pp. 58–60, 2017.

[64] D. P. de Sena, D. D. Fabrício, V. D. da Silva, L. C. Bodanese, and A. R. Franco, "Comparative evaluation of video-based on-line course versus serious game for training medical students in cardiopulmonary resuscitation: a randomised trial," *PloS one*, vol. 14, no. 4, p. e0214722, 2019.

[65] D. Drummond, P. Delval, S. Abdenouri, J. Truchot, P.-F. Ceccaldi, P. Plaisance, A. Hadchouel, and A. Tesnière, "Serious game versus online course for pretraining medical students before a simulation-based mastery learning course on cardiopulmonary resuscitation: A randomised controlled study," *European Journal of Anaesthesiology— EJA*, vol. 34, no. 12, pp. 836–844, 2017.

[66] I. García and E. Cano, "A computer game for teaching and learning algebra topics at undergraduate level," *Computer Applications in Engineering Education*, vol. 26, no. 2, pp. 326–340, 2018.

[67] V. Garneli, M. Giannakos, and K. Chorianopoulos, "Serious games as a malleable learning medium: The effects of narrative, gameplay, and making on students' performance and attitudes," *British Journal of Educational Technology*, vol. 48, no. 3, pp. 842–859, 2017.

[68] P. Haubruck, F. Nickel, J. Ober, T. Walker, C. Bergdolt, M. Friedrich, B. P. Müller-Stich, F. Forchheim, C. Fischer, G. Schmidmaier, *et al.*, "Evaluation of app-based serious gaming as a training method in teaching chest tube insertion to medical students: randomized controlled trial," *Journal of medical Internet research*, vol. 20, no. 5, p. e195, 2018.

[69] A. J. Q. Tan, C. C. S. Lee, P. Y. Lin, S. Cooper, L. S. T. Lau, W. L. Chua, and S. Y. Liaw, "Designing and evaluating the effectiveness of a serious game for safe administration of blood transfusion: A randomized controlled trial," *Nurse education today*, vol. 55, pp. 38–44, 2017.

[70] K. J. M. Klit, K. S. Pedersen, and H. Stege, "A prospective cohort study of game-based learning by digital simulation of a pig farm to train agriculture students to reduce piglet mortality," *Porcine health management*, vol. 4, pp. 1–8, 2018.

[71] P. Phungoen, S. Promto, S. Chanthawatthanarak, S. Maneepong, K. Apiratwarakul, P. Kotruchin, and T. Mitsungnern, "Precourse preparation using a serious smartphone game on advanced life support knowledge and skills: randomized controlled trial," *Journal of Medical Internet Research*, vol. 22, no. 3, p. e16987, 2020.

[72] D. Rodríguez-Cerezo, A. Sarasa-Cabezuelo, M. Gómez-Albarrán, and J.-L. Sierra, "Serious games in tertiary education: A case study concerning the comprehension of basic concepts in computer language implementation courses," *Computers in Human Behavior*, vol. 31, pp. 558–570, 2014.

[73] C.-Y. Tsai, H.-s. Lin, and S.-C. Liu, "The effect of pedagogical game model on students' pisa scientific competencies," *Journal of Computer Assisted Learning*, vol. 36, no. 3, pp. 359–369, 2020.

[74] R. A. Tubelo, F. F. Portella, M. A. Gelain, M. M. C. de Oliveira, A. E. F. de Oliveira, A. Dahmer, and M. E. B. Pinto, "Serious game is an effective learning method for primary health care education of medical students: A randomized controlled trial," *International journal of medical informatics*, vol. 130, p. 103944, 2019.

[75] M. Winter, R. Pryss, T. Probst, and M. Reichert, "Learning to read by learning to write: Evaluation of a serious game to foster business process model comprehension," *JMIR Serious Games*, vol. 8, no. 1, p. e15374, 2020.

[76] M. Yallihep and B. Kutlu, "Mobile serious games: Effects on students' understanding of programming concepts and attitudes towards information technology," *Education and Information Technologies*, vol. 25, no. 2, pp. 1237–1254, 2020.

[77] H. Hosseini, M. Hartt, and M. Mostafapour, "Learning is child's play: Game-based learning in computer science education," *ACM Transactions on Computing Education (TOCE)*, vol. 19, no. 3, pp. 1–18, 2019.

[78] B. Brezovszky, J. McMullen, K. Veermans, M. M. Hannula-Sormunen, G. Rodríguez-Aflecht, N. Pongsakdi, E. Laakkonen, and E. Lehtinen, "Effects of a mathematics game-based learning environment on primary school students' adaptive number knowledge," *Computers & Education*, vol. 128, pp. 63–74, 2019.

[79] M. Boeker, P. Andel, W. Vach, and A. Frankenschmidt, "Game-based e-learning is more effective than a conventional instructional method: a randomized controlled trial with third-year medical students," *PloS one*, vol. 8, no. 12, p. e82328, 2013.

[80] A. Khan, F. H. Ahmad, and M. M. Malik, "Use of digital game based learning and gamification in secondary school science: The effect on student engagement, learning and gender difference," *Education and Information Technologies*, vol. 22, pp. 2767–2804, 2017.

[81] R. Lorenzo-Alvarez, T. Rudolphi-Solero, M. J. Ruiz-Gomez, and F. Sendra-Portero, "Game-based learning in virtual worlds: a multiuser online game for medical undergraduate radiology education within second life," *Anatomical sciences education*, vol. 13, no. 5, pp. 602–617, 2020.

[82] M. Mavromihales, V. Holmes, and R. Racasan, "Game-based learning in mechanical engineering education: Case study of games-based learning application in computer aided design assembly," *International Journal of Mechanical Engineering Education*, vol. 47, no. 2, pp. 156–179, 2019.

[83] S. Perini, R. Luglietti, M. Margoudi, M. Oliveira, and M. Taisch, "Learning and motivational effects of digital game-based learning (dgbl) for manufacturing education–the life cycle assessment (lca) game," *Computers in Industry*, vol. 102, pp. 40–49, 2018.

[84] O. Ku, S. Y. Chen, D. H. Wu, A. C. Lao, and T.-W. Chan, "The effects of game-based learning on mathematical confidence and performance: High ability vs. low ability," *Journal of Educational Technology & Society*, vol. 17, no. 3, pp. 65–78, 2014.

[85] S. B. Bayram and N. Caliskan, "Effect of a game-based virtual reality phone application on tracheostomy care education for nursing students: A randomized controlled trial," *Nurse education today*, vol. 79, pp. 25–31, 2019.

[86] O. Ak and B. Kutlu, "Comparing 2d and 3d game-based learning environments in terms of learning gains and student perceptions," *British Journal of Educational Technology*, vol. 48, no. 1, pp. 129–144, 2017.

[87] M. S. Jong, "Does online game-based learning work in formal education at school? a case study of visole," *Curriculum Journal*, vol. 26, no. 2, pp. 249–267, 2015.

[88] C.-L. D. Chen, T.-K. Yeh, and C.-Y. Chang, "The effects of game-based learning and anticipation of a test on the learning outcomes of 10th grade geology students," *Eurasia Journal of Mathematics, Science and Technology Education*, vol. 12, no. 5, pp. 1379–1388, 2016.

[89] M.-T. Cheng, W.-Y. Huang, and M.-E. Hsu, "Does emotion matter? an investigation into the relationship between emotions and science learning outcomes in a game-based learning environment," *British Journal of Educational Technology*, vol. 51, no. 6, pp. 2233–2251, 2020.

[90] G. Y.-M. Kao, C.-H. Chiang, and C.-T. Sun, "Customizing scaffolds for game-based learning in physics: Impacts on knowledge acquisition and game design creativity," *Computers & Education*, vol. 113, pp. 294–312, 2017.

[91] N.-Z. Legaki, N. Xi, J. Hamari, K. Karpouzis, and V. Assimakopoulos, "The effect of challenge-based gamification on learning: An experiment in the context of statistics education," *International journal of human-computer studies*, vol. 144, p. 102496, 2020.

[92] T. Radnai, T. T. Juhász, A. Juhász, and P. Jenei, "Educational experiments with motion simulation programs: can gamification be effective in teaching mechanics?," in *Journal of Physics: Conference Series*, vol. 1223, p. 012006, IOP Publishing, 2019.

[93] M. Ortiz-Rojas, K. Chiluiza, and M. Valcke, "Gamification through leaderboards: An empirical study in engineering education," *Computer Applications in Engineering Education*, vol. 27, no. 4, pp. 777–788, 2019.

[94] M. Jurgelaitis, L. Čeponienė, J. Čeponis, and V. Drungilas, "Implementing gamification in a university-level uml modeling course: A case study," *Computer Applications in Engineering Education*, vol. 27, no. 2, pp. 332–343, 2019.

[95] M. Dicks and F. Romanelli, "Impact of novel active-learning approaches through ibooks and gamification in a reformatted pharmacy course," *American Journal of Pharmaceutical Education*, vol. 83, no. 3, p. 6606, 2019.

[96] L. Gutiérrez-Puertas, V. V. Márquez-Hernández, P. Román-López, M. J. Rodríguez-Arrastia, C. Ropero-Padilla, and G. Molina-Torres, "Escape rooms as a clinical evaluation method for nursing students," *Clinical Simulation in Nursing*, vol. 49, pp. 73–80, 2020.

[97] I. Yildirim, "The effects of gamification-based teaching practices on student achievement and students' attitudes toward lessons," *The Internet and Higher Education*, vol. 33, pp. 86–92, 2017.

[98] C. H.-H. Tsay, A. Kofinas, and J. Luo, "Enhancing student learning experience with technology-mediated gamification: An empirical study," *Computers & Education*, vol. 121, pp. 1–17, 2018.

[99] A. Ahmad, F. Zeshan, M. S. Khan, R. Marriam, A. Ali, and A. Samreen, "The impact of gamification on learning outcomes of computer science majors," *ACM Transactions on Computing Education (TOCE)*, vol. 20, no. 2, pp. 1–25, 2020.

[100] D. R. Sanchez, M. Langer, and R. Kaur, "Gamification in the classroom: Examining the impact of gamified quizzes on student learning," *Computers & Education*, vol. 144, p. 103666, 2020.

[101] G. Goehle and J. Wagaman, "The impact of gamification in web based homework," *Primus*, vol. 26, no. 6, pp. 557–569, 2016.

[102] J. A. Stansbury and D. R. Earnest, "Meaningful gamification in an industrial/organizational psychology course," *Teaching of Psychology*, vol. 44, no. 1, pp. 38–45, 2017.

[103] M. D. Hanus and J. Fox, "Assessing the effects of gamification in the classroom: A longitudinal study on intrinsic motivation, social comparison, satisfaction, effort, and academic performance," *Computers & education*, vol. 80, pp. 152–161, 2015.

[104] J. Jo, H. Jun, and H. Lim, "A comparative study on gamification of the flipped classroom in engineering education to enhance the effects of learning," *Computer Applications in Engineering Education*, vol. 26, no. 5, pp. 1626–1640, 2018.

[105] C. Kroustalli and S. Xinogalos, "Studying the effects of teaching programming to lower secondary school students with a serious game: a case study with python and codecombat," *Education and Information Technologies*, vol. 26, no. 5, pp. 6069–6095, 2021.

[106] S. I. Ch'ng, Y. C. Low, Y. L. Lee, W. C. Chia, and L. S. Yeong, "Video games: A potential vehicle for teaching computational thinking," *Computational thinking education*, pp. 247–260, 2019.

[107] M. L. Wu and K. Richards, "Facilitating computational thinking through game design," in *Edutainment Technologies. Educational Games and Virtual Reality/Augmented Reality Applications: 6th International Conference on E-learning and Games, Edutainment 2011, Taipei, Taiwan, September 2011. Proceedings 6*, pp. 220–227, Springer, 2011.

[108] M. Papastergiou, "Digital game-based learning in high school computer science education: Impact on educational effectiveness and student motivation," *Computers & education*, vol. 52, no. 1, pp. 1–12, 2009.

[109] M. Prensky, "Digital game-based learning," *Computers in entertainment (CIE)*, vol. 1, no. 1, pp. 21–21, 2003.

[110] D. Weintrop, N. Holbert, M. S. Horn, and U. Wilensky, "Computational thinking in constructionist video games," *International Journal of Game-Based Learning (IJGBL)*, vol. 6, no. 1, pp. 1–17, 2016.

[111] C. Kazimoglu, M. Kiernan, L. Bacon, and L. MacKinnon, "Learning programming at the computational thinking level via digital game-play," in *Procedia Computer Science*, vol. 9, pp. 522–531, Elsevier B.V., 2012.

[112] J. M. Wing, "Computational thinking and thinking about computing," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 366, no. 1881, pp. 3717–3725, 2008.

[113] A. Ater-Kranov, R. Bryant, G. Orr, S. Wallace, and M. Zhang, "Developing a community definition and teaching modules for computational thinking: accomplishments and challenges," in *Proceedings of the 2010 ACM conference on Information technology education*, pp. 143–148, 2010.

[114] M. Berland and V. R. Lee, "Collaborative strategic board games as a site for distributed computational thinking," *International Journal of Game-Based Learning (IJGBL)*, vol. 1, no. 2, pp. 65–81, 2011.

[115] S. Bjork and J. Holopainen, *Patterns in Game Design (Game Development Series)*. USA: Charles River Media, Inc., 2004.

[116] C. Alexander, *A pattern language: towns, buildings, construction*. Oxford university press, 1977.

[117] E. Gamma, R. Helm, R. Johnson, J. Vlissides, and D. Patterns, "Elements of reusable object-oriented software," *Design Patterns*, 1995.

[118] B. Softworks, "The elder scrolls." [PC, Consoles], 1994.

[119] P. D. Studio, "Stellaris." [PC, Consoles], 2016.

[120] F. Games, "Civilization v." [PC, Consoles], 2010.

[121] M. Games and E. F. Studio, "Thea 2: The shattering." [PC], 2018.

[122] JoyBits, "Doodle god." [PC, Consoles], 2010.

[123] S. Enix, "Final fantasy x." [PlayStation 2], 2001.

[124] L. Haaranen and R. Duran, "Computer science in online gaming communities," in *Computers Supported Education: 9th International Conference, CSEDU 2017, Porto, Portugal, April 21-23, 2017, Revised Selected Papers 9*, pp. 279–299, Springer, 2018.

[125] K. Ferguson, "Everything is a remix." `https://vimeo.com/14912890`, 2011.

[126] A. Lenhart and M. Madden, *Teen content creators and consumers*, vol. 2. Pew Internet & American Life Project Washington, DC, 2005.

[127] Q.-Y. Yin, J. Yang, K.-Q. Huang, M.-J. Zhao, W.-C. Ni, B. Liang, Y. Huang, S. Wu, and L. Wang, "Ai in human-computer gaming: Techniques, challenges and opportunities," *Machine intelligence research*, vol. 20, no. 3, pp. 299–317, 2023.

[128] K. Arulkumaran, A. Cully, and J. Togelius, "Alphastar: An evolutionary computation perspective," in *Proceedings of the genetic and evolutionary computation conference companion*, pp. 314–315, 2019.

[129] C. R. Madan, "Considerations for comparing video game ai agents with humans," *Challenges*, vol. 11, p. 18, 8 2020.

[130] S. Davern and M. Haahr, "On the interactions between narrative puzzles and navigation aids in open world games," in *International Conference on Interactive Digital Storytelling*, pp. 259–275, Springer, 2023.

[131] E. Tomai, "Extraction of interaction events for learning reasonable behavior in an open-world survival game," in *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[132] W. Min, B. W. Mott, J. P. Rowe, B. Liu, and J. C. Lester, "Player goal recognition in open-world digital games with long short-term memory networks.," in *IJCAI*, pp. 2590–2596, 2016.

[133] A. Gupta, D. Carpenter, W. Min, J. Rowe, R. Azevedo, and J. Lester, "Enhancing multimodal goal recognition in open-world games with natural language player reflections," in *Proceedings of the aaai conference on artificial intelligence and interactive digital entertainment*, vol. 18, pp. 37–44, 2022.

[134] Z. Wang, S. Cai, G. Chen, A. Liu, X. Ma, and Y. Liang, "Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents," *arXiv preprint arXiv:2302.01560*, 2023.

[135] I. Borovikov and A. Beirami, "From demonstrations and knowledge engineering to a dnn agent in a modern open-world video game.," in *AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering*, 2019.

[136] M. Müller-Brockhausen, G. Barbero, and M. Preuss, "Chatter generation through language models," in *2023 IEEE Conference on Games (CoG)*, pp. 1–6, IEEE, 2023.

[137] A. Nantes, R. Brown, and F. Maire, "A framework for the semi-automatic testing of video games," in *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 4, pp. 197–202, 2008.

[138] A. Albaghajati and M. Ahmed, "Video game automated testing approaches: An assessment framework," *IEEE Transactions on Games*, vol. 15, pp. 81–94, 3 2023.

[139] M. Aung, S. Demediuk, Y. Sun, Y. Tu, Y. Ang, S. Nekkanti, S. Raghav, D. Klabjan, R. Sifa, and A. Drachen, "The trails of just cause 2: spatio-temporal player profiling in open-world games," in *Proceedings of the 14th International Conference on the Foundations of Digital Games*, pp. 1–11, 2019.

[140] W. Min, B. W. Mott, J. P. Rowe, B. Liu, and J. C. Lester, "Player goal recognition in open-world digital games with long short-term memory networks.," in *IJCAI*, pp. 2590–2596, 2016.

[141] M. A. Gómez-Maureira, I. Kniestedt, G. Barbero, H. Yu, and M. Preuss, "An explorer's journal for machines: Exploring the case of cyberpunk 2077," *Journal of Gaming and Virtual Worlds*, vol. 14, pp. 111–135, 4 2022.

[142] Obsidian Entertainment, "The outer worlds," 2019.

[143] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen, L. Wang, A. T. Luu, W. Bi, F. Shi, and S. Shi, "Siren's song in the AI ocean: A survey on hallucination in large language models," *CoRR*, vol. abs/2309.01219, 2023.

[144] S. Bjork and J. Holopainen, *Patterns in game design*, vol. 11. Charles River Media Hingham, 2005.

[145] M. Lafond, "The complexity of speedrunning video games," in *9th International Conference on Fun with Algorithms (FUN 2018)*, Schloss-Dagstuhl-Leibniz Zentrum für Informatik, 2018.

[146] X. Cui and H. Shi, "A*-based pathfinding in modern computer games," *International Journal of Computer Science and Network Security*, vol. 11, no. 1, pp. 125–130, 2011.

[147] E. Alonso, M. Peter, D. Goumard, and J. Romoff, "Deep reinforcement learning for navigation in aaa video games," *arXiv preprint arXiv:2011.04764*, 2020.

[148] E. Tomai, R. Salazar, and R. Flores, "Mimicking humanlike movement in open world games with path-relative recursive splines," in *Proceedings of the aaai conference on artificial intelligence and interactive digital entertainment*, vol. 9, pp. 93–99, 2013.

[149] L. Zheng, J. Chen, J. Wang, J. He, Y. Hu, Y. Chen, C. Fan, Y. Gao, and C. Zhang, "Episodic multi-agent reinforcement learning with curiosity-driven exploration," *Advances in Neural Information Processing Systems*, vol. 34, pp. 3757–3769, 2021.

[150] X. Chen, T. Shi, Q. Zhao, Y. Sun, Y. Gao, and X. Wang, "Wild-scav: Benchmarking fps gaming ai on unity3d-based environments," *arXiv preprint arXiv:2210.09026*, 2022.

[151] A. Hintze, R. S. Olson, and J. Lehman, "Orthogonally evolved ai to improve difficulty adjustment in video games," in *Applications of Evolutionary Computation: 19th European Conference, EvoApplications 2016, Porto, Portugal, March 30–April 1, 2016, Proceedings, Part I 19*, pp. 525–540, Springer, 2016.

[152] P. Spronck, I. Sprinkhuizen-Kuyper, and E. Postma, "Difficulty scaling of game ai," in *Proceedings of the 5th International Conference on Intelligent Games and Simulation (GAME-on 2004)*, pp. 33–37, 2004.

[153] R. Hunicke, "The case for dynamic difficulty adjustment in games," in *Proceedings of the 2005 ACM SIGCHI International Conference on Advances in computer entertainment technology*, pp. 429–433, 2005.

[154] Paradox Development Studio, "Stellaris," 2016.

[155] Paradox Development Studio, "Hearts of iron iv," 2016.

[156] M. Buro, "Real-time strategy games: A new ai research challenge," in *IJCAI*, vol. 2003, pp. 1534–1535, 2003.

[157] Larian Studio, "Divinity: Original sin ii," 2016.

[158] D. Livingstone, "Turing's test and believable ai in games," *Computers in Entertainment (CIE)*, vol. 4, no. 1, pp. 6–es, 2006.

[159] J. R. Searle, "Minds, brains, and programs," in *Machine Intelligence*, pp. 64–88, Routledge, 2012.

[160] N. R. Jennings, "Coordination techniques for distributed artificial intelligence," *Foundations of DAI*, 1996.

[161] Y. Nakajima, "Task-driven autonomous agent utilizing gpt-4, pinecone, and langchain for diverse applications." `https://yoheinakajima.com/task-driven-autonomous-agent-utilizing-gpt-4-pinecone-and-langchain-for-diverse-applications/`. Accessed: 2024-05-01.

[162] L. M. Csepregi, "The effect of context-aware llm-based npc conversations on player engagement in role-playing video games," *Unpublished manuscript*, 2021.

[163] N. Akoury, Q. Yang, and M. Iyyer, "A framework for exploring player perceptions of llm-generated dialogue in commercial video games," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2295–2311, 2023.

[164] Y. Liu, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, vol. 364, 2019.

[165] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "Goemotions: A dataset of fine-grained emotions," *arXiv preprint arXiv:2005.00547*, 2020.

[166] D. Yang, E. Kleinman, and C. Harteveld, "Gpt for games: A scoping review (2020-2023)," *arXiv preprint arXiv:2404.17794*, 2024.

[167] R. Gallotta, G. Todd, M. Zammit, S. Earle, A. Liapis, J. Togelius, and G. N. Yannakakis, "Large language models and games: A survey and roadmap," *arXiv preprint arXiv:2402.18659*, 2024.

[168] M. U. Nasir and J. Togelius, "Practical pcg through large language models," in *2023 IEEE Conference on Games (CoG)*, pp. 1–4, IEEE, 2023.

[169] J. Roberts, A. Banburski-Fahey, and J. Lanier, "Surreal vr pong: Llm approach to game design," in *36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, vol. 1, 2022.

[170] S. Hu, T. Huang, F. Ilhan, S. Tekin, G. Liu, R. Kompella, and L. Liu, "A survey on large language model-based game agents," *arXiv preprint arXiv:2404.02039*, 2024.

[171] H. Shakeri, C. Neustaedter, and S. DiPaola, "Saga: Collaborative storytelling with gpt-3," in *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*, pp. 163–166, 2021.

[172] M. Elgarf, S. Zojaji, G. Skantze, and C. Peters, "Creativebot: a creative storyteller robot to stimulate creativity in children," in *Proceedings of the 2022 International Conference on Multimodal Interaction*, pp. 540–548, 2022.

[173] A. Zhu, L. Martin, A. Head, and C. Callison-Burch, "Calypso: Llms as dungeon master's assistants," in *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 19, pp. 380–390, 2023.

[174] A. Anable, *Playing with feelings: Video games and affect*. U of Minnesota Press, 2018.

[175] D. K. Langlois, S. Drury, and S. Kriglstein, "Press h to help: The impact of prosocial video games on prosocial behaviors by exposure time," in *Proceedings of the 18th International Conference on the Foundations of Digital Games*, FDG '23, (New York, NY, USA), Association for Computing Machinery, 2023.

[176] M. Viggiato De Almeida, "Leveraging natural language processing techniques to improve manual game testing," 2023.

[177] B. Sandhya, *A Learning Based Emotion Classifier with Semantic Text Processing*, vol. 320, pp. 371–382. Springer International Publishing, 01 2015.

[178] F. Anvari, E. Efendić, J. Olsen, R. C. Arslan, M. Elson, and I. K. Schneider, "Bias in self-reports: An initial elevation phenomenon," *Social Psychological and Personality Science*, vol. 14, no. 6, pp. 727–737, 2023.

[179] C. Zhang, F. Zhang, J. Xie, and S. Wang, "Bert based text emotion classification," 2022.

[180] A. Alhuzali, Y. S. Alginahi, and M. A. Alzahrani, "Robust emotion detection from text using advanced deep learning techniques," 2023.

[181] E. Cambria, S. Poria, A. Gelbukh, and M. Thelwall, "Sentiment analysis is a big suitcase," *IEEE Intelligent Systems*, vol. 32, no. 6, pp. 74–80, 2017.

[182] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba, "Large language models are human-level prompt engineers," *arXiv preprint arXiv:2211.01910*, 2022.

[183] E. Delvaux, N. Vanbeselaere, and B. Mesquita, "Dynamic interplay between norms and experiences of anger and gratitude in groups," *Small Group Research*, vol. 46, no. 3, pp. 300–323, 2015.

[184] P. Konieczny, "Golden Age of Tabletop Gaming: Creation of the Social Capital and Rise of Third Spaces for Tabletop Gaming in the 21st Century," *Polish Sociological Review*, no. 206, pp. 199–215, 2019.

[185] J. Arjoranta, V. Kankainen, and T. Nummenmaa, "Blending in Hybrid Games: Understanding Hybrid Games Through Experience," in *Proceedings of the 13th International Conference on Advances in Computer Entertainment Technology*, ACE '16, (Osaka, Japan), pp. 1–6, Association for Computing Machinery, Nov. 2016.

[186] S. Risi and M. Preuss, "From Chess and Atari to StarCraft and Beyond: How Game AI is Driving the World of AI," *KI - Künstliche Intelligenz*, vol. 34, pp. 7–17, Mar. 2020.

[187] V. Kankainen and J. Paavilainen, "Hybrid Board Game Design Guidelines," in *Proceedings of the 2019 DiGRA International Conference: Game, Play and the Emerging Ludo-Mix*, p. 22, DiGRA, 2019.

[188] C. C. Abt, *Serious Games.* Viking Press, 1970.

[189] Niantic Inc., Nintendo Co. Ltd., The Pokémon Company, "Pokémon GO," 2016. [Android, iOS].

[190] S. Björk and J. Juul, "Zero-Player Games Or: What We Talk about When We Talk about Players," in *Philosophy of Computer Games Conference*, 2012.

[191] J. Schell, *The Art of Game Design: A book of lenses.* CRC Press, Aug. 2008.

[192] Blizzard Entertainment, "StarCraft II: Wings of Liberty," 2010. [Windows, macOS].

[193] M. J. Rogerson, M. Gibbs, and W. Smith, ""I Love All the Bits": The Materiality of Boardgames," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, (San Jose, California, USA), pp. 3956–3969, Association for Computing Machinery, May 2016.

[194] J. R. Wallace, J. Pape, Y.-L. B. Chang, P. J. McClelland, T. N. Graham, S. D. Scott, and M. Hancock, "Exploring automation in digital tabletop board game," in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work Companion*, CSCW '12, (Seattle, Washington, USA), pp. 231–234, Association for Computing Machinery, Feb. 2012.

[195] A. Liapis, G. N. Yannakakis, and J. Togelius, "Sentient sketchbook : computer-assisted game level authoring," in *8th International Conference on the Foundations of Digital Games*, May 2013. Accepted: 2018-04-26T10:13:13Z Publisher: ACM.

[196] S. Deterding, J. Hook, R. Fiebrink, M. Gillies, J. Gow, M. Akten, G. Smith, A. Liapis, and K. Compton, "Mixed-Initiative Creative Interfaces," in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '17, (Denver, Colorado, USA), pp. 628–635, Association for Computing Machinery, May 2017.

[197] F. de Mesentier Silva, S. Lee, J. Togelius, and A. Nealen, "AI-based playtesting of contemporary board games," in *Proceedings of the 12th International Conference on the Foundations of Digital Games*, FDG '17, (Hyannis, Massachusetts), pp. 1–10, Association for Computing Machinery, Aug. 2017.

[198] M. Preuss, T. Pfeiffer, V. Volz, and N. Pflanzl, "Integrated balancing of an RTS game: Case study and toolbox refinement," in *2018 IEEE Conference on Computational Intelligence and Games, CIG 2018, Maastricht, The Netherlands, August 14-17, 2018*, pp. 1–8, IEEE, 2018.

[199] Y. Xu, E. Barba, I. Radu, M. Gandy, and B. MacIntyre, "Chores Are Fun: Understanding Social Play in Board Games for Digital Tabletop Game Design," in *Proceedings of the 2011 DiGRA International Conference*, p. 16, 2011.

[200] Anki, "Anki OVERDRIVE," 2015. [Android, iOS, Physical Platform].

[201] L. van Velthoven, "Room Racers: Design and Evaluation of a Mixed Reality Game Prototype," Master's thesis, Leiden University, 2012.

[202] E. M. Lang, "XCOM: The Board Game," 2015. [Android, iOS, Tabletop Platform].

[203] Berserk Games, "Tabletop Simulator," 2015. [Windows, macOS, Linux].

[204] R. Abbott and E. Rothman, "Disrupting creativity: Copyright law in the age of generative artificial intelligence," *Fla. L. Rev.*, vol. 75, p. 1141, 2023.

[205] D. Baidoo-Anu and L. O. Ansah, "Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning," *Journal of AI*, vol. 7, no. 1, pp. 52–62, 2023.

[206] E. Shein, "The impact of ai on computer science education," 2024.

[207] R. Yilmaz and F. G. K. Yilmaz, "The effect of generative artificial intelligence (ai)-based tool use on students' computational thinking skills, programming self-efficacy and motivation," *Computers and Education: Artificial Intelligence*, vol. 4, 1 2023.

[208] F. Deriba, I. T. Sanusi, O. O Campbell, and S. S. Oyelere, "Computer programming education in the age of generative ai: Insights from empirical research," *SSRN*, 2024.

[209] F. A. Sakib, S. H. Khan, and A. R. Karim, "Extending the frontier of chatgpt: Code generation and debugging," in *2024 International Conference on Electrical, Computer and Energy Technologies (ICECET*, pp. 1–6, IEEE, 2024.

[210] D. Spinellis, "Pair programming with generative ai," *IEEE Software*, vol. 41, pp. 16–18, 5 2024.

[211] B. Idrisov and T. Schlippe, "Program code generation with generative ais," *Algorithms*, vol. 17, 2 2024.

[212] Özgen Korkmaz, R. Çakir, and M. Y. Özden, "A validity and reliability study of the computational thinking scales (cts)," *Computers in Human Behavior*, vol. 72, pp. 558–569, 7 2017.

[213] V. Ramalingam and S. Wiedenbeck, "Development and validation of scores on a computer programming self-efficacy scale and group analyses of novice programmer self-efficacy," *Journal of Educational Computing Research*, vol. 19, pp. 367–381, 1998.

[214] C. Cao, "Scaffolding cs1 courses with a large language model-powered intelligent tutoring system," in *International Conference on Intelligent User Interfaces, Proceedings IUI*, pp. 229–232, Association for Computing Machinery, 3 2023.

[215] T. Phung, V.-A. Pădurean, J. Cambronero, S. Gulwani, T. Kohn, R. Majumdar, A. Singla, and G. Soares, "Generative ai for programming education: Benchmarking chatgpt, gpt-4, and human tutors," in *Proceedings of the 2023 ACM Conference on International Computing Education Research-Volume 2*, pp. 41–42, 2023.

[216] A. Baytak, "The acceptance and diffusion of generative artificial intelligence in education: A literature review," *Current Perspectives in Educational Research*, vol. 6, no. 1, pp. 7–18, 2023.

[217] K. Kanont, P. Pingmuang, T. Simasathien, S. Wisnuwong, B. Wiwatsiripong, K. Poonpirome, N. Songkram, and J. Khlaisang, "Generative-ai, a learning assistant? factors influencing higher-ed students' technology acceptance," *Electronic Journal of e-Learning*, vol. 22, no. 6, pp. 18–33, 2024.

[218] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS Quarterly: Management Information Systems*, vol. 13, pp. 319–339, 1989.

[219] D. Melhart, A. Azadvar, A. Canossa, A. Liapis, and G. N. Yannakakis, "Your gameplay says it all: Modelling motivation in tom clancy's the division," in *2019 ieee conference on games (cog)*, pp. 1–8, IEEE, 2019.

[220] D. Yang, E. Kleinman, and C. Harteveld, "Gpt for games: An updated scoping review (2020-2024)," *arXiv preprint arXiv:2411.00308*, 2024.

[221] Y. Wang, S. Guo, L. Ling, and C. W. Tan, "Nemobot: Crafting strategic gaming llm agents for k-12 ai education," in *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pp. 393–397, 2024.

[222] S. E. Huber, K. Kiili, S. Nebel, R. M. Ryan, M. Sailer, and M. Ninaus, "Leveraging the potential of large language models in education through playful and game-based learning," *Educational Psychology Review*, vol. 36, 3 2024.

[223] A. Isaza-Giraldo, P. Bala, P. F. Campos, and L. Pereira, "Prompt-gaming: A pilot study on llm-evaluating agent in a meaningful energy game," in *Conference on Human Factors in Computing Systems - Proceedings*, Association for Computing Machinery, 5 2024.

[224] I. Steenstra, P. Murali, R. B. Perkins, N. Joseph, M. K. Paasche-Orlow, and T. Bickmore, "Engaging and entertaining adolescents in health education using llm-generated fantasy narrative games and virtual agents," in *Conference on Human Factors in Computing Systems - Proceedings*, Association for Computing Machinery, 5 2024.

[225] Y. Shao, L. Li, J. Dai, and X. Qiu, "Character-llm: A trainable agent for role-playing," *arXiv preprint arXiv:2310.10158*, 2023.

[226] A. Kong, S. Zhao, H. Chen, Q. Li, Y. Qin, R. Sun, X. Zhou, E. Wang, and X. Dong, "Better zero-shot reasoning with role-play prompting," *arXiv preprint arXiv:2308.07702*, 2023.

[227] W. Zhang, J. Sheng, S. Nie, Z. Zhang, X. Zhang, Y. He, and T. Liu, "Revealing the challenge of detecting character knowledge errors in llm role-playing," *arXiv preprint arXiv:2409.11726*, 2024.

[228] Y.-M. Tseng, Y.-C. Huang, T.-Y. Hsiao, Y.-C. Hsu, J.-Y. Foo, C.-W. Huang, and Y.-N. Chen, "Two tales of persona in llms: A survey of role-playing and personalization," *arXiv preprint arXiv:2406.01171*, 2024.

[229] Y.-S. Chuang, K. Nirunwiroj, Z. Studdiford, A. Goyal, V. V. Frigo, S. Yang, D. Shah, J. Hu, and T. T. Rogers, "Beyond demographics: aligning role-playing llm-based agents using human belief networks," *arXiv preprint arXiv:2406.17232*, 2024.

[230] S. Höhn, J. Nasir, D. C. Tozadore, A. Paikan, P. Ziafati, and E. André, "Beyond pretend-reality dualism: Frame analysis of llm-powered role play with social agents," in *Proceedings of the 12th International Conference on Human-Agent Interaction*, pp. 393–395, 2024.

[231] H. Elsayed, "The impact of hallucinated information in large language models on student learning outcomes: A critical examination of misinformation risks in ai-assisted education," *Northern Reviews on Algorithmic Research, Theoretical Computation, and Complexity*, vol. 9, no. 8, pp. 11–23, 2024.

[232] Infocom, "Zork." [PC, Consoles], 1977.

[233] H. Heyen, A. Widdicombe, N. Y. Siegel, M. Perez-Ortiz, and P. Treleaven, "The effect of model size on llm post-hoc explainability via lime," *arXiv preprint arXiv:2405.05348*, 2024.

[234] A. Liu, B. Feng, B. Wang, B. Wang, B. Liu, C. Zhao, C. Dengr, C. Ruan, D. Dai, D. Guo, *et al.*, "Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model," *arXiv preprint arXiv:2405.04434*, 2024.

Summary

Video games have emerged as one of the most interesting media of the turn of the century. The peculiar characteristic of video games is the biunivocal interaction between the computer and the human player. This dissertation starts by exploring the historical development of games and video games as media for serious purposes. Then, it dives into the traditional intersection of video games and education, with a specific focus on computer science and the development of computational thinking skills. We analyse current research in the field, reporting on its limits and trying to draw conclusions on the impact of video games in scientific education. Moreover, we highlight the affordances between video game play and computational thinking skills. We see that computer science education can greatly benefit from game elements.

The biunivocal interaction fostered by video games becomes particularly fascinating with the rise of artificial intelligence. We illustrate how video games have defined the development of intelligent systems since the beginning. Moreover, generative artificial intelligence creates new opportunities for video games as complex interactive environments for humans and AI. We explore some of these opportunities and how humans can be influenced by artificial intelligence in video games.

Finally, we bring together all the elements we introduced, analysing the impact of generative AI on education and singling out its limits and dangers. Although unrestricted AI is particularly problematic for students' learning, video games can moderate it, creating complex interactive learning environments. While technology is not necessarily developed enough to allow for the mass production and use of AI-powered educational video games, we highlight how the rapid progress we are currently experiencing can soon make this a reality.

The thesis analyses the impact and opportunities for game research in the era of human-AI interaction studies. Specifically, looking at the potential for a better and positive impact of AI in education.

Samenvatting

Videogames zijn rond de eeuwwisseling ontpopt tot één van de meest interessante media. Het bijzondere kenmerk van videogames is de biunivocale interactie tussen de computer en de menselijke speler. Dit proefschrift begint met een verkenning van de historische ontwikkeling van games en videogames als media voor serieuze doeleinden. Vervolgens gaan we in op het traditionele snijvlak van videogames en onderwijs, met een specifieke focus op informatica en de ontwikkeling van computational thinking skills. We analyseren het huidige onderzoek op dit gebied, brengen verslag uit over de beperkingen ervan en proberen conclusies te trekken over de impact van videogames in het wetenschappelijk onderwijs. Daarnaast belichten we de mogelijkheden tussen het spelen van videogames en computational thinking skills. We zien dat het informaticaonderwijs veel baat kan hebben bij game-elementen.

De biunivocale interactie die wordt bevorderd door videogames wordt nog intrigerender met de opkomst van kunstmatige intelligentie. We illustreren hoe videogames de ontwikkeling van intelligente systemen vanaf het begin hebben bepaald. Bovendien creëert generatieve kunstmatige intelligentie nieuwe mogelijkheden voor videogames als complexe interactieve omgevingen voor mens en AI. We verkennen enkele van deze mogelijkheden en hoe mensen kunnen worden beïnvloed door kunstmatige intelligentie in videogames.

Tot slot brengen we alle elementen die we hebben geïntroduceerd samen, analyseren we de impact van generatieve AI op het onderwijs en definiëren de grenzen en gevaren ervan. Waar onbeperkte AI bijzonder problematisch is voor het leerproces van studenten, kunnen videogames dit remediëren door complexe interactieve leeromgevingen te creëren. Hoewel de technologie nog niet voldoende ontwikkeld is om de massaproductie en het gebruik van AI-aangedreven educatieve videogames mogelijk te maken, benadrukken we hoe de snelle vooruitgang die we momenteel meemaken dit binnenkort reëel kan maken.

Samenvatting

Deze dissertatie analyseert de impact en mogelijkheden voor game-onderzoek in het tijdperk van de mens-AI-interactie studies. Specifiek kijken we naar het potentieel voor een betere en positieve impact van AI in het onderwijs.

Acknowledgements

Those who know me are aware of how uncomfortable I am at showing emotions (besides annoyance, frustration, boredom, and general angst). However, this PhD is not a product of my work only, but of many, many people who, throughout the years, supported me both professionally and personally. Therefore, here we go, time to show genuine gratitude in the best way I can.

First, I want to thank my family. Thanks to my parents, who supported and followed every single one of my weird fixations and decisions (not without concerns). They not only encouraged me but also recommended that I follow my path regardless of where it would lead me. Thanks to my brother, who showed me (and my whole family?) that sometimes remaining calm is the best course of action. Finally, I want to thank the rest of my family, also those who are not with us anymore; thanks to my grandmother who taught me not to care for other people's opinions, and to my grandfather who taught me the value of knowledge.

Next, I want to thank my husband, Daniel, who tolerated me (starting to see a pattern here) while challenging me and showing me the importance of standing up for myself and other people.

I want to express my gratitude to my best friends Adriano and Giuseppe for being there at every step of this journey, through adventures and misadventures, and Jonne, who has been my partner in crime since my bachelor's.

Coming at the end of this PhD, I cannot forget about my colleagues. Thank you, Matthias, for being the most reliable, curious, and kind office mate; through all these years, you have been not only an amazing colleague but also an incredible friend. I hope we will continue to have fun together for many years. Thanks also to Tom, who has been there to challenge me and be challenged every single day (and I mean every single day). I find it so fortunate that I was able to start and finish this journey together with you both. Thanks also to all the colleagues who made these years lighter:

Nathan, Roberta, Arina, Lennard, and many others.
I want to thank my supervision team. Thanks to Marcello, who throughout these years continued to trust me and to support me both in research and education. Thanks also to Felienne, who not only entrusted me with this job, she also taught me the importance of education in academic research. Finally, I want to thank Mike, my daily supervisor, who has been a mentor, a role model, and, most importantly, a friend.
If I did not mention your name on this page, know that I did not forget about you. Thank you to all the people in Italy and the Netherlands who talked, encouraged, and challenged me. Now that I have finally finished this page, I can go back to designing the cover page, a job at which I am much more comfortable.

Curriculum Vitae

Giulio Barbero was born on the 23rd of December 1992 in Acqui Terme, Italy. He moved to the Netherlands in 2014 where he completed his BSc in Industrial Design Engineering at the Haagse Hogeschool in 2017. He then obtained his MSc in Mediatechnology at Leiden University in 2019. In the fall of the same year, he continued working for Leiden University as a technical education assistant for the courses *Advances in Datamining*, *Data Structures*, and *Concepts of Programming Languages*. He started his PhD in 2020. As part of his PhD trajectory, he taught courses including *Introduction to Programming*, and *Video Games for Research*. At the same time, he followed a variety of courses in order to obtain his University Teaching Qualification (UTQ), including Testing and Assessment and Teaching in Practice. His research interest lies in the use of video games for education both as a medium and as a topic of study.