UNIVERSITY OF CALIFORNIA
SANTA CRUZ

NPCs to Believe In: Value-based Morality in Video Game Characters

A dissertation submitted in partial satisfaction
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

In

COMPUTATIONAL MEDIA

By

Rehaf A. AlJammaz

December 2024

The Dissertation of Rehaf AlJammaz is approved:

_____
Professor Noah Wardrip-Fruin

_____
Professor Micheal Mateas

_____
Professor Mike Treanor

_____
Professor Elin Carstensdottir

_____

Peter Biehl

Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures, Tables and Illustrations

Abstract

NPCs to Believe In: Value-based Morality in Video Game Characters

Rehaf Al Jammaz

In our dissertation, we investigated three areas in video game characters: believability, a character's sense of beliefs, and designs for moral reasoning. Character believability is an often misunderstood and ambiguous area of research. Literature usually uses “realism” and “believability” synonymously to describe believable characters, whereas the terms include inherent differences in methodology, agent structures, designs, and evaluation techniques. Thus, one agenda item is to clarify and define what constitutes a believable character.

Our research dives deeper into notions of believability. A set of believability criteria is established from the literature to include a character's personality, emotion, context, roles, change, and sociability, among others. By examining the field, we noticed a gap exists in developing a character's sense of change. Furthermore, a gap exists in how agent architectures utilize values and beliefs, especially as they tie to a character's moral judgments.

While morality as a field is extensive, our dissertation limits morality to believable video game characters, focusing on standard methodologies used in constructing moral systems, including linear scales and state machines. In later chapters, we reexamine morality through a faction lens and present a taxonomy of faction characters, including how morality is perceived and conveyed to the player.

Through our examination of morality in games, we argue that standard reputation scales shift moral systems into a somewhat binary state of good or evil with little shades of gray, whereby morality often includes many diverse notions and personal beliefs. We believe that a character's values and beliefs provide a promising avenue for constructing a multifaceted moral system, especially for background characters.

Thus, we introduce our system and subsequent studies of Argument Box (AB). AB is a system modeled after Lakoff's *Moral Politics*, which highlights morality as a collection of metaphorical values. In AB, characters often judge others based on surface-level calls and multiple deep-rooted values, held at varying levels.

Our primary goal in AB lies in exploring and understanding the implications of value-based morality and its relationship to notions of believability. Through our studies in AB, we discovered that players can perceive values (in many ways), and that a relationship exists between an NPC's perceived values and believability. Specifically, a strong connection exists between a character's set of values and their perceived personality. We also discovered that a character's values influenced the character's perceived motivation, moral beliefs, and propensity to change.

Acknowledgements

بِسْمِ ٱللَّهِ و ٱلْحَمْدُ لِلَّه

I would like to thank the following people who have been instrumental in my PhD journey. Thank you for all of your love and support. We did it!

To my advisors, Michael Mateas and Noah Wardrip-fruin, for always providing a guiding hand and for your patience, kindness, and support. You have helped me grow both academically and personally, and I will forever cherish these lessons. Thank you for being wonderful mentors and amazing human beings. Thank you for all of this and for much more beyond what I can name.

To my dad, Abdulrahman, for your unconventional shenanigans and everlasting encouragement and support, thank you, dad! Thank you to my mom for always being a fantastic role model, guiding me, and supporting me all these years. To both of my parents, thank you for your guidance, love, and encouragement; I would never be here without you.

I am especially grateful and thankful to my amazing friend Yasheng She for his support, kindness, and patience. Thank you for being who you are and for being a wonderful person. My PhD journey would not have been the same without you! I will do better and be better!

To my amazing friends, Aljoharah Alfayez, Aljohara Almoammar, Yasmin Badrudin, and Athoug Alsoughayer, I love you; thank you for always being there.

To the first group of friends I made in the States, thank you for making the transition easy and providing me with a home away from home. Thank you for being amazing people. Thank you to Cherise Datu, Brittany Williams, Aubrey Hill, Caitlin Friess, Morgan Epstein, Daniel Perricca, Kyle Mitchell, Kirby Cofino, and Ffion Goss-Alexander.

To my friends, Devi Acharya, Mirek Stolee, and Elisabeth Oliver thank you for all your support and love.

To my lab mates and friends, thank you for our wonderful discussions, games, projects, and experiences. Thank you, Shi Johnson-Bey, Alex Calderwood, Kyle Gonzalez, Hongwei Zhou, Maxwell Joslyn, Jack Kelly, Jñani Crawford, Max Kreminski, Dylan Lederle-Ensign, Kevin Weatherwax, James Fey, Allen Riley, Batu Aytemiz and Issac Karth.

Thank you to my mentors and professors who helped guide me through this long journey: Chris Totten, Lindsay Grace, Jim Whitehead, and Adam M Smith. Thank you, especially to Josh McCoy and Mike Treanor, for your support and encouragement. You inspired me to follow in your footsteps.

I would also like to thank SACM, Saudi Arabia's cultural mission their your support.

To my siblings, Shadin, Yasir, Shahad, Tarig, Mohammad, Sultana, Rasha and Basel, thank you for your love and support. Thank you Basel for all of your advice! To my nieces and nephews, thank you for being you. Thank you to Albandry Alrobaiaan for being an amazing person; thank you, Dudu. My thanks to my aunt and cousins,

special thanks to Sara Aljammaz, Marwan AlBawardi, and Sadeem AlJammaz.for their continued support. A special gratitude goes to my best friend, closest confidant and sister, Rasha, for always being there and for being you. Thank you, sis! Lastly, thank you to the smallest members of the family for pestering your parents, showing love, and being amazing nieces and nephews; thank you, Nono, Khalodie, and Nayofey.

Thank you, Luisa Squires, for all the advice and encouragement.

Last but not least, thank you to my committee, Noah Wardrip-Fruin, Michael Mateas, Mike Treanor, and Elin Carstensdottir, for your guidance and support.

Introduction

> "... character is the most interesting phenomenon anywhere. Every character represents a world of his own, and the more you know of this person, the more interested you become." _EGRI

Characters exist in all forms of media; we can see them in books, plays, movies, interactive experiences, and video games. Characters have the strength to transfer us into unimaginable worlds, make us feel joy or hatred, drive us, and inspire us in our daily lives. Characters are so paramount of a concept that stories are conveyed through them. When asked, Lajos Egri, an expert in character designs and the author of *The Art of Dramatic Writing,* writes, “The interface is unmistakable: character creates plot, not vice versa.” He then highlights the importance of a character by mentioning this example:

> “Which is more important, plot or character? Let us trade the sensitive brooding hamlet for a pleasure loving prince, whose one reason for living is the privileges his princehood affords him. Would he avenge his father’s death? Hardly. He would turn the tragedy into a comedy.” _EGRI

Characters, without a doubt, hold the power to mold stories and alter the narrative. This is particularly evident in interactive experiences, such as games, where the world is in a constant state of flux.

Defining what constitutes a character can result in a multitude of elements taken from many different fields and perspectives. At a bare minimum, a character's basic structure [47, 30] includes three dimensions: a character's physiology, sociology, and psychology. These elements combine to contextualize and create characters in popular literature and media. However, contemplating what makes a "character" is different from what makes a "good character." In other words, what makes a character suspend us from our world into theirs, what gives us that sense of suspended disbelief, what makes them *believable*?

This dissertation will focus on believable characters within the context of video games and interactive experiences. Video games differ in how they present and portray characters compared to other media forms. For one, video games create worlds that are often explorable, giving users a larger space for interaction. Meanwhile, extending the interactive space creates a more burdensome task for an author or creator. Characters in interactive spaces create additional complexities such as communicating the world and character affordances to the player in a reasonable way.

With that, we arrive at our first research question and subquestions:

- **RQ1: What is believability for video game non-player characters (NPCs)?**
    - What frameworks exist for analyzing and understanding character believability?
    - What opportunities present themselves for extending character believability?

As we will explore in **Chapter One**, *character believability* is an ambiguous term at best. Multiple theories, systems, and artifacts in both academia and industry use the term believability to describe different elements of a character. Is a character believable because it can mimic a human being almost perfectly, or is it believable because it can suspend disbelief?

To understand what makes a character believable, we must first clarify our notions of character believability. In **Chapter One**, we will differentiate between different characters, categorizing them into two broad categories: that of *realistic characters* and that of *believable characters*. Our focus in this dissertation is the latter, the *believable character*.

In general terms, a *believable character* is a character that extends beyond reality, mimicking life and pulling the viewer into their world while abiding by a combination of believability criteria. The list of believability criteria [3] includes personality, emotion, agency, growth, and sociability, among others.

As we will explore in future chapters, one area of character believability pertains to a character's ability to portray autonomy and life-like qualities. These qualities include a character's sense of self and capacity to act according to their own volition. With that, we reviewed autonomous and character-based systems with the criterion that they use believability-based characteristics in their design.

Upon surveying believability-based systems, a few elements were apparent. First, *change*, an aspect of character believability involving the idea that a character gradually grows and develops, is an often underdeveloped and understudied concept.

Second, we discovered that a common implementation of autonomous and believable systems often involves values and beliefs. However, values are often taken in the form of knowledge facts (such as a character believing that a book's color is red) or relationship beliefs (e.g., character A believes character B likes them).

We believe we can utilize values and beliefs to invoke change. One area where change can have an effect is that of character morality.. We believe there is an opportunity for enhancing moral characters through their values and beliefs. Our upcoming system description (AB) highlights this in the form of opinion change.

But before that, we have to understand how moral values and depictions are operationalized in video games. To that end, **Chapter Three** explores common moral designs, highlighting common issues with standard reputation systems such as reputation scales and state machines. These standard morality systems seem to result in a binary state of right and wrong, with little in terms of shades of grey. Characters' values are often expressed through rich authoring, conveying single-minded NPC perspectives or shallow reactions for (most) background NPCs, with a few exceptions. Through values, we can create varied NPC moral designs across multiple characters, whereby a character's morality is represented via their internal beliefs. What's more, we consider how an NPC can change their views of others through opinion change.

Thus, this dissertation aims to investigate *systemic* approaches to value-based moral design. We believe that we can create systemic change that affects player perceptions of differing NPCs through character values. As part of our research, we designed our system Argument Box, which incorporates moral values at two levels: (1)

a surface level in the form of topic beliefs and (1) deep values, which focus on arguments that utilize up to 18 different values.

Argument Box's moral values are based on cognitive linguist George Lakoff's book, *Moral Politics [77]*, in which he differentiates between conservative and liberal rhetoric through two family system metaphors: the strict father and the nurturant parent family systems. Each of these metaphors includes a subset of moral metaphors (values) that each system finds moral or immoral. We used Lakoff's work as a basis for constructing our NPC's moral values and opinions. But how do values affect our perception of morality? Are these communicated—and do they aid in character believably?

With that, we arrive at our second research question (followed by three subquestions):

- **RQ2: In what ways can a systemic approach to value-based design affect perceptions of character morality and believability?**

In order to understand the relationship between value-based design and that of morality we need to first clarify what we mean by morality within the context of this dissertation. Thus we pose the following set of subquestions:

  - In what ways are morality and values portrayed and operationalized for video game NPCs?
  - What is the relationship between value-based design and perceived morality?
  - What is the relationship between value-based design and believability?

Argument Box (AB) examines the effects of value-based design in a close-up manner through NPC conversations through which the player attempts to dissuade NPCs by appealing to their deeply held values. By playing to an NPC's deep moral values, players can change the NPC's opinion about other NPCs living in their world, changing their moral perspective. However, as AB does not represent a conventional game genre, one may wonder how we can apply a systemic approach to value-based moral design across multiple NPCs and abide by the rules of conventional genres.

One conventional character system applicable to many game genres that inherently includes values in its designs are factions! Factions are included in many genres, such as RPGs, MMORPGs, RTS, and FPS games. Factions represent a group of characters following some ideal or shared cause; they are usually formed as a result of conflict between opposing groups based on a violation of an ideal or belief.

With this in mind, we arrive at our last research question and subquestions:

- **Q3: How can we utilize a systemic approach to morality in existing game genres?**

As this dissertation argues, factions present an ideal system for highlighting change as an exciting game genre, but faction designs are numerous. In an effort to understand how a systemic approach to morality can be applied, we examine faction further, by asking these subquestions:

    - What roles do factions play in games?
    - How can systemic approaches highlight growth and change through a large faction system?

- In what ways can we develop character growth in a large-scale faction system?

Dissertation Outline and Contributions:

Unfortunately, the term "believability," when referring to game characters, is used in two incompatible ways: realism (mimicking realistic humans) and character believability, as used in the character arts. In **Chapter One,** we aim to clarify the areas of realism and believability, covering their goals, research areas, and the main differences between the two camps.

Chapter Two dives deeper into character believability, covering what it means to be a believable character. This chapter provides background research, theoretical frameworks, analysis, system reviews, and case studies on believability.

Chapter Two also establishes the gap in the literature, highlighting the fact that change, an aspect of character believability, is an often understudied topic; this chapter also introduces values as a common design element of agent architectures. We argue that value-based design can be utilized as an alternate method for modeling morality and moral reasoning, presented in our systems as opinion change.

To understand how morality is commonly implemented, **Chapter Three** provides background information on morality, particularly as it relates to video games. This chapter covers morality in industry and academic systems by highlighting examples and case studies. From our review, most moral systems present character values as non-changing NPC perspectives (with some exceptions in academic systems, as we will soon discover). We also argue that common moral methodologies (state

machines or reputation scales) eventually make characters feel binary, whereby their moral stance is usually perceived as good or evil.

Through a set of moral values, we can make characters appear morally varied, with characters caring about different agendas based on their individualized beliefs. Thus, in this chapter, we argue that incorporating values as moral arguments at a systemic level can introduce new alternative systems for morality. Furthermore, we can utilize beliefs to invoke character change. Our upcoming system, as we will soon discuss, attempts to do just that, in the form of opinion change, whereby NPCs can change their mind about others based on the player appealing to an NPC's deeply held values.

Chapter Four covers beliefs and values as a topic. Additionally, **Chapter Four** provides a review of belief- and value-focused systems. Through this chapter, we learn more about value-based design and suggest examining cognitive and value-based literature as a basis for systemizing values. We then select and introduce Lakoff's *Moral Politics*, which showcases morality as a collection of metaphorical values used as a basis for our value-based moral project, *Argument Box* (AB).

Chapter Five introduces our first iteration of *AB*, an argument simulator with a value-based system for moral opinions. Following Chapter Five, **Chapter Six** introduces our current version of Argument Box. The current iteration of AB focuses on incorporating values and opinion change at a systemic level. The game is generative, with each NPC at initialization caring about different topics (surface values) based on their current patterns and some RNG. NPCs in this game judge or applaud others based

on other NPCs adhering to or violating their surface beliefs. NPCs use their deeply held values to argue their position; the player's goal, in turn, is to dissuade NPCs of their opinion using what they think the NPC cares about (deeply held values) in their arguments.

Each NPC in AB references a collection of 18 deeply held values modeled on Lakoff's moral metaphors. Each NPC can *gradually change* their opinions about other characters as the player works on convincing them through multiple topics and utilizing their deeply-held values against them.

Chapter Five and Chapter Six cover the system's implementation and overall design. We also discuss the differences between the two systems, covering lessons learned and highlighting authoring and design considerations.

Chapter Seven and Chapter Eight present our qualitative and mixed method studies in AB to understand the implications of value-based moral designs; we review the details of our research, including recruitment, methodologies, data analysis, and results. Before we can understand how a systemic approach affects the player's perception of morality and believability, we must first ask whether players can perceive values in the first place.

Thus, as presented by **Chapter Seven's** study analysis, we discover that players could indeed perceive values at multiple levels, including those defined by Lakoff, where players mentioned conservative and liberal rhetoric (equated to Lakoff's strict father and nurturant parent metaphors) as well as identify specific deep values that the NPC holds, such s valuing strength or community.

After establishing that values are understood by players, **Chapter Eight** highlights the results of our second study; here, we examine values in-depth, covering what players thought of specific values and how they relate to morality and believability.

In terms of believability, our value-based morality system indicated an association between an NPC's perceived personality and values. Some players, for instance, used specific values in their personality descriptions (e.g., a character that followed *authority*), while others referred to the overall moral model, stating descriptors like a character perceived as being *old-fashioned.* Values also played a role in the player's perception of other believability-based criteria, such as change and motivation.

While our system implements *change*, it communicates it at a superficial level. Because of our system and design limitations (e.g, lack of animation, additional authoring), *change* is simply displayed in the play through intro and outro statements. NPCs reflect their feelings about an NPC on five levels, but the arguments used in most of the game remain unchanged, as they are tied to the values themselves. That said, we noted that change can be seen in terms of character persuadability or in terms of the player's perception. Character persuadability is often seen as a difficult task when conversing with a character with opposing player beliefs (SF), while an NP character, in most cases, is seen as easily persuaded. Some players perceived the character growing in terms of personality. In contrast, others might perceive a lack of change as a personality trait (e.g., seeing a character as stubborn and refusing to change).

In terms of morality, players could identify and correctly map the NPC's deep values to their overarching moral models (SF and NP models). Some players also perceived morality as multi-varied, explicitly authored, or following a branch-like structure. Generally, the majority of our players felt that value-based morality made our characters more engaging and more believable.

We note that both studies in **Chapter Seven and Chapter Eight** present additional elements and lessons learned that emerged as a result of value-based moral design. In addition to the above summary, **Chapter Seven and Chapter Eight** discuss the emergence of storytelling, the effects of value-based design on the remaining believability criteria, value identification, difficulty, and the differences between an NP and SF character, among others.

As described earlier, AB implements a value-based approach at a systematic level; it depicts morality and *change* (via character's opinions) at a zoomed-in level, taking place conversationally. We acknowledge, however, that an argument simulator is not a typical game genre. As such, we envisioned factions as an ideal scenario that can impact morality and growth at a zoomed-out level.

Thus, **Chapter Nine** presents and defines what constitutes a faction. Through our analysis, we discover that factions inherently contain values and ideals. Furthermore, faction members often fight others because of a violation of those ideals and values.

To understand factions better, we created a faction taxonomy that categorizes factions in games according to their base design as implicit or explicit factions. We

then taxonomize factions further into subcategories under each umbrella, extrapolating key features, including the faction's moral range, the player's main choices, possible character interactions, the faction's use of hierarchical structures, the faction goals, the differences between representative and background NPCs, and lastly, the unfolding story structure.

To better understand how change appears, we examined both factions in games and those in other media. Unfortunately, growth and change are underrepresented aspects of character believability within faction games (with the exception of some representational characters that **Chapter Nine** highlights).

Through media analysis, looking beyond games, we noted different ways in which change often appears and develops within faction characters. For instance, we discovered that:

- Change can happen to any character (not just protagonists).
- Change is often gradual.
- Change can be triggered in multiple ways (chance, choice, or force).
- Personality factors can make characters susceptible to change.
- Events and story beats can push characters to change.
- Change is not all-encompassing—some older beliefs and values could remain unchanged.

Lastly, we end our dissertation by introducing our current simulation prototype that focuses on character growth and change in a faction game while considering the above change elements. **Chapter Nine** further illustrates the system in greater detail.

Summary of research questions:

Here we summarize our research questions, with subquestions, as presented earlier in the introduction.

RQ1: What is believability for video game non-player characters (NPCs)?

- What frameworks exist for analyzing and understanding character believability?
- What opportunities present themselves for extending character believability?

RQ2: In what ways can a systemic approach to value-based design affect perceptions of character morality and believability?

- In what ways are morality and values portrayed and operationalized for video game NPCs?
- What is the relationship between value-based design and perceived morality?
- What is the relationship between value-based design and believability?

RQ3: How can we utilize a systemic approach to morality in existing game genres?

- What roles do factions play in games?
- How can systemic approaches highlight growth and change through a large faction system?
- In what ways can we develop character growth in a large-scale faction system?

Chapter One: Believability vs. Realism

Acknowledgement: This chapter has been published elsewhere [6].

1. What Is Character Believability?

Before we dive into character believability, let us examine the broader notion of *character*. Generally, one would assume a character represents a fictional person or entity. If we look into established literature by screenwriters and playwrights such as Egri [47], we find certain elements come together to make a character, including psychological, physiological, and sociological components. Characters are further developed by aspects such as motivational factors and growth.

Unsurprisingly, these definitions and usages are closely related to what research from the arts calls *character believabilit*y. Character believability has its origins in the arts [47,70]. It was further developed by Carnegie Mellon University's (CMU) Oz Group, whose researchers defined a believable character as one that can portray an illusion of life and allow the player to suspend disbelief. This Group [93] developed a list of characteristics borrowed from Loyall's [87] work (e.g., personality and motivation) deemed necessary for creating believable characters. This type of character is an exaggeration of life. We will return to the subject of believable characters in greater detail in **Chapter Two.**

The term believability is also sometimes used in reference to human-like behavior. For instance, Togelius et al. [150] identified character believability as a bot

perceived as an actual human or living being; they also defined player believability as the situation in which players believe that a human controls a character. Other works often use the term believability to gauge a character's realistic traits.

Understanding what believability entails at this point becomes taxing and confusing for researchers and creators. It is not uncommon for different communities to use the term character believability to convey both realism and character believability.

2. The Seemingly Blurry Line Between Believability and Realism

Thus, the term believable character (or believability) is used in two incompatible ways: either as the ability to mimic realistic human-like behavior, or as exaggerated and expressive characters that create the illusion of life [47,70]. We argue that these two areas are too dissimilar to use the same language and terminology. These distinct areas—consisting of different definitions, artifacts, and design and evaluation methodologies—should not be grouped under the same term.

In this section, we aim to clarify the areas of believability and suggest an alternative terminology to *character believability*. This section is not intended as a comprehensive review but rather to showcase the diverse schools of thought under each category. In categorizing and identifying notions of believability , we can foster better communication and grow as a research community. Table 1 summarizes how the term character believability is being used in our research communities.

Research using the term *Character believability*									
General Areas	Realistic characters				Believable characters				
Research using the term *believability*	Areas of affect, visual representations, player modeling, cognitive and psychological models.				Oz philosophy, character categorizations and typologies, believability patterns, life-like and emergent behaviors.				
Research Focus	Character appearances and physiological aspects.	Ability to replicate scenarios/ results.	Realistic/ plausible models.	Psychological, cognitive models of human behavior.	Behavioral modeling.	Audience perception.	Emergent behavior.	Focuses on an area of believability (characteristic)	Story/ narrative plot.
Examples	Virtual Agents, virtual beings work.	Player modeling, Realistic agents in FPS.	Human behavior via statistical modeling.	Areas of a Affect, Emotion modeling, Empathy models, (e.g., OCC)	Reactive planning.	Believability is based on the audience perception of characteristics	As a result of agent interaction, unique behaviors.	Character specific personalities, social simulations.	Narrative systems work, drama managers and their use with agents.

Table 1: The table depicts an overview of areas where character believability is mentioned. Specific examples and analyses are presented in a later section.

2.1 Realism Research Areas

In examining believability, we noticed that the realism camp often compares an agent's behavior to that of a human, e.g., [9, 86 ,82]. The research areas involved focus on characteristics and contributions that differ from the notion of believability in the realm of the arts, as we will soon discuss. We recommend using the term *realistic character* instead of *believable character*, due to the inherent comparison to actual human behaviors embedded in the research. Here we describe the research areas that are relevant to *realistic character* research:

- **Player modeling and competitive autonomous NPC agents research.** Player modeling is a significant area concerned with creating agents that behave similarly to human players. The agents in this area are often composed of computational models that consider behavioral, cognitive, and emotional attributes [169]. There are numerous examples and classifications of player modeling, including agents used in AI competitions, FPS (first-person shooter) agents, and agents used for game design or debugging purposes [9, 82,170].
- **Psychological agent modeling research**. Research in this area usually employs specific psychological models in an agent's design to mimic human-like responses. Research often focuses on the area of affect [49], [138]. Examples include emotional and empathetic modeling, such as the Ortony, Clore, and Collins (OCC) and Pleasure-Arousal-Dominance (PAD) models [138], to elicit a human-like response from an agent.
- **Cognitive agent modeling research.** Similar to psychological modeling, cognitive modeling focuses on implementing an agent's problem-solving capabilities in a human-like fashion. Research in this area often focuses on modeling aspects of human cognition, including motivational facets, memory management, reasoning, and perception capabilities. Examples include implementing the agent's motivational and metacognitive control in CLARION [142,143] and social perception in Casper [60].

- **Research focusing on physiological character aspects.** Research in this area often focuses on the authenticity of human-like behaviors and animations. Examples include realistic expressions and gestures [28] as well as virtual beings research [158], such as that by Miquela Sousa, a robotic digital influencer on Instagram [103], and Mica [99], an AI assistant developed by Magic Leap, with hyper-realistic features.

Generally, work in the areas mentioned above seeks to achieve plausibility and replicability. The successful evaluation of a character truthful to cognitive or psychological models is usually compared to that of human-like responses, statistical data, or a human player (in the case of games and player modeling) [9, 82, 49, 123]. Furthermore, the models are usually grounded in reality, constricted by what the employed theory deems as an appropriate true-to-life response. As we will soon note, this contrasts with believability characteristics, where a believable character is often exaggerated, delivers emergent behavior, and is evaluated based on the audience's perceptions [93].

2.2 Believability Research Areas

Now that we've looked at the areas involved in the realism camp, we will examine the research areas associated with the second camp, which we call *character believability*. Character believability is not realism [93]. A concept taken from the arts [70, 48], believable characters exaggerate life-like qualities and support a "willing suspension of disbelief," emphasizing characteristics such as personality,

emotion, motivation, change, social relationships, and the illusion of life [87], as we will soon discuss in greater detail.

Research areas that are relevant to believable characters as seen from an arts perspective, as follows:

- **Story-centric social simulation research.** Social simulation architectures mainly focus on supporting and defining social aspects between interacting agents. As we will elaborate in detail in a later section, social relationships between characters are crucial for believability, representing one of the key characteristics. Furthermore, social architectures, such as the ones seen in [51, 61, 97], inherently incorporate other areas of character believability, such as motivations and personality, to varying degrees.
- **Autonomous character research.** This domain covers characters that can act and interact with objects and the world around them. Autonomous agent architectures help give characters the illusion of life [87], whereby characters seek their actions according to their own goals and agendas. Contrary to autonomous and FPS agents under realism, these characters involve other believability characteristics in their design. Unsurprisingly, researchers often combine autonomous agent and social simulation architectures to cover character goals, social aspects, and other believability characteristics. These elements are incorporated by many academic systems, such as Prom Week, MKUltra, Talk Of The Town, and Façade [97, 66, 94, 132], among others.

- **Story generation and narrative systems research.** Story generators such as story planners often involve characters that act according to a plan or a generated plot. Planners often set and manage a given character's goals and conditions. They can advance the plot, provide emergent gameplay and contextualize characters within their stories, covering all aspects of believability [164, 163].
- **Character modeling and game research.** Character modeling may use cognitive or emotional theories to enhance a character's believability characteristics. Unlike agents focused on realism, the research here does not focus on whether the emotional/cognitive model is authentic but on whether it can create unique and emergent character behavior. Modeling in this research area is usually combined with other criteria and produced (often) as an interactable experience. Examples include using the OCC model in Cif-CK [61] and Talk of the Town [132].

Work in the areas above usually creates unique emergent experiences that enhance the overall believability of a character. Studies often focus on the resultant experience as the audience interacts with the produced artifact. Unlike agents mentioned in the realism domain, evaluation is based on the audience's perception of the characters involved, if they are contextually appropriate and fit the overall narrative. Examples can be seen in systems such as Versu, PromWeek, and Façade [51, 97, 94]; all of these unique systems take into account the player's subjectivity rather than comparing character behavior to an objective standard of human behavior.

2.3 The Not-so-blurry Line between Believability and Realism

In comparing the two research camps, we noticed key differences in research goals, evaluation, artifacts, questions proposed, and (if applicable) the agent's overall architecture. Here we summarize the differences across research areas by examining the systems, theories, and architectures involved. We looked at research that often uses the term "believability" or focuses on characters and agents in their system's architectures. We note that this is not intended as a comprehensive review but a way for us to differentiate the research areas. Samples of research are mentioned in table 2. Here we review the key differences:

- **The human in the room, research goals.** One prominent distinction between the two camps lies in measuring the agent's relatedness to human behavior. *Realism* asks whether an agent's behavior mirrors how a human would act in a given situation. *Believability*, on the other hand, examines specificity in character and often takes the player's subjective response to the character into account (e.g., [61]). This result is further mirrored by Mateas's conclusions in [93] about the difference between believable agents and AI research agents' goals.
- **Evaluation**. Realistic agents are often evaluated based on their ability to replicate human behavior. Realistic agent research can involve objective measures (e.g., [82]) and statistical data (e.g., [163]). Evaluations are sometimes framed as a Turing Test [153], to see whether the player can tell the difference between a human and the agent. In contrast, believable agents

are usually evaluated in the context of the interactive experience in which they appear, with a focus on how effectively the agent contributes to the interactive media experience. The evaluation focuses on the player's interaction with and subjective response to the character.

- **Part or whole.** In studying agents as part of different architectures and systems, we noticed that realistic agent research tends to focus on particular theory implementations or specific character attributes as the focal point of the research. For instance, agents in PsychSim [123] are created to implement a theory of mind whereby the research is focused on the implemented model and scenarios. On the other hand, believable character research often involves multiple components within one agent architecture. While the agent architecture may focus on a novel computational component, it does not dismiss other believability components from its scope or experience. For instance, Versu [51] pays close attention to social practices (how a character reacts during social situations). However, it does not dismiss other character components within an experience, such as modeling personalities or supplying characters with motivations and goals.

In the context of recent work in deep learning such as large language models (LLMs) and deep reinforcement learning, one might question how such work relates to believability. Are the created agents realistic or believable—or can they be both? One quick rule of thumb in differentiating believability from realism is presented in the artifact and the nature of the experience. If the artifact is intended to simulate real-

world characteristics without the heightening and compression of media, we consider it realism. Otherwise, if the audience is meant to interact with a "character" as a piece of media, we qualify it as believability. For example, by our definition, virtual assistant AIs such as Siri and Alexa fall under the realism category. We note that the goal of creating them is to elicit a response similar to a human working as an "assistant." While they may have some notions of believability (such as personality traits), their core is more focused on authenticity, task orientation, and replicable behavior, all of which are elements that contrast with the exaggerated believable character. For example, the characters in Façade [94], an interactive drama showcasing an escalating fight between two NPCs, Trip and Grace, while making use of natural language interaction that is somewhat similar to current chatbot interactions, demonstrate exemplary believability characteristics (which will be expanded upon in the next section). The type of interaction between a player and an NPC such as Trip vs. the player interacting with an AI assistant such as Alexa is vastly different: One is concerned with believability, the other with realism. Table 2 summarizes the differences between believable and realistic agents, as described in this section.

	Realistic agent research	**Believable agent research**
Research goals	Agent comparable with human/player behavior.	Creating expressive and emergent agents.
Use of cognitive or psychological models	-Constricted by model theories, the model is sometimes the focal point of the research, e.g. [123]. -Core component of the agent.	- Models are used in addition to other components, usually as part of the overall agent architecture. e.g. [105]. - Supplementary to the agent.
Type of experience	-Can be based on objective and subjective experiences. -Able to replicate with somewhat consistent results, e.g. efficiency in patrolling agents.	-Often subjective experiences, -seeks emergent character behavior. -Usually focused on creating unique experiences and interactions.
Answers the question	-Is this how a "human" would act? -Did this character behave realistically? - Did players perceive the agent as a real agent? -Does the path planning work? -Do the agents move in a realistic manner?	-Is this how this particular "character" would act? -Is this character exciting and fun? -How did the character fit with the overall flow of the system?
Used to/for	Validity, effectiveness, and testing architectures or theories, such as validating an emotional theory in specific scenarios or player models.	Interactable experiences such as characters in games.
Evaluation	Evaluation may be based on authenticity, statistical data [19], relatedness to human behavior/ player (e.g. player modeling in	Evaluation is based on an audience's perception, e.g. [117].

	[150]or perception in [89])	
Reality	Grounded, in reality, may be constricted by cognitive and emotional theories [123].	Provides an abstraction of reality or an exaggerated reality contextualized by the provided experience e.g. [97].
Example fields and areas	Virtual assistants, virtual beings, agents focused on the authenticity of an area of affect, FPS tournament agents, agents used in debugging games, and player modeling.	Storytelling, reactive agents, characters in use by drama managers, character personality modeling, characters created for stories and games.

Table 2: Presents the general differences applicable to areas of realism and believability

3 Conclusion

In this chapter, we argued that the term believability is often ambiguous, as it means different things to different communities, which leads to misunderstanding, confusion, and communication issues. As a result, we suggest separating character believability definitions into two groups: one of realism and one of believability. Thus, this chapter defines believability and realism, illustrating the different kinds of artifacts, methodologies, and research areas involved in their creation.

Chapter Two: A Deeper Understanding of Character Believability

Acknowledgement: This chapter has been published elsewhere [6].

Now that we've covered the different aspects of believability and realism, it's appropriate to examine what it means for a character to be believable. In this chapter, we will discuss some theories and frameworks of believability, provide examples and use cases, and review some academic systems that highlight aspects of believability. Before we begin, we would like to note that the terms character and Non-Player character (NPC) are used interchangeably in this thesis. However, we note that characters are included in broader media artifacts (including games), whereas NPCs are limited to game experiences.

1. Background

Mateas [93] broadly described the research guidelines for interactive drama, an agenda that includes believable characters, story, and presentation. Mateas et al's research [93] lists a set of requirements for a believable character as borrowed from Loyal's dissertation [87], including personality, emotion, motivation, change, social relationships, and the illusion of life. The following section will describe the six characteristics of life-like characters, providing examples from academia and

industry showcasing how these characteristics may appear in games and computational systems.

1.1 Believability Characteristics

- **Personality.** "Personality is king," as emphasized by Mateas [93]; a believable character should display a complex and unique personality. One way to consider a character's uniqueness is by comparing it to an animation silhouette. Ideally, a character's personality should be as immediately recognizable as an animated character's silhouette. Characters with rich personalities are often limited to just a few NPCs within a game. The game industry often focuses on companion characters, such as the characters Elizabeth from *Bioshock* [1] and Ellie from *The Last of Us* [107]. In academia, examples of strong character personalities often appear in research games such as *Façade* [94], where both NPCs exhibit individual personalities as the couple argues, responding to each other or to the player in a unique way; this is portrayed verbally or via facial gestures and animation.
- **Emotion.** Characters should be able to express and respond to the emotions of other characters. Emotion under the believability category is tied to a character's personality; it should be expressed in a way that is at least relatively unique [93]. In contrast, most theories of emotion, as implemented in emotion models (e.g., [49, 123, 57]), tend to focus on general emotional responses that can be replicated in psychological experiments. We also note

that believable character architectures will often make use of emotion models, but in the context of a broader architecture, with a focus on emotion presentation, such as the emotional model used in *CiF-EX* , or the emotional states conveyed in *Prom Week* [97].

- **Self-motivation.** Self-motivation represents the character's ability to act of their own volition and pursue their own goals beyond simply reacting to the player or other characters [93] . *The Sims* [95] presents an excellent example of a character's self-motivation. *Sims 3* [95] is a social simulation game whereby residents (*Sims*) of a given town can act according to their needs and desires (with or without the player's involvement!); Sims even have aspirations they try to meet. In academia, we can see examples such as *Prom Week* and *Versu* [97, 51], whereby characters are driven by their goals and have incentives to achieve those goals. We note that other goal-driven agents, such as those in RTS games (e.g., *Starcraft* [167]) or the machine learning literature employ motivation but disregard the other characteristics mentioned in this section, putting them in the category of realistic modeling of player behavior rather than believable characters.
- **Change.** Change represents the character's ability to adapt and evolve with time. A character's sense of change can be seen through growth of their personality (gradual change that matches their base persona), linear change such as aging, or physical change. A story plot also involves character change based on the challenges they face [47]. Change through time has yet to be

explored that much in terms of character believability. While games that use time as a feature exist, such as *Braid* [114], the game's mechanics focus on puzzle-solving rather than character development. Some episodic games, such as *Assassin's Creed* and *Telltale: Batman* [147, 155], have characters that appear to grow as time goes on. In actuality, the player experiences a snapshot of that time frame, showing abrupt changes without actual continuous character growth. One good example of change is in *Beyond: Two Souls* [124]. In *Beyond,* the player plays Jodie, a character with supernatural abilities. The player, throughout the game, experiences life through Jodie's eyes—from childhood to adulthood—through episodic structures; recurring characters (NPCs) have shown personality growth (albeit limited), manifesting aging and progressional development based on circumstances. In academia, *Talk of the Town* [132] accounts for time factors such as establishing historical roots for characters and allowing characters to forget. *Prom Week,* a social simulation game in which players interact with characters in the week leading up to prom [97], also shows a character's advancement through episodes achieved by character goals. Another example is in *Façade* [94], where the player notices an escalation in the drama surrounding Trip and Grace.

- **Social relationships.** Social relationships are measured by the character's ability to interact and react with other characters in a given environment; furthermore, change should happen as a result of that interaction [93]. Social

systems are one of the more explored areas in character believability in both industry and academia. In industry, games such as *Harvest Moon*, *Redshirt*, and *Stardew Valley* [91, 36, 122] feature intricate relationships among characters; these social systems influence in-game events (such as birthdays or town festivities) and dialog trees. In academia, social systems have made significant advancements, such as CiF [98] and its implementation of social physics in *Prom Week* [97], *Neighborly* [69], a sandbox reconstruction for large-scale social systems such as *Talk of the Town*, and *Kismet* [141], an authoring platform for social scenarios and simulations. Social systems are a great source of character believability. Not only do social systems reflect relationships, they often connect other believability characteristics, such as motivation, change, and personality, although in varying degrees.

- **The illusion of life.** The illusion of life combines different requirements, such as the character's ability to remember, react to situations, and maintain and manage different goals simultaneously. The Oz philosophy [93] is closely associated with behavior-based AI approaches, as a character needs to broadly integrate multiple capabilities such as goals, perception, and memory. These lists of believability characteristics remain influential today. Researching believability, we found other factors that are important for believability, as established in the literature.
- **Contextualizing characters and referencing expected roles.** Characters in games and story worlds often portray particular roles, usually specific, aided

by the character's personality. Warpefelt [165] described behaviors related to character roles (such as Vendors, Services, and Questgivers) to help contextualize characters and provide helpful specific behaviors that are expected in the context of their portrayed roles, which helps structure player interaction.

    - **A quick side note on morality and general roles:** In relation to this dissertation's morality theme, we highlight that characters in (most) games, by nature, often take differing roles in relationship to the player. While NPC roles are specific, another dimension often exists, one of morality. At a basic level of interaction, we notice that NPCs can either support the player, oppose them, or act neutral toward them. While not all NPC actions towards the player are moral, morality or moral actions exist as a subset of game genre, where NPCs and players often undertake actions and choices with moral consequences. **Chapter Three and Chapter Nine** highlight morality further, reflecting how they appear and how they are operationalized.

- **An NPC should portray its role convincingly.** It is often not enough for an NPC to simply have an assigned role. It needs to act it out in a way that is genuinely convincing. Characterhood [165] borrows Dennett's [139] notion of personhood, emphasizing the importance of an NPC's ability to act rationally (perceived as a rational being) and intentionally (actions are made with perceived intentions) to qualify as a basic NPC or character. Like a

character’s personality, uniqueness separates an NPC’s Characterhood from other in-game entities [165]. One factor differentiating NPCs or characters from other static objects is an NPC’s agency, which helps NPCs perform their assigned roles. This helps us understand why abstract voiceless characters such as those in *Lim* [102] can be perceived as believable characters that act with intent.

1.2 Design Patterns for Believable Characters

Here we review believability from a believability pattern perspective. Like game design patterns, these patterns provide recipes for a designer to follow so as to solve design problems [22] and detect character believability within games [166], [81]. Lankoski and Bjork [79] focused on believable character categories obtained from cinema studies that include such factors are the following: a character's presentation (human body), a character’s state (including self-awareness, intentions, and self-impelled actions), a character's ability to express emotions, a character's utilization of natural language, and finally the character's ability to have persistent states (such as recalling facts). As our readers may have guessed, unsurprisingly, most of these qualities can be mapped to the believability characteristics mentioned above.

To derive these qualities, the authors [79] analyzed Claudet Prttivk, a shopkeeper in *Oblivion* [16]. They chose Claudet as she represented a standard NPC that is also capable of performing several actions (such as acquiring rumors and path

following) or being a participant in other NPC social exchanges. Their analysis of Claudet identified several believability patterns[1] under the four categories, noting how Claudet fails and succeeds and how these patterns relate to one another. For example, Claudet succeeded at the believability pattern in *Own Agenda* (supporting the self-awareness quality), where Claudet highlighted *self-preservation* goals, as the character can equip gear and defend herself from harm. While *Own Agenda* was successful, the authors also included examples of Claudet failing to hold believability patterns, such as her failure at acquiring the pattern in *Emotional Attachment*, as she exhibits apathy whenever the player jumps around her shop.

Lankoski et al. [81] further elaborated on NPC patterns, focusing on architectures and AI rather than appearances. Their categories were expanded to include perceptual activity,[2] non-mechanical behavior,[3] and goal-driven behaviors,

[1] The authors included an extensive list of patterns as defined under each of the five qualities, which can be referenced here [79].

[2] Perceptual activity includes patterns that are planning or observation-based, as well as patterns that account for the NPC's behavior and reaction to stimuli. The patterns include Memory of Important Events, Awareness of Surroundings, Emotional Attachment, and Sense of Self.

[3] The authors coined the term non-mechanical behavior to mean "unpredictable behaviors" that match the character's base persona. The list of patterns includes ambiguous responses and open density.

with each category defining and recategorizing their list of patterns. For example, goal-driven behaviors include goal- or task-oriented patterns, such as the *Own Agenda* pattern explained in an earlier example.

As the authors imply in [79], it is probably impossible to design a long-lasting, fully believable character, as one cannot cover all patterns. However, we firmly believe that a character should hold the believability qualities and characteristics mentioned in earlier sections at least to some extent. It is no mere coincidence that the extracted believability patterns have a strong affinity to that of the Oz characteristics, suggesting the significance of intentionally designing characters with these qualities.

2. Surveying the Field

2.1 Social Simulation and Autonomous Architectures

This section introduces representative autonomous character and social simulation architectures that often incorporate believability characteristics. As discussed in **Chapter One**, social simulation and autonomous architectures are two areas relevant to character believability. In a later section, we will include a summary of our review of systems, followed by a short discussion in which we identify possible gaps in the literature.

Let us take *Prom Week* as our first representative system. Prom Week [97] is a social simulation game in which players interact with their classmates in the week

leading up to the prom. *Prom Week* utilizes Comme il Faut (CiF), [96], a social simulation architecture that enables intricate character relationships by focusing on different character aspects such as a character's attributes and social standing. *Prom Week*'s gameplay consists of the player solving "social puzzles" through which the player can manipulate NPCs into taking social actions. For example, the player can manipulate characters into making a character popular; however, that is dependent on the character's volition. Through CiF, *Prom Week* can manage a character's complex social state, involving such aspects as a character's relationships, emotions (i.e., feelings toward characters or objects), attributes (e.g., competitive), and social status.

Before moving onto other systems, we want to provide additional information on CiF [96], as CiF has been applied in some of the upcoming social systems described in this section. CiF's architecture includes knowledge representation as seen through the social fact and cultural knowledge bases. A social fact includes information about characters such as the identify of the initiator and the target of social exchanges.[4] These social facts store information about previously simulated social exchanges to influence future social exchanges. We should note that social facts can affect a character's relationships. The cultural knowledge base helps players detect the NPC's personality. It includes shared information among all characters,

[4] A social exchange can be seen as the social interaction between characters: for example, Mike asks Sarah to the prom but Sarah rejects his invitation [97]

such as how characters evaluate world objects. For example, viewing an object as "lame" in *PromWeek*. CiF includes intricate character relationships whereby each character shares a connection with other characters; the types of connections are romantic, friendship, and respect-based. CiF employs social status rules, whereby conditions must be met before a social exchange or change in social state can occur. CiF also develops a character's personality through the previously mentioned knowledge bases and social networks, in addition to individual character traits and needs. Lastly, CiF reasons over social exchanges (e.g., flirting with a character for a relationship goal) by calculating both the interacting and recipient character goals, desires, reactions, and types of social games.

CiF-CK [61] adapted the CiF architecture and implemented it in *Skyrim* [17], a commercial open-world RPG where the player plays the role of a Dragonborn, blessed with dragon-like powers to save the world. The authors chose Skyrim for its popularity and modding (modifying) capabilities. Their architecture combined CiF with Creation Kit, a Bethesda modding tool that layers mods on the original game rather than rewriting the base game. Their work connected CiF's social exchanges (outcomes based on knowledge representation, social state, and personality attributes) to that of quests; for example, the preconditions of CIF's social rules were adapted into the quest start conditions. Skyrim characters were also modified with variables to accommodate their CiF modification (e.g., adding permanent values for character traits). Generally, they modified CiF to better integrate into the mechanics of Skyrim. Their main contribution lies in adding a belief model alongside the social

networks, by which a character may believe in false information and act accordingly. For example, character A likes character B despite character B hating A. A will advance toward B, as A believes that B likes her. Lastly, the authors ran a study focused on the NPC's believability, player engagement, and experience, with positive results. Compared to non-CiF-CK characters, players noted that CiF-NPC actions seemed more comprehensible, less mechanical, and expressed higher levels of enjoyment.

[115] presents another system that integrates CIF's social capabilities. However, their system, which was built[5] specifically for modeling merchants in game worlds, was tested using the game *Conan: Exiles* [55]. Like CiF and CIF-CK [61, 96] the model shares many base similarities (such as a social fact and a cultural knowledge base); however, the authors have expanded their model to account for merchant-specific capabilities. For example, the enhanced model includes a player preference knowledge base used for tracking a player's liked items (based on facts or merchant assumptions about the player). Another example is changing the variables in the social network to reflect how a merchant sees the player (e.g. in terms of monetary interest, perception, and social bond) as represented by the aforementioned knowledge bases. Lastly, the authors ran a player interaction study focusing on

[5] They used Devkit, a modding tool that provides model, sound files, Al, and blueprints for Unreal, with the game Conan Exiles.

believability, engagement, and enjoyment, which found that the social and believability elements improved the overall player experience.

Comme il Faut - Exiles (CiF-EX) [105] is another system inspired by Cif and Cif-CK whereby the authors used *Conan Exiles'* NPCs as a testbed as well. CiF-EX shares many similarities with the aforementioned systems, such as reasoning about character traits and maintaining social states. However, one of CiF-EX's main contributions lies with improving Cif-CK's belief system, whereby characters can now have beliefs toward other characters beyond the interacting character. Taking the example mentioned in Cif-CK, when character A advances toward character B with A believing that B likes A, this modified version allows for an additional character, C, to infer the relationship between A and B. They also improved upon the emotional state of the characters. Instead of a binary character social state, they extended it to become a continuous scale, allowing for varied emotional responses.

Talk of the Town (TotT) [132, 131, 130] is a historical town simulator that incorporates characters, town information (roads and buildings), social networks, and town events and businesses. In contrast to CiF, *TotT*'s unique contribution lies in how characters and the player share knowledge (e.g., gossiping or misremembering) whereby CiF does not model knowledge dissemination. Characters in *TotT* can communicate both false and accurate information when conversing with the player; the player also has the power to spread misinformation (such as falsely describing a character) and, in turn, corrupt the town's knowledge stream (i.e., the town's gossip network). Not only that, characters even have the ability to forget or misremember

information over time. Characters can reason using their held beliefs, viewed in this case as information (e.g., a belief that a character's hair color is brown). Furthermore, a character's held beliefs can change depending on the evidence supporting the belief or because of the character's misremembering information. However, beliefs are portrayed on a scale that increases depending on the strength of the evidence provided. Lastly, characters have mental models[6] and social networks, whereby relationships between characters can affect a character's beliefs. For example, characters are likely to believe false information if they have a trusting relationship with another character that is lying to them.

One interesting application of *TotT* is Bad News (BN) [133], a playable experience that allows the player to interact with characters from the simulated TotT town by interacting with a live actor portraying different NPCs at different times. The actor receives instructions from a Wizard (a developer operating the simulation for queries) behind the scenes. An interesting takeaway from BN is that real actions may be perceived as NPCs. BN also allows a greater portrayal of emotion, as presented by the actor's tone and body language. According to the author's results, some players experienced discomfort initially, as it was hard for players to improvise; nevertheless, they eventually adapted well to the role-playing aspect of BN. Lastly, BN allowed

[6]Mental models include knowledge modeling of other characters or locations such as a character’s name or where an employee resides.

players to express greater agency, portray greater emotion, and experience a unique play experience.

Another architecture that incorporates knowledge and emotions as main features is the work by Lankoski and Bjork et al. [81]. Lankoski et al. 's architecture includes knowledge, perception, decision, and emotion components. What is interesting about this architecture is how it handles incoming knowledge. Several factors affect knowledge. First, knowledge is disseminated and ranked by the perception component[7] and stored in either the working memory or the long-term memory elements[8] of the knowledge component. Furthermore, incoming information is associated with an agent's emotional state, which can be retrieved later, depending on the agent's current emotions. Emotions are a key factor for the agent's decision process, with agent relying on an effective, emotional model, dubbed Emotional Behavioral Network (EmoBN)[9]; this model includes states, goals, and behaviors. The

[7] The perception component can be visual (what the agent sees, as per the current model), although in the future, the authors aim for the cognitive approach, which centers on the objects an agent is thinking about (for example, if the agent searches for apples, it will notice apples, rather than other objects in the area).

[8] The authors factor in emotions when storing information, which is related to how an agent later retrieves a memory.

[9] The emotional model uses behavioral networks (BN), an area often associated with areas of affect and realistic behaviors. Above we noted key concepts

states constitute an agent's beliefs about the world, with their project modeling it as a continuous value from 0-1, rather than a strict binary element. Lastly, the authors note how their system addresses the believability patterns mentioned in section 3.1.2. For example, in "Awareness of Surroundings," the agent can detect objects depending on their emotional state, affecting what the agent focuses on. The object is then stored in working memory and later moved to the long term, depending on the object's significance, at which time the agent can finally reason about the object (e.g., Claudet views an object as stolen vs. legally purchased).

Façade [94] is an interactive drama experience similar to an act of a play. In *Façade*, players visit the couple, Trip and Grace, in their apartment and witness the couple's escalating tension as the drama unfolds. *Façade*'s experience enables players to interact using natural language, supporting the player's freedom of expression and agency. Furthermore, the agents in *Façade* are reactive[10] to player actions; their actions are guided by beats (beats influenced by [100] are actions with preconditions and postconditions used to advance the plot) selected by a drama manager.[11]

---

for our review. However, BN and behavioral networks are beyond the scope of this literature review (for more information, please refer to [46]).

[10] The agents are built using ABL, a reactive planning behavior language based on Hap [88], which supports an agent's behavior, such as the ability to use parallel actions.

[11] A drama manager assembles the story using the mentioned beats.

Moreover, the agents have a broad range of capabilities beyond simply reacting to the player. Agents can initiate a conversation or interact with objects in the world as well as portray emotions (via animation and tone). The interactions themselves (presented by beats) can hint or guide the story along. For example, an agent can offer players drinks; when doing so, the agent hints at the conflict of their marriage. We believe that *Façade* represents an example of the kind of playable experience that expresses believable characters well.

2.2 Storytelling Systems and Narrative Planners

While our primary analysis looks at social simulation and agent architectures, we note that other avenues can improve our understanding of believability.

As mentioned in Chapter One, storytelling systems and narrative planners are some of the predominant areas of believability-based work. Story and characters work hand-in-hand to enhance the overall character and situate them within their worlds.

For one, storytelling systems [163,164] often guide characters in their world; they help give characters a sense of autonomy and motivation, an essential criterion of believability. They fuel a character's motivation by establishing and maintaining each character's unique goals and conditions.

Planners also play a part in creating dynamic interactions between different characters. Like a director on a stage play, planners have the potential to direct

characters in engaging directions, ensuring stories unfold in interesting and often climactic ways [94].

To elaborate, planners have the potential to make characters appear not only as social entities (a believability criterion) but also contextualize characters with unique roles within the overarching story. For instance, depending on the state, *Façade's* [94] system can have characters appear as a normal couple, a bickering one, or an awkward one all the while situating the player and the NPCs as characters in a dramatic scenario.

The following paragraphs will examine exemplary planners and current research within the field, drawing lessons learned, unique approaches, and outcomes.

Our first example looks at Cavazza's work in *Madame Bovary on the Holodeck* [31]. The author's work is based on excerpts from the Madame Bovary classic novel, where players roleplay a specific character (Rodolphe) and interact with real-time animated characters running on an immersive CAVE-like display. The main feature of the story engine is its ability to connect the player's real-time actions, both auditory (e.g., speech) and physical (e.g., gestures), into the story's context, influencing the character's emotions and actions (seen as story plot).

Another unique feature is that the engine operates on a *character first* basis, meaning the engine runs on a character-based approach rather than a plot-driven approach; this feature adds autonomy to characters, as they operate and act according to their own volitions and feelings. While the resultant gameplay, due to technical

limitations, is shortened to two minutes, it still enables a dynamic and emergent experience for the player.

In an earlier section, we discussed *Façade* [94] through a social simulation and character lens; here, we look at *Façade* from a storytelling perspective. *Façade* uses natural language to consider the player's utterances; through the concept of story beats and the system's drama manager (explained earlier), the player can interact more dynamically with the characters, participating in a live drama.

The system's structure makes the narrative highly dynamic and reactive to the player's responses, creating a sense of immersion and involvement. Not only can the player play a character, they can play *any* character, including themselves. The system allows the player to feel as if they are part of the story, escalating or alleviating dramatic moments. Through the gameplay, characters often fit into their roles (i.e., characters in a dramatic fight) but also situate the player in any role they want to play, an important aspect of believability.

Cavazza et al. [30] developed an interactive narrative prototype based on sitcom themes. What's unique about the system is its ability to formalize character roles as plans, thus allowing the character's behavior to change dynamically at runtime.

Additionally, the prototype highlights unpredictability, an element that when balanced can aid in character believability. According to the author, unpredictability can be achieved in numerous ways, including the character's location, the interactions

between different character plans, the character's mood, and the player's intervention or character influence in the game.

Similar to the author's earlier work, the architecture follows a character-based approach rather than a plot-based one. According to the authors, one problem arising from character-based approaches is balancing story variability while adhering to genre conventions, as conventions help player understandability.

Another planner-based work examines the characters created for Fear-Not [11], an anti-bullying dramatic interactive experience with the goal of teaching players about bullying. The characters and interactions in this project utilize concepts from interactive drama and performances. As mentioned, interactive drama is associated with believability under the OZ project's definitions and goals.

The system uses emergent narrative through planners, allowing the characters to act autonomously. The characters are also structured using the OCC (Ortony, Clore, and Collins) model for emotion generation and appraisal. The experience also splits audience members into groups, with each group assigned responsibility for a character, creating roles and dynamic interactions. This project is further discussed in **Chapter Three** as we explore morality in games and academia.

Other work by Riedl & Young [125] further discusses the current state of narrative systems and planners, highlighting interesting and important concepts. The authors highlight two problems facing a narrative system's understandability, which include a logical flow of the generated story and the character's believability.

To achieve believability, the authors communicate *intentionality* (as mentioned, intentionality is associated with character goals and motivations, a criterion of believability). An interesting point the authors make is separating the goals of the agent (the narrative character) from those of the author. Not only should characters be perceived to have intentions, but these intentions should also be *observable* to the players, meaning that the narrative planner should provide enough context and information that players are able to interpret the character's goals and motivations. Thus, the authors researched a system that considers a mock "audience" perspective, ensuring that character intentionality is understandable relative to the overarching plot.

We consider storytelling systems and planners an essential avenue in character believably studies. We believe that studying these systems can help us develop ways of improving the character's overall believability. Through representative examples and theories, we highlighted the importance of an NPC's intentionality, autonomy, and roles, all aspects that mirror the aforementioned believability criteria.

Additionally, the systems highlighted here bring the audience's experience to the forefront. We saw examples in which "fake audiences" were taken as an actual parameter in creating character plans or instances in which the player was allowed more agency in choosing their roles within different systems, adding dynamism and emergence to a player's gameplay experiences.

While planners and storytelling systems are instrumental in creating believable worlds, different mediums may call for different methodologies. For instance, drawing from performance studies, the authors [10] argue for a need for narrative theories that prioritize interactivity in platforms such as virtual reality (VR). The author highlights that story planners can be unsuitable for interactive VR environments and proposes instead a methodological approach inspired by improvisational theater and live role-playing games to create flexible and engaging narrative experiences [10].

2.3 Virtual Agents and Human-Robot Interactions.

Another avenue worth mentioning is the body of work under virtual agents (VA) or human-robot interaction literature. At this point, the reader may wonder, "But you mentioned that most of these areas are associated with realism," to which we largely agree. However, we note that one of the differences between reality and believability is dependence on the nature of the experience and artifacts themselves. While most of the research in this area falls under realism due to the nature of the artifact and the author's intent on creating realistic human-like behaviors, other artifacts fall under the believability camp as a result of the components and the nature of the interaction.

Additionally, VA work often shares many components used in believability research, such as modeling agent emotions [93,121,84] sociability [40,54,134], or creating emergent behaviors [40,84] (at least in some cases), though contextualized

within a realistic character. Nonetheless, examining these areas can further enhance our understanding of specific believability-related components as we reframe them for believability-based characters.

Our first example explores a VA [41] created to study the link between believability and emotions expressed in multiple formats (verbal and nonverbal) by the VA. While the authors consider the user's engagement and suspension of disbelief as a believability criteria, they blur the line with realism in their design and situated simulations.

The study [41] asks users about the VA / agent's reaction in a simulated scenario in which the agent Greta plays cards, assists, or advises its simulated user. Greta [39, 121] has two main components, one reflects her personality while the other displays her expressive behaviors, aided by the use of XML languages. Participants were asked to evaluate the agent's emotional responses. Their results generally indicated that believability is associated with the agent's emotional response appropriateness and correlates with the VA's warmth and competence. Additionally, the agents that acted with multiple modalities (speech and gesture) were perceived to be more believable than those acting in one. We believe that studies such as these can aid us in considering other minor avenues regarding believability, such as exploring emotion in greater detail and the expressivity of that emotion. We note that while studying elements like appropriateness is valid in its own right, under a believability lens, we should contextualize it by the game or artifact itself; after all, believability

usually seeks greater expressivity and exaggerated behaviors, leaving the designers to deem what is appropriate.

Another VA situated in realism is KRISTINA [160], a knowledge-based, emotional, conversational virtual agent acting as a medical assistant. An interesting aspect of Kristina is the emergent nature of the system, where the VA's logic is guided by a reasoning-based dialogue planner rather than predefined responses. Additionally, the agent can understand users by reading their emotions (through gestures or facial features) and searching the web for information, contextualizing the experience to the user.

Although the agent is primarily realism-based, some features can be extrapolated to create more believable characters. For one, we can learn more about expressivity through the agent's planner, which allows dialogue flexibility and responsiveness to unexpected inquiries, an aspect often encountered in many academic and believable-based agents. Additionally, how the agent responds to emotions and facial gestures can enhance the player experience, especially as it allows more dynamic interplay between a character and a player.

Another interesting VA is the *Affective Guide with attitude [84]*. The VA is a guide that guides users on outdoor attractions. The agent utilizes two interesting features: an emotional system that is more emergent and biologically inspired and a storytelling system.

The emotional model is influenced by the agent's built-in motivation factors (competence and certainty[12]). These motivators affect different modulated values (such as *resolution level, arousal level, and selection level[13]*), which interact with the user's feedback to produce dynamic and emergent emotions.

The other interesting part is the storytelling feature of this agent. The agent utilizes beliefs (beliefs about the user's interests) along with its personality, emotion, and memory components to dynamically tell stories about the attraction. The researchers studied different variations of the guide (guide with emotions and attitude vs. ones without either element). They discovered that emotions played a part in enhancing the experience, but due to the study's limitations, the results were insignificant in drawing formal conclusions between agent types.

In another study, researchers [134] dived deeper into user experiences by examining the player's posture through a social robot playing chess with the player.

[12]Competency according to the author is the agent's ability to handle differing opinions with the user on presented issues whereas certainty masseur's the user's predictability levels

[13]These represent system elements that are affected by the agent's motivational factor and the agent's current state. For instance, the arousal level represents the agent's readiness to act and is inversely related to the resolution level and proportional to the selection threshold. For more information on how these parts interact please refer to [84].

The study highlights player engagement in interesting and new ways, including the player's recorded posture and movement as indications of engagement. While this study falls under the category of realism, it still highlights the importance of engagement, leading to new ways of looking at player engagement.

While most of the represented examples from our point of view fall under realism, we can still learn more about specific components, especially if they contribute to emergent and dynamic character interactions as discussed earlier.

2.4 Mind the gap: a Discussion on Social Simulation and Autonomous Architectures

In this section, we take a more focused approach to believability, focusing mainly on representative examples from our above-mentioned social simulation and autonomous architectures.

Generally, we note that the systems above account for most believability characteristics; however, the characteristics are usually focused in one or two areas and presumably shallow in others (with limited exceptions, as we will discuss). For example, CiF-influenced systems [115, 61, 105] focus on social interactions (social relationship characteristic), paying close attention to the relationship modeling between characters and a character's social goals (self-motivation characteristic). On the other hand, CiF representations include limited emotional behaviors (e.g., conveying emotion as a binary value whereby a character is sad or happy without depicting varying degrees of emotions). To be fair, we believe that supporting all the

factors of believability is a complex task; there are limitations such as time, technical capabilities, and resources to consider. However, in the following section, we hope to provide fruitful directions with regard to emphasizing character believability across multiple believability facets.

2.1.1 On personality

In covering the believability characteristics across the mentioned systems, we noticed that personality modeling in research games often factors in as a result of assigned character traits or via psychological models. For instance, *TotT* [131], [130] uses the Five-Factor personality model [58] in influencing how character-to-character social interactions change the friendship and romantic relationships between characters. On the other hand, some models use trait-based characteristics usually assigned by keywords or traits that a character possesses (e.g., a "friendly" personality in Cif-CK [61]); in some cases, these personalities are influenced by social relationships. Furthermore, these personalities can either be public or private (viewed by other NPCs) and permanent or changing. Lastly, we noticed that these personalities could be conveyed in three main ways: by textual information, by tone, or by animations and expressions (including UI expressions)—with the exception of Bad News [133], as it was a live performance.

2.1.2 On emotion

We find that character emotions are usually modeled through psychological frameworks (such as the OCC model in [105]), cognitive frameworks (use of signals

in [81]), or as a result (or reaction) to other characteristics such as social interactions affecting NPC responses. Furthermore, we noticed a trend whereby presented architectures increased an NPC's emotional state from binary values (e.g., happy or sad) to ternary (e.g., happy, neutral, or sad) to cumulative values across the surveyed systems (e.g., increasing levels of fear in [81], by having emotions influence other emotions cumulatively). We note that emotions are conveyed similarly to the presented personality characteristic, except for the time during which emotions are temporary.

2.1.3 On social interactions and motivation

Social interactions are one of the core features of believable NPC architectures. Through social interactions, players can gain knowledge about the world and the characters. Furthermore, social interaction allows unique player experiences, such as increasing biases an NPC may have about the player based on their past social interactions (e.g., a merchant learning about player preferences in [115]).

We also noticed that social interaction and motivation characteristics strongly correlate with one another, whereas most of the NPC goals in the surveyed systems are social goals. We also observed how character motivations could be somewhat limited, as a result of the type of interactions. To elaborate, we distinguished two primary ways a character's motivation presents itself. The first is *player-centric*, with characters reacting to players. For example, the NPC's goals in *Prom Week* [97] are identified after the player chooses a character with whom to interact. Although

characters in *Prom Week* consider their own volition before completing a social exchange, the player may not perceive the NPC's goals without first engaging them (however, we note that CiF can handle these autonomous interactions). The second is *NPC-centric*, where NPCs do not differentiate between players and NPCs; in other words, NPCs pursue goals regardless of the player's interaction. We believe this second method supports emergent behaviors arising from character interactions, a motion supported by believability aspects.

2.1.4 On beliefs and values

One common element that emerged from our surveyed systems is the use of belief modeling in believability-based systems. From our review, we noticed that most social systems either focus on social aspects with shallow belief modeling (which can be non-existent) or focus on beliefs as the main contribution without much effect on other characteristics.

For example, some systems, such as CiF-CK [61], employ the Theory of Mind to model beliefs; however, when applied to social interactions, they are treated as knowledge facts that a character has or does not have. Furthermore, the presented social system beliefs are mostly tied to relationship beliefs rather than world or personal beliefs, especially as those beliefs change, which presents a gap in the literature, as we will discuss in **Chapter Four**.

Chapter Four dives deeper into beliefs and values and includes reviews of additional systems that highlight beliefs in their design. The use of beliefs as an

influential element for believability is an understudied topic, especially as it relates to morality.

2.1.5 On change, a gap in the literature

There exists a gap relating to the implementation of the change characteristic. Most of the surveyed systems do not account for real character growth or how NPCs handle situations differently based on character development. Furthermore, elements of change in the surveyed systems are implemented at a surface level. For example, characters can change their social standing [97], forget things [130], and change their beliefs as seen in *Versu* [51] (*Versu* is further discussed in **Chapter Four**), affecting their social interactions. However, it appears as an abrupt change rather than a result of a character's development with respect to time (i.e., time as in the duration of the game, story, or game loop). When implemented, changes in systems are usually limited to the player character, as we will discuss in **Chapters Three and Four.**

2.1.5 On the illusion of life

Certain elements worthy of consideration fulfill the illusion-of-life category. As a reminder, the illusion of life defines an NPC's ability to partake of broad functions [93]. An interesting application of the illusion of life lies in how NPCs handle knowledge. The surveyed systems included characters that can remember, forget and lie [132]. These aspects further enhance other characteristics such as social interactions and gameplay. We further observed that interactive NPCs (as defined in the GAM [165]) in systems such as *Façade* tend to incorporate more characteristics

than other surveyed systems. We believe this is a result of their interactive nature (beyond simply reacting), where NPCs can initiate conversations, address in-world objects, and participate in player initiations.

4. Case Studies in Believability: from Bioshock to Lim

Now that we have discussed what believability entails in academia and theory, we will examine a few specific characters in popular games to cement our understanding of character believability.

We will first look at Elizabeth, a companion character from *Bioshock Infinite* [1]. Elizabeth was first introduced to the player through tiny glimpses, literally, whereby the player peeks at her through blinds and windows. These tiny glimpses show her passion, personality, and goals, even at an early stage. Figure 1 shows Elizabeth's curious and artful personality and her desire to visit the outside world. Upon joining the player, we get even more details about her personality. Her actions mirror her rich personality; they even appear unusual compared to the other NPCs around her. Figure 2 shows Elizabeth inspecting a boat. Compared to the background NPCs, this simple scene shows Elizabeth's curious personality, making her stand out.

Figure 1: Elizabeth from Bioshock, expressing personality traits and goals early in the game.

Figure 2: Elizabeth from *Bioshock*, inspecting the boat.

Throughout the game, we see glimpses of her following her own agenda and maintaining an overall illusion of life; she often comments on other NPCs, actively engages with the player (e.g., gives the player discovered items), and even contributes her thoughts about various elements in the world. For instance, when the player is presented with two styles of brooch (as seen in figure 3), Elizabeth comments on their designs, showing a mild preference for the free bird brooch (perceived as such by the player).

Figure 3: Elizabeth from *Bioshock*, interacting with the brooch[14].

We also note that Elizabeth maintains a friendly relationship with the player. We can see her portraying various emotions, depending on the situation and with whom she interacts. At this point, although we can see a collection of believability characteristics delivered here, one may wonder: Elizabeth is a heavily authored, developed, and animated character—Is that needed for all NPCs to showcase believability? To answer this, we will briefly examine a few other characters.

The first character to mention in this context is the Valentine butcher, a background vendor in *Red Dead Redemption 2* (RDR2) [126]. One wonders how this background character is believable if he is not as heavily authored as Elizabeth. We argue that this NPC actually fulfills most of the believability characteristics despite its minor role in the world. Upon interacting with the player, the butcher usually comments on items, showing a preference for those that would sell well. Actions such

[14] All Bioshock images are captured and composed from gameplay videos, found https://www.youtube.com/watch?v=7-WH-i2Piy4

as these reveal the butcher's agenda and display his personality, making the player perceive his personality as standoffish or greedy.

After all, perception, as hinted in **Chapter One**, is a key to evaluating believability. We also note that the butcher depicts other believability characteristics that contribute to his overall believability. For instance, if the player forgoes the Valentine butcher for another vendor and comes back to the butcher after some time, the butcher comments on the player's abandonment of him, hinting at feelings of jealousy and betrayal. Lastly, we can see a few elements of change, such as the butcher's shifting appearance, depending on what transpired with the player or the world. Figure 4 shows a few different representations of the same vendor. We note that *RDR2* used uncomplicated tricks and behaviors that are expected within the role of the character, making them seem believable as they lie within the player's expectations while, at the same time, exhibiting characteristics that are appropriate to the assigned role, contributing to the character's overall believability.

Figure 4: The Valentine butcher as seen on different occasions[15].

In the next couple of examples, we investigate how believability is not limited to humanoid characters. After all, we consider how believability is linked to the "character" and not the authenticity of human-like behavior. To that end, we briefly examine the characters Dogmeat, Glados, and Lim NPCs.

Dogmeat is a companion K9 character from the game *Fallout* 4 [18]. This furry friend is one of the player's earliest companions in a post-apocalyptic world. Throughout the series, Dogmeat acts in support of the player, depicting diverse emotions and reactions to different scenarios (e.g., attacking enemies and howling mournfully), illustrating believability characteristics. Another non-human example is Glados. Glados, a menacing AI with a sense of humor, is the player's main adversary in the game *Portal* [157]. Glados maintains a watchful eye on the player and often

[15] Images of the Valentine butcher were retrieved from Google's search engine, then assembled and modified here.

comments as the player solves the various puzzles in the world. Glados demonstrates a humorous personality, motivated by the desire to stop the player, portraying a variety of emotional responses (despite being a robot). Lastly, we look at NPCs in *Lim* [102]; despite being abstracted squares, they depict believability characteristics that the player perceives as such. For instance, in Figure 5, we can see square NPCs pushing the player around in what the player perceives as aggressive behavior. These squares portray emotions and personality through simple movements and color changes. In fact, Heider and Simmel's early animation work [65] seen here [168] showed how abstracted shapes can tell a story and could be perceived as characters with goals and personalities.

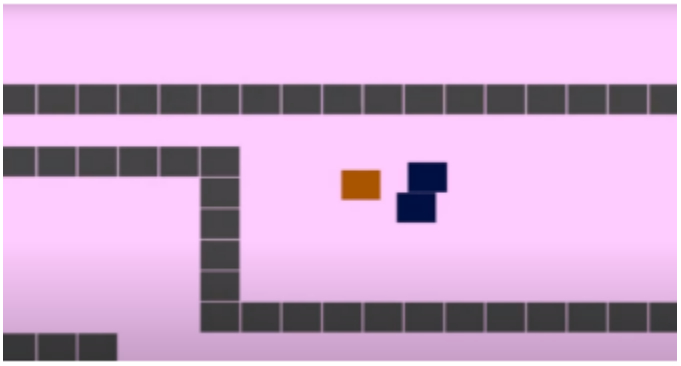

Figure 5: Squares pushing the player in Lim[16].

To some extent, all of the characters presented here present believability characteristics. There are, of course, limitations to incorporating all of the

[16]Image captured from video https://www.youtube.com/watch?v=Dr0LdXhSnFkt=45s

believability characteristics, because of design, authoring, or technical limitations. Some combination of these characteristics often portrays a believable character. We also note that some characteristics innately elicit other characteristics based on our perception. For instance, a character's social capability can provide perceived personalities by the player as antisocial or standoffish.

4. Believability Evaluation

Now that we have seen several believability examples and how they appear, one may wonder how we can evaluate them. What tools can we use to assess our characters as believable?

4.1 Evaluation Based on the Player's Perception.

As noted by Mateas [93], an essential component in the success of a character's believability is the audience's perception. Therefore, considering the audience's subjective response in evaluating character believability makes sense. However, the type of questions asked matters in assessing believability. Care should be taken to avoid confusing believability questions with those of realism. The nature of the realism-based questions does not fit well for believable characters [93]. For example, one should avoid asking, "Does this seem realistic?" or "Is this what a real human would do?" These types of questions do not apply to the exaggerated believable character. As we argued in **Chapter One**, believable characters are far from realistic ones; in fact, designing scenarios that lean more toward realism can

raise the user's expectation, in turn causing skepticism when they cannot explore their intended actions [162].

4.2 Evaluation based on Qualitative Measures.

[150] describes and collects methodologies from different articles for realistic characters that can be adapted for believable characters. In particular, the author examines subjective assessment and forced assessment. Forced assessment is information gained from questionnaires and surveys, while subjective assessment focuses on open responses during play/interaction. The latter is similar to the known think-aloud method [80, 150] used in game design studies; as the name suggests, participants are encouraged to speak their minds as they play. Through these methodologies, we can gain rich qualitative data that helps us identify and understand how characters are perceived.

4.3 Evaluations based on the Believability Characteristics

Gomes et al. [59] focused on evaluating characters' believability in interactive narrative by adapting and expanding on similar criteria referenced in this article. The authors [59] note that simply asking the audience, "How believable is the character?" leads to uncertainty. Instead, the authors encourage the players to partake in answering questions about the believability dimensions themselves. The dimensions include aspects of believability gathered from various literature, including the following: personality (individual behavior that is unique [87]), emotional

expressiveness (how a character expresses emotion [87, 117]), social aspects (can the player identify social relationships [87]), predictability (maintaining the proper balance between both predictable (repeatable) and unpredictable behaviors) [83], [117]), visual impact (an agent can draw attention [83]), behavior understandability (can the player understand the NPC's thought process? [117]), awareness [83, 87], change with experience (e.g., character growth [100]), and finally behavior coherence [14, 117]. The authors [59] have provided templates for quantifying believability.

We note that these evaluations mirror the believability characteristics discussed earlier, with a few exceptions and elaborations. The questions cover personality, emotion, social relationships, predictability (roles and player expectations), change, motivation, and the illusion of life (awareness, behavior, and coherence). However, we dismiss visual impact as an evaluation criteria; our earlier examples demonstrated that we could perceive simple elements as characters without a strong visual or a feature-based indicator, given of course its role and coherence with the game-world.

We caution evaluators to think as designers when evaluating their systems. We should define what we mean when we measure believability. As suggested by [93], believability should not be measured by statistical or predictive data but rather by works such as [150] that focus on believable characteristics. Furthermore, as designers, engineers, and creators, we should focus on how our system meets specific standards regarding believability. Does it cause player enjoyment? Does the character

lose believability aspects—and if so, when and how? How can we better support our character's believability?

5 Conclusion

In conclusion, this chapter defined what constitutes a believable character and how one evaluates believability. Through a literature review, we determined that change is an underdeveloped and often underrepresented topic in the realm of autonomous and believable agents. We also uncovered common implementation methodologies of values and beliefs at a high level.

As we will discuss in upcoming chapters, we believe that an opportunity exists to extend change via character values and beliefs and, in turn, affect character morality.

Our next chapter examines morality further, covering how moral values often appear and how morality is operationalized in video games. In later chapters, we present our system Argument Box (AB), which incorporates moral values and beliefs at a system level in relation to our current topic and portrays change via character opinions.

Chapter Three: Character Morality

Please note: This chapter is not intended as a complete literature review of morality or moral concepts. We will, however, review morality as it pertains to selected games and academic systems due to their relevance in our system designs and this dissertation.

This chapter will first give a preliminary definition of morality within our writing. We will then establish a framework highlighting typical implementations of moral systems within commercial games, mainly that of reputation scales and state machines. Next, we will present a series of case studies of well-known games defined by our framework. Through our case studies, we will review how character morality appears, how it is operationalized and portrayed, and present our key findings. Following that, we will discuss morality in academic systems and games, covering additional bases. Please note that **Chapter Nine**, which covers additional systems, dives deeper into morality by reframing it through a faction lens.

While we acknowledge that moral games (and their characters) are well-designed, we believe the overall structure of moral systems and characters could be improved. First, we argue that typical implementations (reputation scales or state machines) eventually lead the system to predominantly limited states (e.g., good, neutral, and evil). In contrast, morality is complicated, often incorporating many shades of gray. Secondly, through our analysis, we argue that there is an apparent discrepancy between player and NPC affordances in morality-based games. NPCs

are usually reflective (mirroring) and scripted, lacking in moral reasoning capabilities; this is especially true for background characters.

We believe values and beliefs can improve NPCs in moral games, especially when applied to background characters. However, to our knowledge, character values and beliefs (Chapters Two and Four) represent an underexplored topic for moral characters, especially when viewed through a believability lens. We hope to learn more about value-based design and its implications through this dissertation. **Chapter Four** covers values in depth, while this chapter illustrates lessons learned from typical moral systems and games.

1. Moral systems in Commercial Games

Before we begin, we should define what morality means in our dissertation. Morality as a concept pertains to a wide array of topics (e.g., moral theories, philosophies, and ethics). Here, we limit morality to video games that highlight moral actions via character interactions or opinions and portray their consequences in some form to the player (i.e., games that depict good and evil to differing degrees).

In examining and reviewing morality-based games, we run across many games with seemingly different moral systems. Through our review of popular morality-based games, we noticed that moral systems are often structured based on an accumulation of points on a linear (or multidimensional) scale, based on a state machine, or a combination of both.

Let us first review reputation scales. A reputation scale counts "morality" points based on moral and immoral actions, mainly through the player character. These reputation scales can be:

- **Global reputation scales.** Global reputation scales highlight the player's accumulated points and reflect them to the player (usually) in the form of a bar or numerical values. Most player interactions in the game's world count toward the player's overall moral standing. Examples include *Fallout 4's* Karma scale and *Red Dead Redemption's* honorable scale [18, 126].
- **Local reputation scales**. Local scales function similarly to global scales, except, as the name suggests, the reputation points are directed at a specific character or faction. We note that multiple instances of local scales are relatively common in RPG-based games. These include character friendship/opinion scales in *Dragon Age Origins,* specific faction scales in Fallout 4 and follower reputations in *Cult of The Lamb* [21, 18, 92].
- **A combination of local and global scales**. Games may use an overarching global moral scale and local character/faction scale at the same time. The overarching scale counts global player actions, while the local scale highlights a specific relationship between an entity and the player. The entity in this case can be an NPC or a smaller scale local to a faction or location. From our review, most morality-based games utilize smaller scales as a friendship/ romance scale in the form of character opinions (usually depicting gray areas of morality). Examples include *Fable's* [85] fame/infamy scale and local

reputation scales with town NPCs (e.g., romancing town NPC), *World of Warcraft*'s [23] faction alliance or horde scales and character-specific or micro faction scales (e.g., allied factions or special character quests) and *Midnight Sun'*s global dark/light scale and local hero/team member friendship scale.

We note that reputation points can be weighted differently depending on specific character or game-world interactions. We also note that some scales embody multiple dimensions rather than a single linear scale. In most cases, the added dimensionality counts toward customization, character abilities, or added narrative flavor. Instances include *Fable* [85], which has a purity scale on top of the game's global fame/infamy scale that adds cosmetic changers.

Next we review state machines or choice-based (or branching) structures. In a state machine or a branching structure, player actions (choices) place the player in a particular game state. These states usually reflect the player's moral actions as responses in a narrative branch or alter the game state in some form (e.g., placing the player in a particular path on a branching structure, unlocking modes or alternate game scenes). These state machines can be based on:

- **States based on gameplay style**, whereby each core option reflects a particular playthrough or state within the overarching game. These states are commonly referred to as playstyles or runs. We note that our focus here is on morality based games; as such, playstyles like speedrunning are dismissed from this category. Examples include *Undertale's* [148] pacifist, genocide,

or neutral run, *Infamous's* [140] hero or villian runs, *Injustice 2*'s [110] chosen side runs (Batman (hero) or Superman (villian)).

- **A set of embedded state machines within larger state machines** (or branching structures.) These embedded states usually depict the player's current moral standing in a localized environment (e.g., a small scene, event, quest, or chapter). The results of the embedded states usually feed into a larger state machine within the overarching game. Due to authoring and technical limitations, embedded states can alter the state of the game in a limited number of ways; a collection of smaller states often serves as flavor or serves a particular character or event resolution. Examples include choices in *Life Is Strange*, *Telltale Series*, *Dragon Age: Origin*, *Beyond Two Souls* and *Until Dawn* [44, 147, 21, 124, 144]. We note that these embedded states can structurally appear as part of:
    - **A series of continuous cinematic scenes.** These types of states often focus on smaller character interactions portrayed to the player via scenes. The scenes literally have a cinematic quality to them, often panning and focusing on the characters involved while giving the player an opportunity to participate in a (moral) choice.
    - **Interactive scenes within a larger game world.** These states often come as part of an unfolding narrative tree based on the player's request (unfolding after the player interacts with an NPC). These states may or may not be a conditional element required in the game

and they may or may not affect the game's resolution, often used to add flavor.

- **A combination of reputation scales and state machines.** These can be seen as paths on a branching structure or content on a branch unlocked as a result of accumulated reputation points. A combination of player options or reputation scales leads to the player's defined moral state. Examples include Fable 2 and Fallout 4 [85, 18]. For example, Fallout 4's companion disapproval of player actions or dialogue choices act separately (for some companions) from faction scales, or a certain location permanently altered in *Fable* (state changed), based on the player's current reputation.

Please note that narrative branching structures are beyond the scope of our dissertation; nevertheless, we acknowledge that differing branching structures can exist within a moral system and be utilized to reflect a particular state. These structures include but are not limited to linear, branching, hubs, fallback, and waterfall, among many others [171]. We also note that other moral systems exist (such as linear story or PCG-based games) due to evident moral themes or situations; however, this chapter focuses on the most common implementation of moral systems in games that emphasize moral consequences based on player or NPC actions. Chapter Ten further examines additional game types through a faction lens, noting how morality and moral situations appear within their systems.

Our next section presents an analysis of established morality-based games within our framework as case studies. We argue that each selection features a unique

addition to the moral system or utilizes a notable morality-based mechanic embedded within its design. The chosen games for our case studies are selected based on their popularity[17] and our own familiarity as a player. We note that each of these games is designated as a representative example of similarly structured games. For instance, *TellTale* and the *Life is Strange* series [147, 44] both employ a state-based morality structure; thus, our case studies highlight the former as a representative example of state-based structures in moral games. Through our case studies, we will examine how moral situations appear, who acts in them, and how they are operationalized and communicated to the player.

2. Case Studies in Morality

In this section, we will review representative examples of reputation and state-based games, one of the most common implementations of moral systems today. Through these case studies, we hope to gain insight into how morality is operationalized and presents itself across characters. We note that each subsection summarizes our findings first then follows up with a discussion. After this section, we will summarize our findings across all of these cases, coming to a broader conclusion.

[17] Popularity is determined based on awards won by the game (e.g., Game of the Year awards), reviews or discussion in academia or industry, or popular reception.

Please note that each case study discussed here may operate based on multiple moral systems; however, our analysis focuses on the system's primary implementation or unique features. For instance, *Fable 2* [85] keeps track of the player's state to determine the outcome of a specific region; it does not utilize state tracking as its primary focus. Therefore, we focus on the overarching global reputation system in its discussion.

2.1 Reputation-focused Games

2.1.1 A look into Fable 2

Fable 2 [85] presents a standard example of a global reputation system. Through our discussion, we can see that what is unique about *Fable* is its vast player action space, where every action in the game (via weights) affects the player's moral standing. While the game is enjoyable, its moral system is essentially reduced to good and evil states. While it's possible for the player to reach either extreme, it is possible to undo the player's actions, making consequences essentially meaningless. Furthermore, background NPCs act as reflective agents of the player's morality, lacking many of the criteria that make them believable. Here we present our study of *Fable*'s moral system.

Fable 2 [85] is an RPG game in which the player takes the role of a young champion fated to fight a corrupted ruler. What separates *Fable* from other games is its expansive moral system. This moral system manifests itself in multiple ways. One method is relatively common in games of this genre: the *fame/infamy alignment bar*.

As the player travels the world, they can take many actions that are either good or bad; each action taken is then reflected in the player's alignment bar. These actions, however, are weighted differently depending on the type of action the player performed.

For instance, delivering a secret love letter to a character in an early-game quest awards the player a positive point, but giving the letter to the character's mother gets a negative point. While this quest awards one point, other early-game quests have higher ramifications. For instance, we can see a five-point discrepancy when the player gets tasked with collecting arrest warrants. Depending on whom the player chooses to hand in the arrest warrants to (a thug or a guard), the game awards the player with possible +/- five points.

Morality in the above scenario appears as an explicitly authored moral dilemma, whereby the player is tasked with two choices, good or bad; these moral scenarios are judged based on the authored scenario, which presents a finite set of options. These scenarios can range from happenstance (e.g., stumbling upon unlawful activities) to circumstantial (e.g., based on an optional/linked quest), or event scenarios (main quests and cutscene options).

Besides authored scenarios, the game also considers the player's actions within the game world. For example, attacking innocent people awards the player with negative points, while donating or giving money to the poor awards the player with positive points. The game increases the possible (good or bad) action space when compared to other games of the time. For instance, the game can award points in a

variety of ways, including social interactions (e.g., praising or punishing the player's dog), transactional interactions (e.g., donating money or stealing), gluttonous actions (eating organic food vs. meat), societal actions (raising and lowering rent/shops), and attacking/defending actions, among others. While the player can take many (good/evil) actions, the player can undo them just as quickly, making the player lose any real consequences for their actions. Furthermore, this encourages the player to game the system, which will eventually help reduce the player's immersion and the believability of the game world.

Now that we have covered how the game counts for and presents moral actions, we can discuss how it communicates these actions to the player. One method is to show the actual numerical points to the player as a floating number accompanied by a small good/evil icon. Another unique way is by physiologically reflecting the accumulated points in the character, depending on the players' good/pure and corrupt/evil bar. The game highlights the player's alignment by changing the player's eyes, skin tone, complexion, and attachments. Figure 6 highlights these elements.

Figure 6 highlights the player's physiological changes based on moral alignments[18].

Morality in *Fable* is primarily player-focused; NPCs are autonomous but lack moral reasoning capabilities. They are mainly reactive to the player based on their moral alignment and actions. NPCs seem to be either authored characters with explicit moral designs (e.g., thugs committing a robbery), part of a main cast (authored for the story), or general characters, living their lives, simply reacting to the player character. General NPCs have a secondary personal relationship scale (local reputation) encoded with the player. They may run away from an evil player or find a corrupt player unattractive. In other words, the player may improve their relationship with a local NPC based on a relationship model.

2.1.2 A look into Red Dead Redemption 2

Generally, *Red Dead Redemption 2 (RDR2)* [126] functions similarly to *Fable* regarding its global reputation scale and the player's action space. What differentiates

[18] Images retrieved from gameplay video and online posts found:.

RDR2 is the player's reflected growth through minute dialogue changes and its superior background NPC design. We note that dialogue changes appear through character interactions as the player journeys in the world or through specific events. Furthermore, NPCs (even background ones) are perceived to judge the player's actions, adding depth to their character. Here we extend our discussion on *Red Dead RDR2*'s moral design:

In *RDR2* [68], the player plays the role of Arthur Morgan, a Van Der Linde gang member, as part of the group of outlaws and misfits in an open Western-themed world. Like *Fable*, this game uses a fame/infamy reputation system themed on honor. Honorable actions such as donating cash, helping the gang out with chores, and surrendering peacefully to law enforcement award the player with honor; other actions such as stealing horses, killing farm animals or humans, robbing shops, and wasting (e.g., killing an animal without skinning it) subtracts honor from the player.

While the player can partake in the listed actions above at any time, like other games, RDR2 offers the player cutscenes and gameplay that is part of the main narrative. In the process of these choices, the player is forced to make a choice that affects their honor (global reputation scale). For instance, on an early game mission, the player confronts a member of the O'Driscoll gang, the player gang's archnemesis. As the cut scene plays out, the player can kill the O'Driscoll member, set him free, or beat him up. The game highlights the moral dilemma by having the NPC plead, appearing to reason about his actions, which led to the following scene playing. The scene starts with a confrontation in an authored scenario in which the O'Driscoll

member iterates through a series of begs, such as: "*Please, spare me." "I promise you, you won't see me again, partner." "Just let me go, come on.*" The player can interrupt the NPC by taking one of three actions: choke, let the NPC go, or knock the NPC out. At this point, it highlights moral actions at three levels: evil, good, or somewhat evil. Scenarios like these add flavor to the text but have minimal impact on the story. However, what is unique about the game is how it establishes the player character's moral growth through the accumulated honor points; in increasing or decreasing the honor, Arthur's speech patterns change to reflect his honor standing. Unlike most games, the text is not only reflected in the main story scenes but as minute changes throughout the game's world (e.g., interacting with NPCs). For instance, notice how these speech patterns change when Arthur interacts with different NPCs in low and high honor standings.

Here we highlight sample speech patterns[19] depending on the player's moral standing.

Apologizing to NPCs (bumping into NPCs):

High honor:

Arthur: *Careful there!*

[19]Gameplay footage analyzed and retrieved from: https://www.youtube.com/watch?v=tWzN6g8qHWg and https://www.youtube.com/watch?v=5uysvcGKcOk.

Arthur: *Excuse me.*

Arthur: *Mind yourself, partner.*

Arthur: *Sorry, mister.*

Low honor:

Arthur: *Just what I need!*

Arthur: *Stay out of my way.*

Arthur: *Open your eyes, fool.*

In a brawl situation:

High honor:

Arthur: *At least try fighting back.*

Arthur :*(chuckles) Look at you, you're a joke.*

Low honor:

Arthur: *You are dead.*

Arthur to the NPC: *You are pathetic.*

In an interaction with Sheriff Maloy :

High honor:

Arthur: *You just look so goddamn tough. It's frightening.*

Maloy: *You got any idea who you're flapping your gums at?*

Arthur: *I mean, do you really wanna test me?*

(player takes action or leaves to advance)

Low honor:

Maloy: *No use playing innocent around here; folks remember you.*

Arthur: *Just give it a rest, will you? What's done is done.*

Maloy: *I don't think you understand who I am, boy.*

Arthur: *You are pathetic, my friend.*

Maloy: *I wanna see you making tracks. Let's go.*

Arthur: *Ok, ok. I'll be on my way.*

(player takes action or leaves to advance)

The changes in dialogue highlight elements in Arthur's development, reflecting his moral standing. The text is accompanied by tonal changes that convey Arthur's cold nature in low-honor and humor in high-honor states. The reported examples lead to the exact unfolding scenario when acted upon, albeit with minor changes; the results remain the same. We noticed that moral actions changed the character's *approach* to a given situation rather than the result. For instance (spoiler alert!), Arthur's story ends with his death at the end of the game; how he dies is determined by his honor state. In high honor, Arthur succumbs to his disease and dies. In low honor, Micah, a major character and antagonist, kills Arthur.

The game reflects the player's moral standing by showing the mentioned dialogue changes and increasing the player's bounty. Additionally, honor (reputation) is reflected in other game elements. For instance, high honor causes a price drop in purchased goods and clothes, unlocking additional quests. As is the case with *Fable*, good actions award the player positive honor points, while evil actions reduce the player's total honor.

The game also involves NPCs in unraveling scenarios that show up depending on the player's honor and actions. For instance, if the player goes on a killing spree, killing random NPCs, they will eventually run into a woman covered in black. The mystery woman cries out, "You bastard, you killed my husband… you know that?" "you probably don't even remember it," "meant so little to you. And now, I am alone with nothing at all." Not giving the player a choice, Arthur replies, "Honestly, I've killed a lot of people, and in my mind, your husband don't stand out," to which the woman replies, "you evil monster!"

NPCs in this game are autonomous and highlight the player's overall standing via minute dialogue changes. The NPCs are also reactive to other NPCs/entities and the world around them. The main characters are heavily authored and feel rich with vibrant personalities, though with set (authored) morals. Interestingly, the minor characters in this game (such as vendor NPCs) are believable and reactive to the world around them. One case is that of the Valentine butcher who in summary, calls out the player for abandoning his shop, appearing to judge the player. **Chapter Two** includes a brief case study of the Valentine butcher, which highlights the NPCs' dynamism in this game. To avoid repetition, please refer to **Chapter Two's** case studies on minor characters and roles.

While most NPCs are reactive, we noticed that not all meet these standards. Other NPCs may react in fleeing, for instance, as a reflection of the player’s evil moral state. Nonetheless, *RDR2* presents a significant leap in designing background NPCs.

2.1.3 World of Warcraft and Oblivion.

Here we additionally examine *World of Warcraft* and *Oblivion* [23, 16]. Through this section and earlier examinations, we note that 1) background NPCs commonly reflect the player's moral stance, 2) they are autonomous and often reactive to the player, not so much on other NPCs, 3) NPCs do not have the same affordances as the player, and 4) accumulated morality points can be reversed, resulting in reduced consequences.[20] Let us next examine *World of Warcraft* and *Oblivion* as reputation-based games.

Reputation scales are one of the most common designs utilized in moral systems. While reputation is common in single-player RPGs, it is also widely implemented in Massively Multiplayer Online Role-Player Games (MMORPGs). One of the biggest games employing reputation is *World of Warcraft* (WOW), an MMORPG in which players belong to one of two opposing factions: the Horde or the Alliance. Thematically, the Horde contains a group of monsters and invaders (e.g., orcs, undead, goblins, and blood elves), while the Alliance contains groups of aesthetically pleasing characters (i.e., elves, dwarfs, humans, and gnomes), reflecting the game's “good” and “evil” states. Reputation in WOW acts as a currency, where

[20] (We note that some games, like Fable and RDR2, provide both state and reputation scales within their games. However, it's most commonly used to secure or unlock specific endings or regions, as we will soon discuss in state-based examples.)

players grind and increase their standing within a faction; it rewards the player with cosmetics or items or increases their standing with a faction.

The overall structure of the system feels binary, where the players and NPCs belong entirely to one particular faction. There are sub-factions in the game; however, they either adhere to the game's two main factions (examples include *allied factions* and *reputation* factions) or as neutral factions, serving to advance a particular storyline. In recent years, the game has allowed a cross-faction feature, allowing players to queue for dungeons or raids with friends. We note, however, that it does not affect the game world itself or race affordances (e.g., horde territory is still violent to an alliance character, as far as we can tell).

WOW is focused on the player's actions. Players take on a key role, whereas NPCs take on a reactive role, having little to no interaction with other NPCs. For example, a WOW quest giver NPC reacts to the player but pays little to no attention to other NPCs within the same faction and vicinity.

Next, let’s look at the *Elder Scrolls IV Oblivion* [16] an open-world fantasy RPG in which the player travels the world on quest-based adventures interacting with NPCs and the world along the way. Researchers Roma et al. [27] analyzed *Oblivion*'s morality system. According to [27], *Oblivion*'s reputation system consists of a fame and infamy scale, marking where the player stands in terms of good or bad (similar to *Fable*'s reputation system). While *Oblivion* is focused on the player as an active agent, it differs from *Fable* in its agent's autonomous capabilities. NPCs in *Oblivion* have the *responsibility* attribute that provides NPCs with a personal scale that denotes

their opinions toward the game's governing laws [27]. This responsibility attribute influences an NPC's behavior, considering their situation and how they feel about the laws to complete an action. For example, an NPC may steal out of hunger if they have a low *responsibility* trait, a clear departure from other games of the same genre.

2.3 State-based Morality Systems

Now that we have covered examples of reputation-based systems, let us next focus on state-based moral systems, covering how morality presents itself, communicates it to the player, and how NPCs act. We will then conclude this section with a system that conveys both state and reputation in its design.

2.3.1 A look into TellTell series.

Telltale games are a series of games that focus on branching episodic storytelling. *Telltale* has developed many stories based on popular IP, such as Batman, The Walking Dead, and the Wolf Among Us [147, 145, 146] *Telltale* redefined this genre for modern audiences.

By reviewing these cinematic state-based systems, we typically find that moral calls come in the form of player choices. Once again, NPCs reflect the player's current moral stance. Usually, these come as NPC approval/disapproval or character resolutions, as part of an unfolding story. Cinematic state-based systems, unlike reputation-based systems, lack autonomous background NPCs. Instead, NPCs appear as part of a reduced cast in a scenario. The presented NPCs usually rely on heavy

authoring and are highly expressive, focusing on emotional expression and physiological changes.

Here we further discuss *Batman: The Telltale Series* [84] as an example of a cinematic state-based game. We note that other games of this genre typically follow the same structure.

In *Batman the Telltale* series, the player takes on the role of Bruce Wayne (Batman) through an unraveling mystery story. The game functions primarily by giving players options to choose from each time in place of a reputation bar. Different cutscenes present the player with different types of narrative-based choices. Some scenes and player choices are informative, giving the player additional context, and some are moral, reflecting the player's moral standing, and some are crucial (often in the form of a moral dilemma) for the developing story. The game also employs a quick-choice structure in which players are pressured to submit an option within a time limit.

The player's character's (Batman) personality is constant. However, the player's methodology changes, depending on the player's moral code (not necessarily Batman's). For instance, in an early scene, the player fights an NPC, at which time Batman, unprompted, states, "Don't make me add your corpse to this graveyard," giving off a cold menacing vibe. The player is then met with a set of timed options: "You did not do this alone," an option to punch the NPC, an option to slam the NPC into the wall, and an option that states, "Talk, and I will spare you." All four options are thematically part of Batman's persona; however, some actions are less harsh than

others, especially when prompted by the earlier statement. The player can shape their Batman/Bruce Wayne via their morality, but the general character remains the same. Figure 7 shows the scenario depicted earlier.

Figure 7: The player responding in conversation with a character.

Of the game's many gameplay options, a few crucial ones (main options) carry a heavier weight on the story; these options also affect the NPCs' development in the game's next chapters. The game communicates these options to the player at the end of each chapter by informing them of their choice as a statistic compared to other players. The game also highlights crucial choices by stating that an NPC will "remember" the player's choice. These statements, while not indicative of true NPC awareness, create the illusion and perception that the NPC is aware and judges the player's actions. NPCs here are authored for their unfolding scenarios with a set of predetermined paths based on a collection of met conditions. The NPC's state is further reflected at the end of the game via their judgment of the player's action (i.e., narrative choices) and possible character resolutions. Figure 8 shows the NPC summary at the end of the game. The player character's morality is a reflection of the player's morality but it is limited by the set of options provided by the developers.

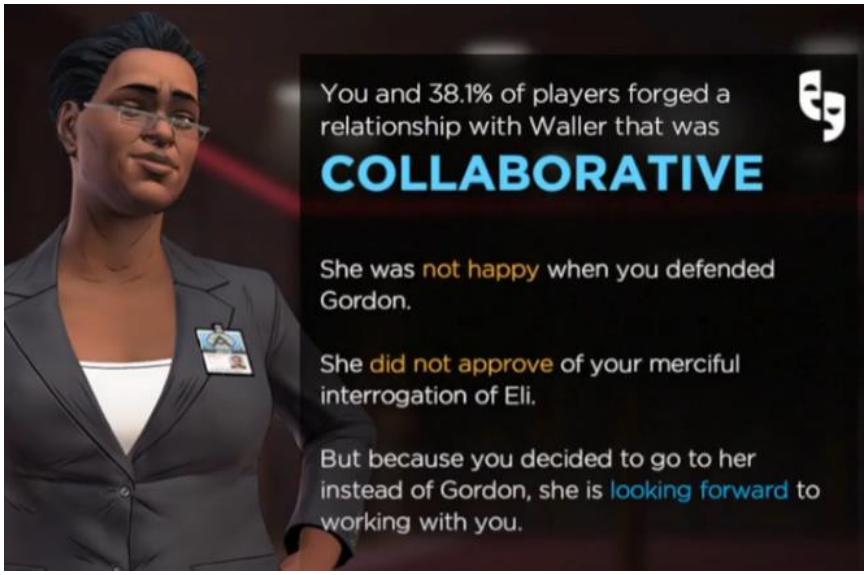

Figure 8 highlights NPC Waller's resolution in Batman: The Telltale Series.

2.3.2 Morality through inaction. A look at Undertale.

One interesting display of state-based morality is in the form of playstyle (or player runs). Unlike the *Batman* example, choices here are predominantly limited yet shape the player's whole gaming experience. What's unique about *Undertale* is it takes inaction as a playstyle (that of pacifism). While the game showcases some dialogue choices, the main states are reflected based on the player's commitment to one of three possible gameplay style choices. In this case, the player's style/choices are as follows: *Kill **some** NPC, Kill **all** NPCs,* or *Do **not** kill **any** NPCs*. Once the player commits to a playstyle, the game proceeds to unfold narratively along one of its core states.

Let us now examine *Undertale* [148] in more detail. In *Undertale*, the player plays a fallen human in a strange underworld. As the game starts, the player's first

encounter is meeting Flowey, a seemingly sweet flower. Flowey's interaction teaches the player about the world. In essence, the flower talks about leveling up and getting stronger, which are common elements in most games. Figure 9 shows the first interaction. The Flower then urges the player to collect these little pixels, informing the player that collecting them strengthens the player.

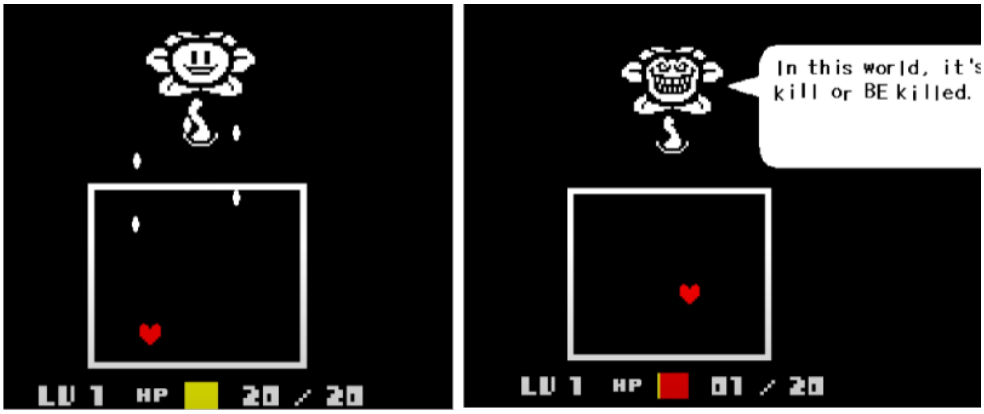

Figure 9 Depicts the player’s first encounter with Flowey.

The player is then shocked to learn that colliding with the pixel lowers their HP; the player is then informed about the truth of this world from Flowey's perspective. Figure 9 shows Flower's surprising transformation from a helpful flower to a wicked monster. The game's combat element is further hinted at in the instructions level before the game starts, when the player is warned that, "When HP is 0, you lose."

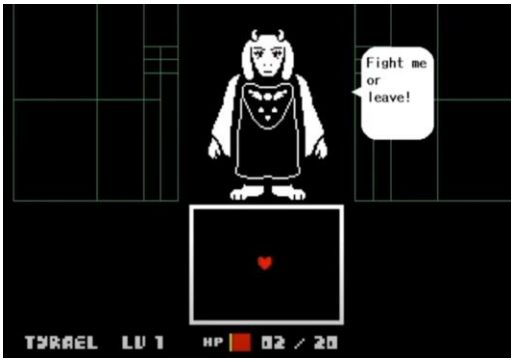

Figure 10: The Caretaker NPC instructing the player on the game's primary interactions.

During Flowey's interaction, the player is saved by a friendly caretaker who takes the player's hand and leads them through the ruins, instructing them about the world. Through the caretaker's instructions, the player learns about the mechanics and puzzles of the world. After the player learns about the game, they are again met by the caretaker, who stands in their way, instructing them to fight her in order to progress. Figure 10 shows the caretaker's instructions. If the player is not familiar with the game, they would assume they need to fight the caretaker or flee as their only options, as instructed by the NPC herself. If the player attacks, the scenario ends, and the player goes ahead. However, if the player persists in the game and ignores the NPC's commands, they learn that another route—that of pacifism—exists in this game. Thus the game informed the player of two play styles. One that harms NPCs and one that does not.

As the game progresses, the player meets various NPCs and has to "win" against them to progress. Almost all encounters with enemies start with combat, in which the enemy attacks the player. These encounters are turn-based, between the

NPC attacking and the player responding from a set of possible actions. During the NPC's turn, the player must dodge incoming projectiles to avoid losing health points and possibly their life.

Mechanically, the player has a set of four actions during combat: *fight*, *act*, *item*, and *mercy*. Fighting and defeating enemies awards players with experience points, making the player stronger. Acting gives each monster encounter a unique scenario, such as telling jokes, complementing, and checking a monster's status. *Item* allows the player to use any items in the game, and mercy enables the player to flee the encounter or spare the NPC (this may not work, depending on previous actions). Fleeing ends the encounter but does not progress the game where sparing is triggered, depending on the scenario's conditions (these can be a combination of *act* and avoiding enemy projectiles). What's noteworthy about *Undertale* is that its moral state relies on the player's actions or inaction as it goes, taking pacifism as the optimal path to the game's best ending/state.

The game's moral system communicates the results of the player's encounter via NPCs and the world itself. If the player uses combat and kills often (one of the first things the game hints at), the world becomes lonely and empty, without life. On the other hand, if the player takes the optimal route, the game becomes more challenging, but the world is lively.

NPCs in this game are authored; each NPC has different textual information depending on the results of an encounter/state. Some special NPCs with recurring

roles, such as Sans, reflect the player's current choices (playstyle), hinting at their moral standing.

2.4 SmallScales and Game-ending States, a Look into Dragon Age: Origins

Lastly, we will use *Dragon Age: Origins* [76] as an example game that uses a collection of states and local reputation scales as part of its moral systems. Local reputation scales add more diversity to the moral system, whereby morality is reflected through the player's companions. These companion NPCs are part of a predetermined cast of highly authored characters. The smaller reputation scales add depth and narrative flavor to the game's main states, sometimes holding the power to unlock a specific choice or state.

Other NPCs (background and enemy NPCs) are role-based and somewhat reactive to the player. Unless specifically authored otherwise, these NPCs play little role or consequence in the overarching story and moral dilemmas. Let us now examine *Dragon Age* further.

Dragon Age: Origin is the first installment of a fantasy RPG game in medieval fantasy. The player is part of the Grey Wardens, a group of duty-bound warriors protecting the realm from Darkspawn, an evil-tainted race.

As the game progresses, the player gathers companions from different races, each with a unique backstory. Companions can join the player based on the player's choice, as an in-game prize, or as part of the main narrative. Like *Telltale* games, the

game's morality often comes in the form of a choice the player takes in the world; choices can be physical actions (e.g., freeing an inmate) or dialogue-based.

The companion's opinions and statements reflect the game's moral system. Almost all significant choices in the game come with the approval of a character and the disapproval of others. Unlike the global fame/infamy scales mentioned earlier, *Origin*'s morality scales are local, reflected via the companion's approval/love (as some characters can be romanced) and disapproval/hate states. The design of these companion characters, each with their own opinions, diversifies the perceived morality, creating shades of gray instead of black and white. We note that these are limited to a set of selected characters because, as we can imagine, this demands a heavy authoring burden.

Some characters in these games come from a specific race or faction. Narratively, these characters represent their race or faction in their judgment calls; let's call these representative characters, as we will discuss them in future chapters. Unfortunately, representative characters in this game have little to no influence over (or from) the factions they are assumed to belong to (where applicable). Let's illustrate this point by taking Wynne, a mage from the magic circle faction, as an example.

While in the magic circle's tower (Wynne's faction location), we can see that the player's choices impact Wynne's approval rating. The player's choices can result in Wynne leaving the player's party, attacking, or staying with the player, but the magic circle's resolution fundamentally depends on the player's choices. Having a

character's local scale change but not influencing the magic circle's faction causes a disconnect; it diminishes the weight of a faction and its members. On the other hand, other games, like *Fallout 4* [69], link the representative characters' approval to that of the faction, confirming the character's role within a faction. While linking characters resolves the disconnect, faction members still seem shallow and irrelevant. We believe these faction areas can improve. However, we reserve this discussion for the coming chapters, particularly **Chapter Nine**, as it frames morality through a faction lens.

Another unusual feature of this game is carrying the choices in one game into the game's sequential series of the same name. While the player does not play the same character, the world still changes, reflecting the player's morality.

3 Findings on Morality

3.1 Summarized Findings

In this section, we summarize our findings and extrapolate additional features based on our case studies of representative examples.

Moral situations appear to the player as:

- Absolute authored game scenarios and cut scenes where the player is forced to choose between gameplay options.
- Happenstance game scenarios in which the player stumbles upon a moral dilemma.

- Optional inquired-upon scenarios, in which the player actively seeks out NPC opinions (e.g., talking to companion characters in *Fallout* or *Dragon Age* [69], [76], unprompted).
- Circumstantial, where the player faces a moral dilemma only after certain conditions have been met.

We note that all of these scenarios can be dialogue-based or action-based (whereby the player takes an action in the game world).

Morality and consequences are communicated and reflected to the player as:

- Graphic changes on the game's UI. These can be a visual change in an alignment bar, change in color, numerical display change, or visual icon pop-ups.
- Physiological character changes; these include scars and cosmetic changes on the player or NPC characters.
- Dialogue changes and employing different speech patterns to a character based on a moral state.
- Auditory changes and notifications such as a character sounding cold while in an immoral state.
- NPCs updating or reflecting the current moral state of the player-character.
- Game world changes; these include changes to the atmosphere of the world (e.g., empty towns on *Undertale*'s [148] genocide route), zone changes (restrict

or unlock access to zones in Dragon Age [21]) and character changes (deaths or abandonments of NPC characters due to the player's current state).

Lastly, let's compare the role of a player character vs. NPCs and how they communicate and handle moral dilemmas.

	Player	NPC
Gameplay	• Active agent	• Reactive agent Scripted agent
Morality can be	• Choice-based • Action-based	• Authored (absolute) mortal persona and opinions
Actions	• Acts upon NPCs and the world • Reacts to NPCs, story, and the world	• Enacted upon by the player • NPCs (mostly) limited to reactive actions, with the exception of some games like *Oblivion* • Specific autonomous actions (e.g., fleeing)
Affordances and agency	• Expansive. In most cases, th player is able to act upon element in the game world.	• Limited. In most cases, NPCs are unable to act upon elements in the game world (with exceptions like some NPCs in *RDR2*'s [126] camp).

	Player	NPC
Gameplay	• Active agent	• Reactive agent Scripted agent
Moral reasoning	• Based on the player's judgmer and perspective.	• Superficial judgment in most cases (e.g., NPCs fleeing in *Fable* from the player [85]). • Authored character and moral standing.
In a moral system, the agent is used to	• Progress through the story • Commit gameplay actions and choices	• Reflect the player's choices. • Provide an opposing/ supporting opinion. • Contextualize a moral dilemma.

Table 3 : Summary of key findings and the differences between NPCs and the player.

3.2 On Good and Evil.

Unsurprisingly, many games employ a reputation bar, which accumulates positive and negative points to reflect the player's moral state in the world. One issue with these designs is they typically make the system predominantly binary, resulting in a "good" or "evil" state. We argue that morality is anything but binary; morality is often a complicated notion that includes many perspectives and shades of gray.

While player morality in many games accounts for many choices and perspectives, an NPC's moral structure is often rudimentary or highly authored, in both cases backed by a simple system. In some cases NPCs seem unable to distinguish between actions, commit to a singular ideal, and exert their authored beliefs upon the

player. Their primary role is reflecting the player's moral state rather than actively participating. We realize that at the end of the day, a character is rich and believable if they are perceived as such. However, we think NPCs, especially the cast used to fill the world (background NPCs), could improve if we improve their moral reasoning capabilities. Our case studies also revealed that some NPCs conveyed diverse moral standings to the player, often expressed through NPC opinions (about the player's actions). Unfortunately, most of the reviewed NPCs are highly authored, restricted as part of a specific cast (e.g., companion NPCs), or performed as part of a cinematic scene, often limiting background NPC interactions.

We believe a gap exists in the literature, whereby we can diversify the player's perceived morality through NPC values and beliefs. We believe that values and beliefs are an understudied area of research, especially when viewed under a believability and morality lens, as we will discuss in upcoming chapters.

We also discovered a clear discrepancy between player and NPC affordances. Most games are focused on the player's actions. Players take on a key role, whereas NPCs take on a reactive role, having little to no interactions with other NPCs or the world. For ease of communication in future chapters, we will call this *player-centric* design.

We believe that the lack of NPC-led (NPC-centric) interactions in open-world games detracts from the overall believability of the game. It creates characters that are mere copies of each other, with seemingly simple goals and little to no personality traits, which, as we stated in Chapter Two, are essential characteristics of believability.

While most moral factions (as we will discuss in **Chapter Nine)** and reputation-based games focus on the player's perspective, there exists a gap in how NPCs react and act to opposing ideals and factions. These are further highlighted in games with two opposing slides such as *WOW, Far Cry Primal*, and *Assassins Creed* [23, 156, 154], in which NPCs often attack members of opposing factions on sight without reason. As we will suggest in our next chapter, by incorporating moral reasoning through beliefs and values, we can enhance how NPCs react to one another and add depth to NPC conversations and actions. More importantly, we can move away from pure moral binary systems as well as linear reputation systems that exist in games today.

Lastly, we note that many moral systems discussed here, framed from a faction perspective (**Chapter Nine**), lack fundamental moral elements compared to their counterparts in other media forms. **Chapter Nine** further analyzes morality under the lens of faction systems, providing fruitful directions and future work that could help improve overall character believability.

Let us next examine morality through academic systems (please note that **Chapter Four** examines values and continues the discussion in depth).

4 Morality and Moral Reasoning in Academic Research

We remind our dear reader that we do not review all areas of morality, as morality is an extensive field, nor do we consider non-systemic approaches in serious games and interactive fiction for scope reasons, instead we highlight representative

examples from academic system implementations of moral theories, reputation systems and agent-based architectures.

4.1 Moral theories and Academic Games

As we have seen, moral actions in games are generally determined by either the players' or the NPCs' discretion. Characters are often judged and judge others based on simple moral alignments. In academic systems, however, moral judgment can be implemented in diverse ways, such as cognitive or political modeling.

Academic morality systems often model particular moral theories. For example, in [149], Togelius asks if it is possible to design a game that follows Kant's categorical imperative, in which the law states, "Act only in accordance with that maxim through which you can at the same time will that it become a universal law." He [149] created a prototype in which a procedural system defines a new maxim rule each time an event triggers. Nelson [108] also implemented different prototypes for the same theory, starting with rules and breaking them, as opposed to Togelius's generating rules from scratch.

Other researchers [136] investigated an agent's ability to decide with intent, inspired by Kholberg's theory of moral development. The theory permits six levels of reasoning, such as one's ability to focus on consequences or best interest. The model is coupled with an appraisal theory that is integrated with Kholberg's six levels of reasoning to provide a more nuanced decision-making agent.

Another example is *Nelson's Moral Calculus* [109]. According to Nelson, morality in games appears in one of two forms: morality that resides within the game world or morality presented by the characters. In game-world morality, the "scorekeeper" or system acts as the one and only judge. It divides player or NPC actions into right and wrong and maintains a clear set of universal rules. On the other end, character-based morality is individualized on a character level, with morality based on the principles a character embodies. Moral situations are also present in systemic interactive fiction; for instance, Harrell's group developed a series of games and systems that deal with specific moral situations. One of these, *Chimera: Gatekeeper* [63], explores character identities within a narrative scenario, while *Greyscale* [64], another *Chimera* application, analyzes gender discrimination via character interaction.

We note, there are other systems that work based on political morality, which we will discuss in **Chapter Four.**

4.2 Reputation Systems in Academic Games.

To cover additional bases, we turn to reputation systems in academia. In academic research areas, reputation systems are often equated with recommendation or social-based reputation systems [135]. In this section, we examine a few representative systems and theories that employ reputation-based systems and morality in NPC designs.

Otello [135] is a reputation-based system used for both social network systems and games. Otello allows NPCs and players to share and rate information. Otello's

reputation system is based on a weighted social graph, within which people (NPCs) have unique reputation standing based on a collective opinion of agents within the graph. Each opinion an NPC submits is subjective, accompanied by a rating that includes a value (how trustworthy the source is) and confidence measure (how accurate the information is). The social graph that Otello constructs highlights trust and interest values. For example, a positive link in the graph shows an increase in trust and interest, while a negative link signifies a decrease in trust. One interesting application shown by the authors is using Otello as a back-channel for NPC town gossip, whereby NPCs can influence and share information based on trust values or can even influence public opinion. We can imagine how a reputation system such as this could be used as the moral basis for NPCs within faction games in which NPCs can form public opinions and attempt to persuade the town they live in (faction members).

Unlike other reputation systems, Mooney et. al.'s [104] reputation system includes explicit modeling of percepts. Percept is a tuple that describes the relationship between objects in the game-world; it accounts for the actor, the action/relationship, and the object. The given example illustrates a scenario in which the player (actor) stole (action) from a shopkeeper (object). This information is accompanied by a confidence value that accounts for how trustworthy the information is. As in Otello, NPCs can share their perceptual information with others (gossip). The truthfulness of the information changes based on the receiving agent's trust, ultimately decreasing until it is idle gossip. The authors included an agent architecture and a Skyrim-like game in which NPCs can share information in a town. Information sharing in their system is

distinguished by the use of perceptual zones. When two NPCs gossip, other NPCs with the perceptual zone can overhear or witness the events. Agents then can choose whether to share that information forward, affecting the player's reputation. The NPC model is also accompanied by memory and prediction models used to predict the player's actions, making their past actions influential.

4.3 Morality and Moral Reasoning in Agent-based Architectures

Agent architectures often involve moral aspects and can be found within the Intelligent Virtual Agent (IVA) community. This section highlights additional areas of morality and presents alternate implementations of moral systems.

One common theme that emerged from the IVA community is modeling morality through empathy within virtual agents (VA). While acting through empathy does not cover what constitutes immoral behavior (in some cases), we infer that empathy is a sentiment that leads characters to act ethically with others, as they perceive them as themselves.

Chapter Two discussed the Fear-Not project [11], which involves creating VAs for an anti-bullying project from a believability perspective. Here, we look at additional literature that used Fear-Not as an application through a moral and architectural perspective. In their other work [118, 42], the authors explain that the basic architecture for an empathetic model should abide by the agent's capacity to recognize emotions, communicate with other agents, process emotions, and have the ability to express them.

In addition to their emotional model, the agents have a personality model (based on the OCC model). These two models dictate how an agent acts or reacts to different emotions. In terms of moral behaviors, agents can appraise actions as *blameworthy* or *praiseworthy*. For instance, a bully character hitting a victim character can be appraised by another with negative connotations.

Additional work by the authors [118, 119] inquires about how a character can build empathetic relationships with users. One way is through the idea of *proximity*. Users engaging with the characters must feel close to them in order to have an empathetic relationship; this cascades into design considerations such as the environment, and the user's familiarity with the character's settings and situations plays a part in feeling related.

Another instance that highlights empathy within its design is the EMOTE project [29]. Similar to other agents within this area, agents (robots in this case) in this project rely on the areas of affect and emotion modeling. The agent can encourage users, detecting uncertainty, and praise them when they succeed. While not directly related to moral behavior, it can facilitate empathy and ethical actions—and respond to a user's emotional state.

Let's look at another article [40] highlighting morality through empathy, culture, and sociability. The author aims to have a deeper manifestation of empathy through an established framework of culture and social dynamics. To that end, the authors have a few requirements for their agents, including that agents should have a shared purpose and attention (framed as ritual), should have different action

interpretations (that change depending on the people or environments), and that agents should infer the status of other agents (by observation or interpretation).

Interestingly, these characters also utilize values as an underlying structure for moral behavior. While the authors use the concept of social norms as guidelines that a group of agents can accept as moral or immoral behavior, what's currently seen as ordinary behavior affects what is viewed as moral behavior. However, it does not affect the character's values significantly. Another interesting facet is the architecture's use of reputation. Reputation in this context is associated with what the authors call a *moral circle (MC)* (a group that the agent feels a moral obligation to), whereby members within the same MC have certain expectations from other agents within the same circle; reputation then acts within the confines of its moral circle. For instance, actions that align with the MC increase an agent's reputation, while contradictory actions harm it.

Another interesting agent architecture [35] utilizes beliefs and desires through assessing different processes, including awareness, evaluation, goodness, and rightness. Each process pertains to different elements and considerations. For instance, the goodness process considers the agent's beliefs, desires, and actions in its assessment. While the agents can use their model to judge their behavior, they can also judge other agents blindly (only taking in an agent's behavior), partially informed (with agent information), and fully informed, where the agent has complete information about the other agent, including its judgment process.

Some researchers focused on moral judgments from different perspectives and areas of thought. For instance, one piece of research [32] connected moral judgements

through concepts in neurosciences and neuropsychology, whereby the study [32] models the decision-making process by mimicking the brain function of different areas, such as the amygdala, hippocampus, and sensory cortices. This model allows the agent to integrate internal motivational and emotional states through a three-phase cycle, which includes assessment of options (evaluating potential choices), “execution” (acting on the selected decision), and “outcome evaluation” (reviewing the results to inform future decisions).

Some researchers modeled morality based on existing character data. For instance, the authors [54] used Goofus & Gallant corpus as a dataset to create labels indicating socially normative and non-normative behaviors (perceived as values aligned as good or bad based on normative behaviors). This study demonstrates how the actions of characters in these stories can train models to classify behavior accurately. The models not only perform well on the Goofus & Gallant corpus but also transfer effectively to other tasks. Their approach is a strong example for value alignment, complementing traditional techniques like learning by demonstration, preference learning, and imitation learning.

Lastly some researchers focused on evaluating the user’s perception of moral agents. One study [56], for example, explores how presenting situations in which an agent behaved ethically or unethically would lead people to attribute virtuous or vicious character to the agent. Participants received scenarios and were able to make ethical judgements based on the virtue domains defined by the study, which include justice, truth, fear, wealth, and honor. However, their findings suggest thatpeople's standards

regarding their evaluation of agents' morality are lower than their standards with regard to humans. Another study [53] attempted to define the threshold of perceived morality in agents. This study extended the definition of agents to non-human entities, such as animals, to seek a better understanding of the conditions under which an agent may be held morally accountable.

We note that other studies mentioned here also evaluate their systems based on their user's responses but from more nuanced/narrow perspectives, such as the study of empathy in virtual agents in Fear-Not [118, 119].

From our review of moral systems within the realm of academia, we noticed that researchers commonly construct moral systems based on specific moral philosophies and theories.

In agent architectures, specifically within the IVA community, morality is often viewed from an ethical perspective, highlighting elements like an agent's appropriate responses. A typical implementation across multiple systems emphasizes emotional modeling in agents.

While the nature of (most) of the cited works depicts realism-based areas (because of the nature of the artifact and questions posed), we can extrapolate key design features and information applicable to believability. We believe that by examining multiple works and depictions, one can learn and enhance avenues of believability, like the use of emotional or social aspects of agents discussed here.

5 Conclusion

This chapter provided contextual information on morality as it appears in industry and academic-based systems. We dove deeper into moral system designs in commercial games by highlighting standard methodologies of reputation scales and state machines.

By examining moral games at close range, we extrapolated key features related to morality, including how moral situations appear in games, how they are communicated, and how NPC and player roles differ within moral designs.

We argue that standard moral methodologies (state machines or reputation scales) eventually make characters feel binary, with their moral stance usually perceived as either good or evil. In the chapters below, we will argue that characters' moral values can make characters appear morally varied, with characters caring about different agendas based on their individual beliefs.

Thus, the next chapter will examine values, beliefs, and moral theories. We propose using NPC values and beliefs as a systemic reasoning device for NPC morality.

Chapter Four: Beliefs, Values, and Moral Reasoning

Acknowledgement: Some elements of this chapter have been published elsewhere [5].

Chapter Two revealed the emerging commonality of beliefs within social and autonomous system architectures. Here, we review additional systems from the believability camp that highlight beliefs in their design, provide a historical definition of beliefs and values in computational models, and explore ways of extending morality systems through beliefs and values.

1. Belief-Focused Systems

This section focuses on a set of representative systems that continues our discussion from **Chapter Two**'s survey of social and autonomous systems. These systems focus on social simulation, character believability, and agent design, while additionally highlighting values and beliefs as a main contributing factor.

Our first example is *Versu* [51], an episodic storytelling simulation whereby players can enact different characters from multiple genres. Like *Cif*-based systems[61, 98, 115], *Versu's* simulation highlights the use of social engines in its design. Versu's

simulation includes *social practices*[21] that describe actions an NPC can take during a social situation. In general, *Versu's* social capabilities[22] are outstanding.

One interesting aspect is how their system treats characters. *Versu* treats NPCs and PCs as interchangeable entities through roles. In other words, the NPC views the player as another NPC with a role. We also note that each NPC in the simulation can play different roles at the same time. For example, an NPC can play as a friend of a character and as a romantic interest of another. Versu evaluates characters based on their relationship model and how convincingly a character can play a role.

What differentiates *Versu*, in our opinion, is their model of belief, whereby agents share a public conceptual view of the world and specific instances of individualized false beliefs. For example, the characters in Versu's *ghost story* can argue about the game's weird circumstances, questioning what causes the story's bizarre incidents to occur, questioning if events were the result of a mischievous ghost,

[21] Social practices are similar [51] to social exchanges (i.e., social games in [19]), where they measure social interactions between multiple characters.

[22] *Versu*'s system can affect the social state of the characters involved in their social exchanges and use various character information at their disposal, such as character names. Their Social practices account for multiple types of practices (such as a "greet conversation" and a "dinner event" running simultaneously. Furthermore, we can separate social practices based on length, such as short and long practices (e.g., events vs. conversations).

or something scientific. Additionally, agents have “self-belief” goals (depending on the outcome of these goals), which can influence and change the arcing narrative. *Versu* is an excellent example of a believability-based system (whereby characters fulfill a combination of the believability criteria mentioned in Chapter Two) that incorporates beliefs as a core feature of its design.

Additionally, we consider *Versu* and *Façade* (discussed in Chapter Two) ideal representatives for NPC believability. *Façade* [94] presented its NPC characteristics through advancements in the plot, which is managed by a drama manager. As the drama escalates, so do the character's reactions and behaviors, matching the plot in tone, utterances, and reactions. *Versu* [51] factors in belief modeling into its decision making, whereby self-beliefs guide character goals. Furthermore, *Versu*'s system divides beliefs into specific character beliefs (false beliefs used for disagreements) and the world-belief (world state) model shared among all characters (rather than separate world views due to memory capabilities). We think these beliefs support more intriguing interactions and character development. For example, characters can have disagreements with other characters that seem legible based on their beliefs.

Next, we examine the system of Azad and Martens, *Lyra* [13]. *Lyra* presents an interesting take on modeling a character's set of beliefs. Their system simulates a town where characters interact in politically charged group discussions (at the time of our review, it was limited to schools or businesses). Their original model is theoretically based on [159], where characters can share their beliefs about a topic; each topic is informed by a private "attitude" local to the agent. Agents can express their thoughts

using "opinions" similar to attitudes; however, they are expressed publicly. Lastly, each opinion has a gauge of "uncertainty" that determines their confidence in their opinion.

Azad et al. extend their model by allowing character biases, whereby characters can learn new biases by introducing them to their knowledge bases (via first-time interactions or informed by self or others[23]). For example, the knowledge model accounts for an "object of discussion" whereby objects are associated with ratings and a source to form a topic (e.g., a computer science student uses AAAI as a valid source). Their model is mainly used for political discussions among simulated NPCs to broaden interactive narrative systems.

The authors then extended *Lyra*'s original model [12] to address a few gameplay issues. The authors addressed their player base's inability to follow an NPC's reasoning by adding a visualized political scale showcasing the character's change of opinion. The authors also ran a believability study [12], with promising results. They noted that some participants reported changes in an NPC's opinion and the NPC's uncertainty as believable behavior. Interestingly, NPC group dynamics also affected believability, whereby players expressed the idea that NPC groups that share similar ideologies were believable. Lastly, players tend to project their own biases on characters as if it is the character's own reasoning.

[23] For example, a character learned from a parent via simulation or as an average of opinions on a topic.

Lastly, we examine another system that highlights beliefs as a main mechanic. *MKULTRA* [66] is an experimental game that uses natural language as the main interaction mechanic between players and NPCs. A prominent feature of *MKULTRA* is "belief injection," through which players can manipulate NPCs by directly injecting false beliefs into their knowledge bases. The player mainly uses belief injection to accomplish goals through NPCs, solving puzzles. For example, the player may influence an NPC by injecting the belief that the player (Betsy) is an inanimate object. Therefore, it is not dangerous to give the player a chance to complete their goal. While belief injection and NPC interactions via natural language sound promising, Horswill notes a few problematic areas. The boundless utterances of natural language caused a problem for players, as players assumed that the problem lies with their logic rather than the limited strings the system actually responds to. Furthermore, the game's cultural view (for example, the NPC's knowledge of CIA agents) does not compete with the player's real-world cultural reference (the player's knowledge of CIA agents), creating an imbalance of viewpoints. Lastly, the game does not match player expectations; for example, the game features an RPG-like setting but operates in a puzzle-like fashion, creating false player expectations.

2. A Historical Lesson on Beliefs

Abelson's work [3], *Computer Simulation of "Hot" Cognition*, helped define the fundamentals of a belief system. In 1963, Abelson realized a gap exists in studying *hot cognition* (affect-reasoning) vs. *cold cognition* (problem-solving), specifically as it

relates to *attitudes*. Attitudes can be seen as a "belief system invested with affect," where affect represents the "evaluation of attitude objects" such as electronic music whereby the mentioned objects (i.e., music) can evoke feelings like joy or excitement. Abelson questioned whether it was possible to simulate a system capable of modeling *attitudes,* including how a system manages attitude changes and resistance to those changes. In other words, can a belief system handle challenges and disputes?

Abelson adapted the general term for beliefs as a broad notion that includes facts and assertions of truth as well as sentences that define relationships between one's goals, methods, and rationality. Their proposed simulation can model attitude change by reflecting on a *direction of change,* accounting for the degree of change from a positive element in a sentence to a negative element. Furthermore, Abelson presented his model for hot cognition using symbolic representation as a basis. Abelson further modeled his system by taking into account a list of *mechanisms*[24]or components that include denial, stopping, thinking, bolstering, and rationalization, among others. For example, the mechanism of denial is seen as “denying the truth or value of a sentence.” His chapter [3] focused on portraying rationalization[25] and integrating it with the overall flow of the system.

[24] Mechanisms are also called “microprocesses,” as borrowed from Miller et al. as “plans for changing images.”

[25] Rationalization is defined by Ableson as the process of accepting a sentence as true, at the same time accounting for any unstated conclusion.

We wanted to note a few things we found of interest in this model:

1. The model presents beliefs and attitudes as sentences or thoughts based on assumed perceptions (as if heard or read).
2. The system checks for validity, first depending on the context of the sentence, a natural language utterance weaved into the system as predicate calculus. If the sentence is incomprehensible, it substitutes the sentence with another sentence from memory. Otherwise, it checks it for *imbalance* (if it contains conflicting positive or negative values) and measures the degree of that imbalance. Assuming the sentence is not imbalanced, it evaluates the sentence for "good or bad" outcomes determined by a positive or negative predicate value.
3. As mentioned, the model attempts denial and rationalization as part of the process using symbolic logic and memory retrieval (for sentence creation in its rationalization).

Abelson and Carrol [2] extend their work on belief systems [3] by focusing on beliefs from an individual's perspective. They viewed an individual's beliefs as "subjectively rational" instead of choosing either rational or irrational viewpoints, as per the psychosocial debate of the time. While Abelson et al. noted the commonality of personality modeling with beliefs (incorporated to this day as seen in **Chapter Two**), they chose to implement their system based on an individual's motivational actions and their response to the actual situation, introducing a unique approach in belief modeling.

Furthermore, the authors identified problems relevant to belief systems that we consider (mostly) problematic to this day. As the authors note, people cannot easily

"believe what they want to believe." The issue lies in balancing a person's self-interest with their perception of reality; this can be explained by asking questions such as how can one balance accepting bad realities vs. rejecting them?

The second problem the authors pondered is how to best model change, specifically asking, at what range of resilience does change take effect? Relative to the second point, the authors consider the following three assumptions: first, the greater the affect (as defined previously) linked with a belief, the greater the opposition to changing that belief. Second, significant changes to a belief can occur solely when a *resistance mechanism*[26] fails or otherwise remains untriggered. Third, the more a belief system is *self-consistent* (defined as a closed system/fewer contradictions), the greater the odds that the *resistance mechanisms* will succeed. Lastly, the authors bring up the problem of choosing an appropriate linguistic structure for the beliefs; this issue remains relevant today, as authoring problems with natural language still persist.

Similar to their earlier model, the model in [2] uses linguistic and symbolic modeling such as the use of sentences, predicates, and concepts in structuring a "belief."" One feature highlighted by the authors is the system's ability to nest beliefs (elements and sentences) inside one another.

One interesting topic the authors bring up is how one sets up the memory structure for beliefs (in other words, how can we best author it?). The authors mention

[26] Resistance mechanisms are specified in their model, generally referred to as the default resistance of "belief change and affect change."

three methods, the first of which is writing it ourselves (or what the author calls "ad-lib") and comparing the simulation's result with that of a test subject. Another method is interviewing people and setting up a belief structure based on the responses; this is the most authentic and labor-intensive option. A final method involves choosing a known person whose beliefs fit well with the defined structure (defined as a "paraphrased" approach).

From our review of earlier systems, we saw how some beliefs were treated as knowledge facts that are either true or false. However, one wonders what exactly differentiates beliefs from knowledge facts. According to Abelson [2], beliefs and knowledge systems are actually similar; only when beliefs have met a certain number of characteristics can we specify them as beliefs. To elaborate, Abelson lists seven characteristics of a belief system; if a system does not include a sufficient number of these characteristics, then it is probably a knowledge system rather than a belief system. The seven characteristics are as follows: First, a belief system's elements should not be *consensual;* in other words, belief systems under the same circumstances should be individualized with varying complexity whereby one person's beliefs differ from another. Second, belief systems are often concerned with existential questions concerning the existence of belief-related beings (such as witches and deities). Third, belief systems should incorporate ideologies of *"alternative worlds,"* such as representations of the current reality and a utopian version of the world. Fourth, as hinted upon earlier, beliefs should rely on *evaluative and cognitive components*, as seen in the hot cognition work [151] in assessing components as good or bad. Fifth, beliefs

should include a person's experiences, or aspects learned from the world (such as methodology). Sixth, a belief system's beliefs are ordinarily *highly open;* in other words, there should not be a clear boundary for a belief system. Lastly, unlike knowledge facts, beliefs should have a degree of *uncertainty*, whereby a person can be committed passionately toward one belief and uncertain about another. According to Abelson, a combination of some of the characteristics mentioned above qualify a system as a belief system.

From Abelson's work, we note the importance of carefully structuring a belief system. We should consider various elements such as belief evaluations and validity checks. We note that our notions of beliefs should be guided by the seven characteristics [2] and Abelson's work [3] whereby we *consider beliefs as a personal biased view on elements, both real and imaginary.* Furthermore, beliefs are affected by and affect an agent's experience as well as real-world events.

In the context of believable characters, open and unbounded belief systems introduce too much complexity to be useful, so care should be taken to avoid edge cases. From a believability perspective, a bounded model may suffice as long as it does not conflict with the game's world. Furthermore, an unbounded model may fall short in system affordances and capabilities; it would break the character's illusion of life as addressed in **Chapter Two.**

Unsurprisingly, many of the issues uncovered by Abelson in belief modeling remain relevant to this day. One prominent issue is that of **authoring**. Unfortunately,

authoring will become more extensive as beliefs become individualized, as we will explore in **Chapters Five** and **Six**.

3. Reviewing Believability and Beliefs

As we discussed in **Chapter Two**, social architecture systems and, in turn, social interactions play a crucial role in character motivation. Some systems discussed in **Chapter Two** and the previous section of this chapter employ belief modeling as a basis for social interactions, such as relationship beliefs (e.g., CIF-based systems discussed in **Chapter Two**). These beliefs and social interactions cascade into character goals, linking beliefs, and character motivations.

While some social systems focus on social mechanics with limited belief modeling (e.g., reading beliefs as character facts), others employ beliefs as their main contributing factors. Systems such as *MKULTRA* [66] (via belief injection) or *Lyra* [13] (beliefs influencing debates on a topic) tend to be strictly scoped (interactions limited to specific topics); nevertheless, they enable newer emergent gameplay and player experiences (e.g., the player discovers group dynamics in *Lyra*). Others, like *Versu*, use beliefs in their design to change the narrative, creating a more immersive experience.

From our review of social and autonomous systems, a gap exists in modeling personal and world belief models, especially when we involve a character's morality as a reasoning device.

4. Moral Reasoning

From our previous sections, we recognize that beliefs and motivations can appear hand in hand. We can perceive instances in which motivations influence a character's beliefs and values, and in turn, shape a character's goal; we have seen some examples of the latter in **Chapter Two, Three** and in this chapter. Here we discuss how beliefs can be implemented through an NPC's reasoning device. We will first look at prominent theories in moral reasoning.

4.1 Motivated moral reasoning

Ditto, a Professor of Psychological Science, uses a metaphor of judges and attorneys to simplify the concept of *motivated reasoning*. Ditto tells us to think of *motivated reasoning* as us being judges looking objectively at facts or as attorneys gathering up evidence for a defense (viewpoint). According to the literature [106], we perceive ourselves and act as unbiased *judges* in some scenarios, while in others, we fail to recognize our biases, gathering evidence that supports our beliefs. We tend to act as *attorneys*, especially when faced with topics we care about (including concerns such as our own self-image, people's perceptions of us, supporting behaviors we like, and reprimanding ones we don't).

According to Skitka [137], moral beliefs are one of the deepest-rooted beliefs a person can hold, as cited by Ditto's survey [43]. Continuing with Ditto's metaphors of attorneys and judges, where we may act as attorneys to reach a favorable outcome, the term *motivated moral reasoning* adds a layer of morality, where we are also motivated

by our inclination to achieve specific moral outcomes. There are three significant features we should consider regarding moral judgments. According to Ditto, *"Moral judgment is deeply evaluative."* We each incline to what we consider moral or immoral behavior; this particular feature is tied to social aspects and group behaviors (people generally want people to perceive them as good or moral). In stating that "moral judgments are inherently intuitive," the author reviews literature arguing for the affective nature of moral judgments as intuitive and responsive (such as hinted by Abelson's work above), rather than as *thoughtful* at a fundamental level. This trait is illustrated in situations in which certain behaviors just feel "right" or "wrong," "good" or "bad." For example, "murder is evil" is a fast, intuitive judgment. However, the authors note that these judgments can be overwritten, but this precedes our moral evaluation phase. Lastly, "*moral judgment is complex and multifaceted.*" Our reasonings should allow for both preferences and will enable us to manage some objectivity.

Before we move into another perspective of motivated morality, one interesting detail that Ditto et al. [43] describe is the elements in our moral evaluations. When we assess a moral situation, we generally focus on either the person (actor), deriving their responsibility and consequences, or the act itself, relying on how the act measures up against the recognized customs of what is deemed moral behavior. Lastly, these judgments on an actor consider our assumptions on their motivation, purposes, and knowledge—all elements of what the authors call *Moral Accountability*.

While Ditto's work presents an exciting system possibly, we chose to create systems based on Lakoff's research [74, 77]. We believe that Lakoff's theory fits better as a representative example of moral reasoning, as we will see in the upcoming discussion.

For more on morality, please note that **Chapter Three** reviews and analyzes morality in games, covering representative moral and reputation systems, while **Chapter Nine** reviews morality through a faction lens. This section (in addition to section 4.2) reviews theories that link beliefs with moral reasoning from prominent experts, choosing one of their theories as a basis for our moral system design. We later discuss how we can apply them in games.

4.2 Moral Reasoning as Metaphors and Frames.

We start this section by examining Lakoff's work in moral politics. Lakoff is a prominent figure in cognitive linguistics; his work examines complex issues such as moral politics and utilizes a unique approach to illustrate how people think and face these issues.

Why politics and morality? As Lakoff mentions, "All politics is moral" [77]; we acknowledge that politics are intertwined with our lives; we argue based on our beliefs and notions. However, as we saw in **Chapter Three**, there is a clear divide between good and evil in most games and reputation systems; we believe moral politics can add a little complexity and, hopefully, shades of gray to moral systems.

Lakoff's [77] moral systems are intricate, creating moral viewpoints based on metaphors and frames. How we frame a point makes a big difference. For instance, Lakoff gives an example of a perceived “evil” leader putting forth a new policy. The hypothetical leader would not explicitly convey a bad policy as, "follow this policy because it is evil." Instead, the leader would frame it as a policy they believe is right. Morality is not simple; it can be different ideas from different people following different perspectives of right and wrong.

Like Ditto's survey [43], we note that moral reasoning is primarily unconscious; however, some people rely on different moral systems in varying situations, termed bi-conceptulism. In this subsection, we will examine a few of the moral systems Lakoff mentions and discuss interesting concepts related to moral reasoning.

In his books *Moral Politics: How Liberals and Conservatives Think* and *Do Not Think of an Elephant* [74, 77], Lakoff describes two metaphorical models of thought that we rely on in arguing for our held beliefs and sense of morality. Lakoff uses these metaphorical concepts throughout his books to illustrate these models. From a gaming context, we encourage the reader to think of character groups, towns, and factions as family units.

Generally, Lakoff explains that there are two types of “family” systems: a Strict Father Morality (SFM) and a Nurturant Parent Morality (NPM) system. At the same time, there is no “middle”; however, there are bi-conceptulists that apply elements of either model depending on the situation at hand.

Let us take the first model, the Strict Father Morality (SFM), as an example. As the name suggests, [75] this model employs a strict figure as the head of the household; they believe that "the world is a dangerous place." Therefore, the parent needs to protect the children, act authoritatively, and teach them "right" from "wrong," reward them if they do good, and punish them if they do wrong. Each family system employs metaphors or values they believe in; for example, the SFM values strength. Anything that reduces strength is seen as weak and immoral. Lakoff offers an example whereby the act of purchasing illegal drugs is immoral, as it emerges from low self-control, according to SFM (i.e., low strength). There are other metaphors and values that the SFM believes in, such as Moral Boundaries (i.e., deviating from the norm is wrong), Moral Wholeness (i.e., uniformity and unity is good), and Moral Purity (i.e., against corruption, "one bad apple ruins the bunch") ,among others (we will explain the subset of metaphors used in our system in **Chapter Five**).

The NPM [76], on the other hand, is a family unit that supports interactions between its members. It believes children grow as a result of nurturance and care. Children (members) learn self-reliance by caring for others. Unlike the SFM, the parents value their child's opinion; children, in turn, respect their parents rather than dread their punishments or anticipate their rewards. Like the SFM, the NPM has metaphors and values it believes in, such as Morality as Fair Distribution (it covers sub-metaphors such as equality of opportunity or fairness but does not inform us what model is needed when), Morality as Social Nurturance (e.g., strengthening social relationships with others and mending those relationships), and Morality as Nurturance

(i.e., being regularly empathetic or putting the child's needs ahead of one's own (to an extent).

These models help us imagine how characters can respond better to various events or situations. One interesting area that Lakoff explored is the concept of self-interest. Self-interest was also raised in modeling belief systems by Abelson [2]. These SFM or NPM systems thoroughly describe how one acts according to one's self-interest (or self-nurturance in NPM). The NPM [76, 77] views self-nurturance as taking care of oneself and then aiding others. For example, one accounts for their happiness (or joy, responsibility, etc.), but they aid others in finding their happiness or help them through nurturance. On the other end, the SFM, as hinted earlier, values strength and self-reliance; aiding others may become immoral as the aid itself robs the person of competition, contentment, or rewards that may come. Our project builds an NPM and SFM model as a basis for our character's reasoning; we will further discuss it in **Chapters Five and Six.**

At this point, the reader may be curious and probably speculating about the validity, as these models seem binary, opposites of one another. Then how is this model any different from any binary system of good and evil? While it is true that there are two main models, each model offers a range of variations. The SFM [75] handles variations through four categories: Linear Scales, Pragmatic-Idealistic Dimensions, Moral Focus, and Moral Order. The NPM [76] shares the former three categories of variations, except Moral Order. Let’s take the idea of *Linear Scales* as an illustrative example in extending moral systems. To elaborate, The SFM's idea of Linear Scales

makes the judgment based on the quality of the act rather than the quantity; it is a matter of perspective on which a person handles punishment or reward, creating various degrees in the process (Lakoff illustrates this by highlighting the difference between punishing a child by not letting them watch their favorite TV show vs. sending them to bed without dinner).

Furthermore, the models can also have perceivably alternative systems. For example, models such as Spiritual progressives and Environmentalist systems exist. However, Lakoff notes that these systems still share common core values of NPM, yet some fail to realize their model as a special case of the grander NPM model. It is also fair to note that the models presented here are not the ultimate go-to models for analyzing moral belief and action. However, this and other similar models could serve as examples of how we can create better characters with greater reasoning capabilities.

Finally, we consider how utilizing moral values can affect change which, as mentioned, is an underdeveloped believability criterion. As we structure characters with varying beliefs, we can imagine scenarios where the player, for instance, persuades an NPC about another character or their opinion. At a basic level, one can use these moral values to change an NPC's judgment.

5. Conclusion and Looking Ahead

As we explored in this chapter and **Chapter Two**, belief modeling in games is commonly associated with character facts and theory of mind. We also explored other systems whereby belief modeling and believable characters intersect. We examined

how beliefs were utilized in NPCs believing in false information (e.g, [61, 132]) used as a game mechanic via belief injection (e.g, [66]) and alongside social systems e.g, [51], creating unique stories.

In **Chapter Three** we explored how morality is often perceived and created in popular games. We discussed two primary avenues for designing morality in games: reputation scales and state machines.

After our discussion, we noted that morality, in some cases, was multifaceted and expressed via NPCs. However, these cases are often limited, based on select interactions, or specifically crafted in unique NPCs (e.g., companions).

We then argued in favor of a systemic approach to value-based moral designs. We believe in modeling NPC values and beliefs; we can create background NPCs that not only seem believable but can also communicate differing morality.

In our next two chapters, we explore two system iterations of Lakoff's model, examining the system's design and limitations, covering what went right and wrong, as well as our technical details, covering how our characters reason and how the system functions. In **Chapters Seven** and **Eight**, we reveal the results of our studies, examining lessons learned, focusing on the link between values, morality, and believability.

Chapter Five: Argument Box: A World of Shapes

Acknowledgment(s): The contents of this chapter have been published elsewhere [4, 5]. Special thanks to Yasheng She for authoring the first version of this system.

From **Chapter Three**, we learned that many moral systems portray mortality as a binary value of good and evil. The reputation of the player character (and sometimes the Non-Player Character (NPC)) is commonly represented as a single-dimensional scale by which players or NPCs land on a spectrum based on their deeds and the morality points they accumulate. Through our examination, we noted that NPCs in these types of games tend to employ surface-level judgments about the characters they encounter. For instance, games such as World of Warcraft and Neverwinter Nights [20, 23] employ NPCs that attack opposing factions or views without any reasoning from the characters involved, usually based on high-level decisions such as the character belonging to an opposing faction. These simplistic moral decisions and binary systems can affect a character's believability, making them feel mechanical, stereotyped, and less life-like.

We believe we can utilize beliefs and values and expand how moral systems are portrayed through a character's moral reasoning. In **Chapter Four**, we reviewed value and belief-based systems, covering how beliefs are defined, utilized, and implemented in believability-based systems. By the end of Chapter Four, we suggested utilizing the

concept of moral reasoning by embedding known theories as a basis for NPC reasoning. We suggested embedding Lakoff's metaphors and frames into a moral system as an example.

This chapter presents our first iteration of Argument Box, a game prototype that employs moral reasoning for NPCs at both surface and deep levels.

1. Initial Concept

In the classic Monty Python skit "Argument Clinic" [34], a man approaches a clinician asking to buy an argument; he then proceeds to argue with the clinician about arguments, debating whether their argument is an argument! Likewise, our prototype features an argument simulator game, in which NPCs walk into a shop called Argument Box (AB) to procure arguments with the player character, focusing on arguments about their and other character's moral behaviors.

The game features characters exported from Talk of the Town [130], a simulation created by James Ryan. TotT, as a reminder, is a historical town simulator that generates characters, including various elements such as character relationships, locations, and jobs. TotT is further discussed in **Chapter Two**. We used the characters as a basis for our world. The characters are thematically modeled as polyhedra to simplify animation and reaction purposes. Thus, in this chapter, we may see references to "shapes," which in our world is the same thing as saying "people."

2. Gameplay

The game starts when an NPC enters the clinic (Argument box) and brings up a rumor about someone living in town. They start gossiping about another character, highlighting their opinions of that character and their feelings about their behavior. At this point, the player can agree or disagree with them. Figure 11 shows an example of a cube coming into AB, stating a rumor they heard about an NPC in town, and expressing an opinion.

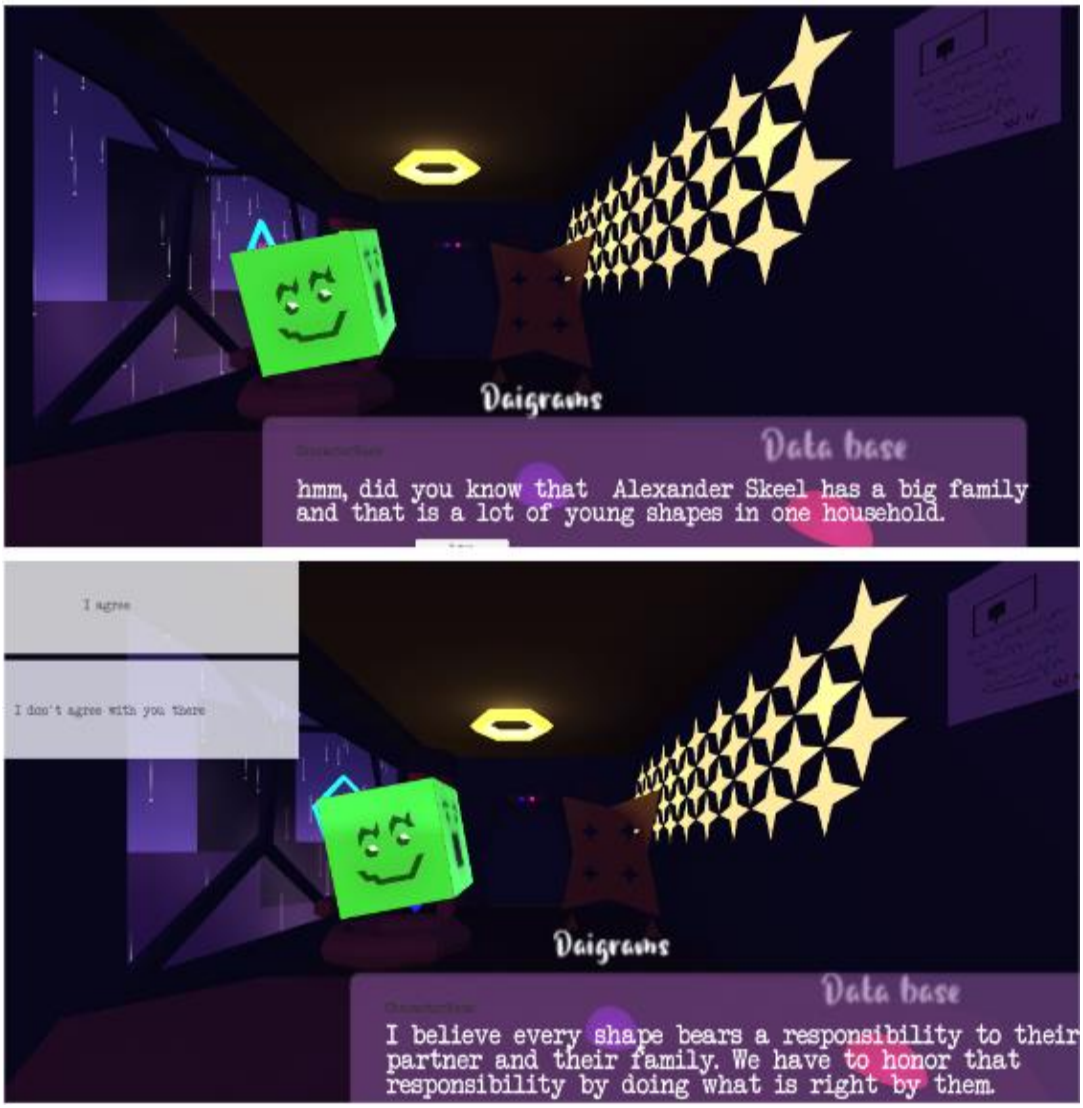

Figure 11 shows an NPC coming into Argument Box and stating an initial opinion.

Depending on the player's response and how passionately the NPC feels about the current argument, the NPC can either bring up a surface-level response or reference their deep-seated beliefs as moral arguments. We note that these deeply held beliefs are

based on Lakoff's moral metaphors, as we will soon explain. Figure 12 shows an NPC's response as a surface-level argument.

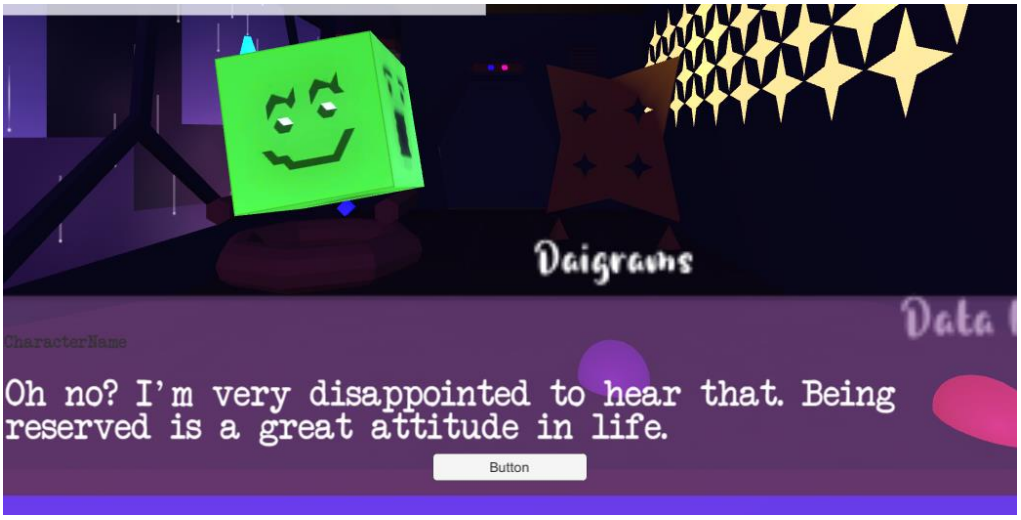

Figure 12 depicts a CNPC's response as a surface-level argument.

While conversing at a deeper level, the player can appeal to the NPC by bringing up multiple reasons (values) that may resonate with the NPC, depending on their moral model. These conversations go back and forth between the player and the NPC until either the NPC is convinced or the player concedes. Figure 13 shows an example of player-based counterarguments on a given topic.

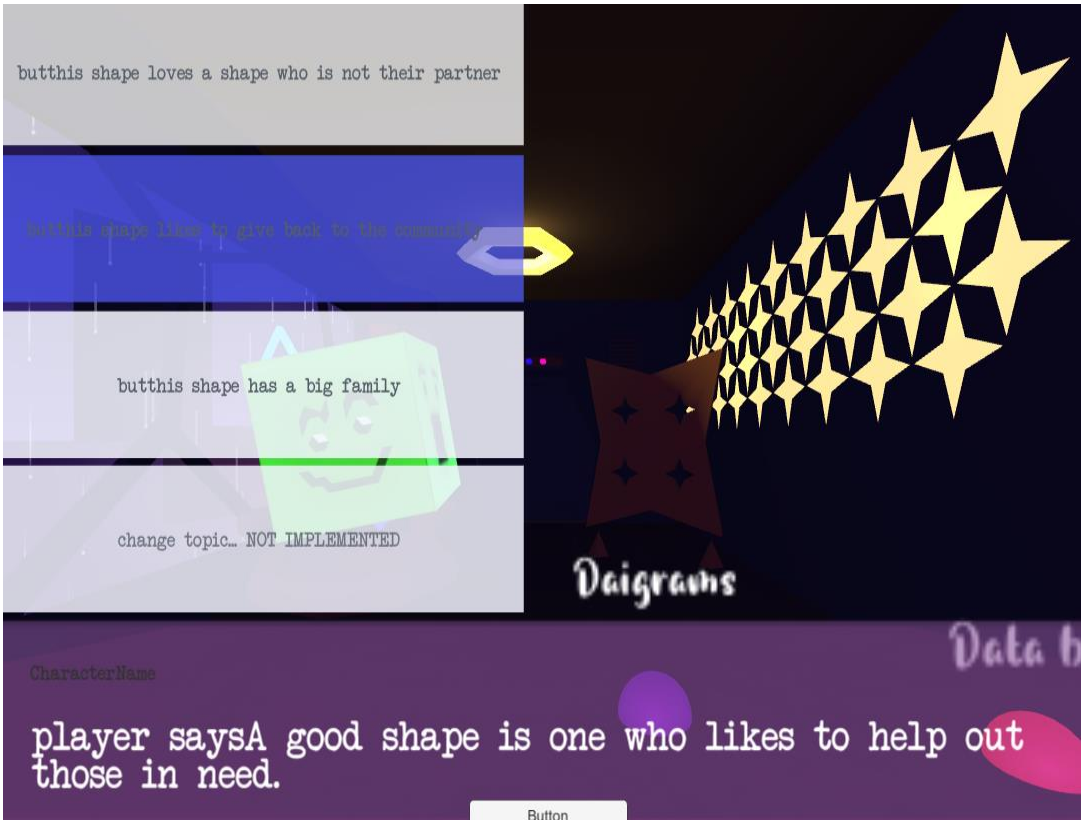

Figure 13 presents player counterarguments as options.

Lastly, this version of the game features a search bar that the player can exploit to look at NPC facts. These facts help the player learn more about the NPCs that live in this world. Upon searching for a name, the UI is filled with patterns belonging to that NPC. Figure 14 shows an example of searching for a character and getting a list of their patterns.

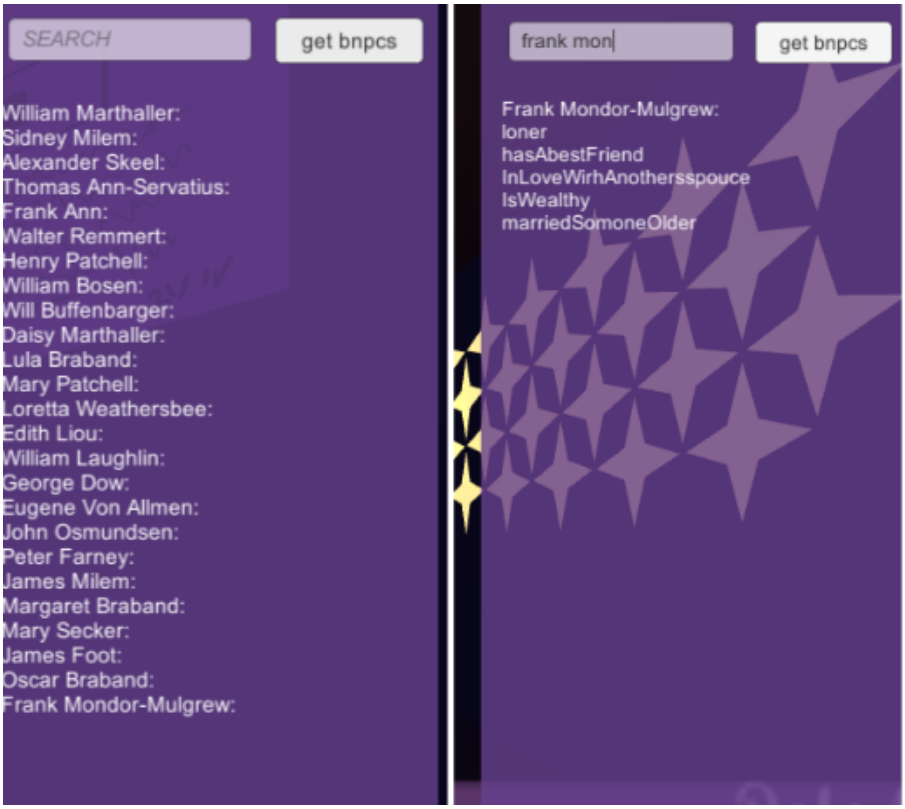

Figure 14 depicts a UI interface for searching and listing character patterns.

In the following sections, we will discuss the system in greater detail. First, we examine the components of our system at a high level; we then look at our NPCs, covering their architecture, design, and belief modeling. We will then illustrate two example arguments: one at a high level and another that references the NPCs deeper mode of reasoning (moral model).

3. System Structure

3.1 Overview

As mentioned earlier, the game features a character coming into AB to state a rumor and gossip about someone. The design of the game differentiates between the conversational NPC and the gossiped-about NPC. For ease of reference, let us call the NPCs the player converses with as Conversational NPCs (CNPCs) and the gossiped-about NPCs as Background NPCs (BNPCs).

The BNPCs include a separate list of NPCs that the player character never interacts with; they are used to seed conversational topics by our CNPCs. These BNPCs are generated from TotT. We import the data from TotT in JSON format. We then search for patterns, such as combinations of attributes or social connections on BNPCs or temporal sequences undergone by BNPCs, in a process similar to story sifting [117]).

BNPCs are assigned tags based on the patterns that match. We then filter the BNPCs by thresholding the number of tags (only BNPCs with enough tags are potential topics of conversation) and place the filtered list of BNPCs into a priority queue based on the number of tags found and the quality of the tags. Tags involving multi-character patterns are weighted more highly than tags resulting purely from within-character patterns. In a later section, we provide examples of the sifting patterns we use.

Once we have our list of BNPCs, the CNPC brings up an appropriate dialogue based on the tags present on the BNPCs. The starting dialogue is unbiased and simply expands the tag into a textual utterance. For example, the tag familyPerson gets translated into, "Have you heard that X has a large family?"

Each pattern in the system is further mapped to the CNPC's surface values. These surface values denote how a CNPC generally feels about a set of tags; there is a many-to-many mapping of tags to surface values, which is described in more detail in a later section. The CNPC holds each surface value with a high, medium, or low strength, indicating how passionately the CNPC believes in that value.

As long as the player agrees with the CNPC's judgments, the conversation will stay at the level of surface values. However, when the player disagrees with the CNPC on a surface value they hold strongly, the CNPC will perform deeper reasoning using their deeper models to back up their claim. This version of this system only utilized the strict father model (SFM) for testing purposes (please note: **Chapter Four** introduces Lakoff's models and theory).

The conversational options provided to the player allow them to agree, disagree, bring up a specific discussion, or change the conversational topic entirely. Figure 15 illustrates the system's general components. Next, we will illustrate the components of the system in greater detail.

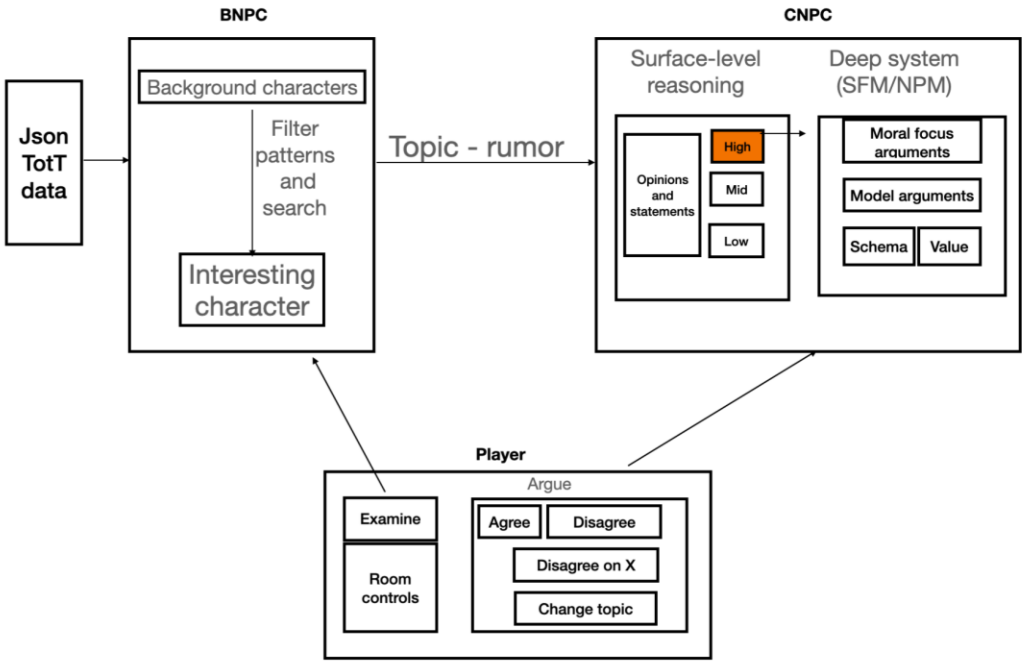

Figure 15: System overview

3.2 Modeling BNPCs

BNPCs are generated using the TotT simulation. Our sifting patterns focus on topics we can have moral debates about, and so generally exclude details such as locations of homes and businesses, street names, and physical character descriptions. We currently use fifty-eight patterns to assign tags to BNPCs, as follows:

- **Patterns that directly map a single TotT raw attribute**. Examples include *isWealthy*, *departed*, and *familyPerson*, indicating if a character has wealth, left town, or has a family, respectively.

- **Patterns that are created by combining different TotT raw attributes**. For example, a retired character that is forty-five years old has the pattern *retiredYoung* assigned.
- **Patterns that are based on TotT character jobs.** These patterns are used to make controversial assumptions about characters, based on the character's career and its potential effect on others. For example, any TotT character with the job miner or cooper is assigned the tag polluterRole. This is later used for conversations related to the environment.
- **Patterns that focus on relationships with other TotT-generated characters** (such as love triangles and backstabbing). Examples include f*riendWithBestFriendsEnemy* and *InLoveWithSpouseOfFriend*.
- **Patterns that combine lower-level tags into higher-level ones.** For example, if a BNPC has the tags *adultButNotworking* and *IsWealthy* (from lower-level patterns), this results in the higher-level tag *notWorkingAndRich*.
- **Patterns that combine with personality traits.** The last category of patterns combines tags found by other patterns with the Five Factor Personality traits [58] that TotT assigns to characters. For example, the pattern *hasAlotOfEnemies* combined with the personality factor of high *agreeableness* results in the tag *tooTrustingOfEnemies*.

As mentioned above, once all pattern matching is completed and the tags have been assigned, BNPCs are placed in a priority queue based on the number and quality

of tags. Starting the conversation about characters at the front of the queue ensures that the CNPC will have a good number of debatable topics to argue about with the player.

3.3 Modeling CNPCs

3.3.1 At a surface level.

The CNPC starts the conversation by picking the BNPC at the front of the queue to talk about, doing this until it has exhausted all the BNPC's tags or the player chooses another character to talk about.

Once a tag has been chosen, the CNPC starts the conversation by commenting on the tag in a value-neutral manner. For example, if our BNPC named Mike had the tag familyPerson selected by the system, the CNPC states, "Oh, have you heard that Mike has a big family?" The CNPC then states how they feel about this tag by relating it to their surface values.

There are currently 28 surface values defined in our system, of which a subset will be held (with varying strengths of low, medium, or high) by a CNPC. Examples include *LoveIsForFools*, *LoveAboveAllElse*, *FamilyPerson*, and *ShapesAreNothingIfNotSocial*.

BNPC tags map in a many-to-many way with surface values. The tags that map to a given surface value are called core tags of that surface value. The core tags have a many-to-many relationship with the surface values. For example, the tag willActOnLove is a core tag of both the surface values BeTrueToYourHeart and LoveIsForFools.

In the event that a core tag maps to two or more mutually held surface values, this provides some non-determinism on how the CNPC will comment on the presence of this tag, depending on which surface value is considered to have been activated.

When a CNPC is instantiated, the system randomly assigns the surface values and their accompanying strength. Some surface values are mutually exclusive, so can not be simultaneously held with high strength. For example, if our CNPC holds *LoveIsForFools* with a high rating, it cannot hold *BeTrueToYourHeart* with any strength other than low.

Additionally, the mapping from core tags to surface values is used for the conversational options presented to the player. This allows the player to bring up BNPC characteristics during the conversation that explicitly agree or disagree with the CNPC at the level of surface values.

The surface values are used for immediate value-laden reactions during the conversation. As long as the player and CNPC agree with each other, the conversation can stay at the surface value level. However, when the player's judgements disagree with the CNPC, the system switches to reasoning about deep values as determined by the Strict Father Model (SFM).

Switching to this deeper model allows the CNPC to marshal arguments by bringing up characteristics that relate to more deeply held values. This prevents the conversation from immediately degenerating into repeated assertions (e.g., "Yes it is! No it isn't! Yes it is! No it isn't") at the surface level.

3.3.2 At a deep level.

We define six deep values drawn from Lakoff's book *How Liberals and Conservatives Think* [75, 77] to specify the SFM:

- **Moral Boundaries** warns about the danger of deviating from the norm. Characters that deviate from the norm are **seen as immoral by a character holding the SFM.**
- **Self Interest** sees seeking one's self-interest as moral and interfering with one's self-interest as immoral.
- **Moral Wholeness** is concerned with unity and conformity among characters.
- **Moral Essence** evaluates a character's past actions as indicators for their future actions, making the assumption that past actions are the result of a character's "essence."
- **Moral Strength** values a character's ability to act in or handle difficult or sensitive situations. Low strength is seen as moral weakness.
- **Moral Order** accounts for traditional hierarchical power relationships, such as rich characters viewed as morally superior to poor ones.

As we mentioned earlier, each surface value includes a set of core tags. Our SFM deep values provide a mapping for each core tag to high morality or low morality. High indicates that the core tag exemplifies the deep value; low indicates it violates the deep value. For example, the surface value *BeTrueToYourHeart* and one of its core tags, *inLoveWithSpouseOfFriend*, is evaluated as low by the Moral Boundaries deep

value. If our CNPC supports the surface value *BeTrueToYourHeart* and is involved in a argument with the player over this value, it would not make use of the BNPC tag *inLoveWithSpouseOfFriend* as an argument for their claim, as this violates a deep value.

Our upcoming examples will further clarify these situations. Furthermore, the SFM can check as to whether a CNPC is searching for a pro or a con argument to back up a claim about a given surface value. Generally, if a CNPC is supporting an argument for a given surface value, the SFM is looking for core tags that evaluate as high against the deep values, possibly switching the argument to another surface value if no such tags can be found for the current surface value (this is an example of deflecting to another topic to continue to argue for the moral virtue or vice of the BNPC if the CNPC can no longer argue for the current surface value). Alternatively, the SFM can look for core tags that evaluate as low against a deep value to provide a cautionary tale in the argument.

For example, if the CNPC holds the surface value *LoveIsForFools*, the SFM can look for a core tag, such as *inLoveWithSpouceOfFriend*, that evaluates as low against a deep value (e.g., *Moral Boundaries*), to make a negative statement about the CNPCs behavior, and thus support their *LoveIsForFools*

4. Example Conversations and Structures

4.1 Surface-level Conversation

Here we will illustrate a typical conversational loop showing what happens when the player agrees with our CNPC on a medium-rated surface value.

At the start of the loop, the CNPC picks a BNPC to converse about, in this case Caroline Milliem. Caroline has three tags available, *inLoveWithSpouseOfFriend*, *familyPerson*, and *willActOnLove*.

The CNPC selects Caroline's first pattern, *inLoveWithSpouseOfFriend*, and opens the conversation by gossiping: *"Oh, have you heard? Caroline Milliem is in love with their friend's spouse!*" Under the hood, the pattern was mapped to the surface value *LoveIsForFools* as "*inLoveWithSpouseOfFriend" is one* of its core tags.

We note that this assignment happens at random; it could have mapped to other surface values, as long as the pattern belonged to that core set of the selected surface value.

It so happened that our CNPC is relatively indifferent to this value, holding it with medium strength. They care about the value but not so much that they would raise a fuss if the player disagreed with their opinion (represented as a "disagree" button). The CNPC then presents its opinion based on this surface value, stating, *"Well, I think that shapes should never let their emotions cloud their judgment."*

The player then has the option to agree, disagree, or bring up another topic or BNPC as clickable options. In our scenario, the player agrees: *"I couldn't agree with*

you more." The CNPC, satisfied with the result, says, *"Yeah, it's good that you see it my way!"* We note that this scenario presents a medium stance on a surface value; if the player disagrees, the CNPC simply notes its displeasure with the player's choice.

The system then checks to see whether the other tags of the current BNPC have been explored; if they haven't been explored, the system proceeds with the same loop but with the newly appointed tag *familyPerson*. Otherwise, the system moves on to the next BNPC in the queue. Figure 16 highlights the surface-level loop mentioned here.

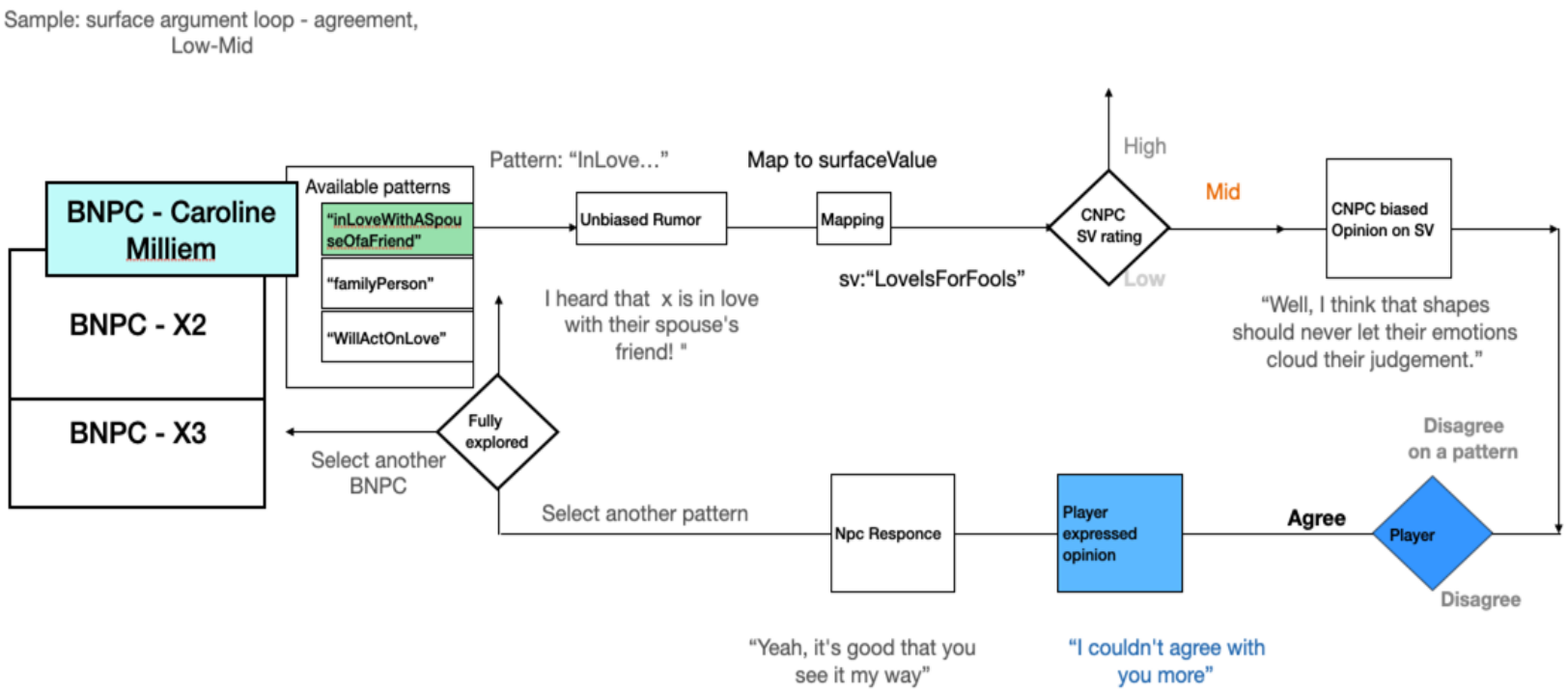

Figure 16: Surface level conversational loop, player agrees

4.1 Deep Rooted Conversation

Let us take the previous example but assume that our CNPC actually cares deeply about the surface value LoveIsForFools. The CNPC conveys this as, “Well, I think that shapes should never let their emotions cloud their judgment AND doing what’s right is always better than doing what feels right.” As a note on authoring, we use AND and BUT to emphasize the CNPC’s high or low stances by extending the same sentence with a modifier that enhances how strongly or weakly they feel about a given topic. This helps minimize the combinatorial amount of dialog we have to write and signals the underlying model more strongly to the player. Authoring components are expanded upon in a later section.

In this example, the player disagrees with the CNPC, selecting the dialog option: *“Are you kidding?? Love is the best thing ever.”* the tone mimics that of the CNPC. The CNPC then responds with, *“You’re joking! Love is for fools.”* It consults the underlying model, in this case the SFM, to back up why love is indeed for fools.

The currently selected tag *inLoveWithSpouseOfFriend* is validated as a core tag of LoveIsForFools, and determined to score low against the deep value Moral Boundaries. Thus the SFM validates the surface value stance, and allows the CNPC to make an argument based on deep values: *“Love can be immoral. Honestly, they have no boundaries. This shape went after their friend’s partner; that’s just wrong.”*

We note that the text is written in a way that references the specific tag, illustrates the deep value, and provides reasoning as to why the surface value *LoveIsForFools* should be held. The player can then agree or disagree with the pattern,

choosing another BNPC pattern to bring up, passing it as an argument for the current surface value.

If the CNPC is presented with a tag that the SFM maps in a way that is contradictory to the surface value argument being made, it then searches through the BNPC's available tags for an alternative argument. If any of the remaining tags are core tags of the surface value and the SFM mapping supports the argument, it presents it as a backup argument. Otherwise, the CNPC mimics a person backed into a corner, randomly firing off defenses based on the tags found. This results in more generic arguments, for example, stating, "But that BNPC has a family!" in response to the situation where the BNPC is in love with the spouse of a friend. The diagram (Figure 17) illustrates the deep conversation loop discussed here.

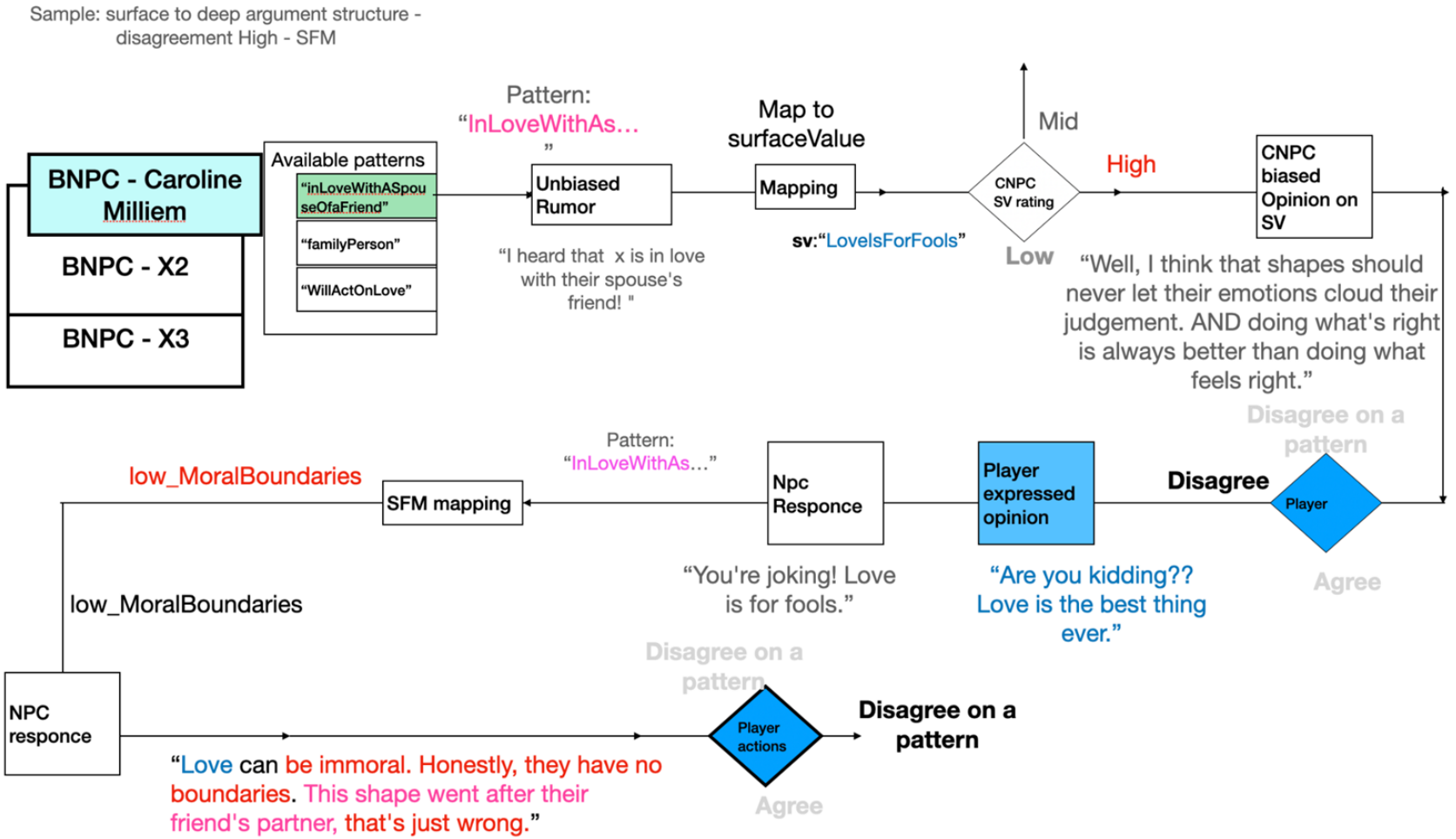

Figure 17: Deep value conversation loop - player disagrees with high surface value

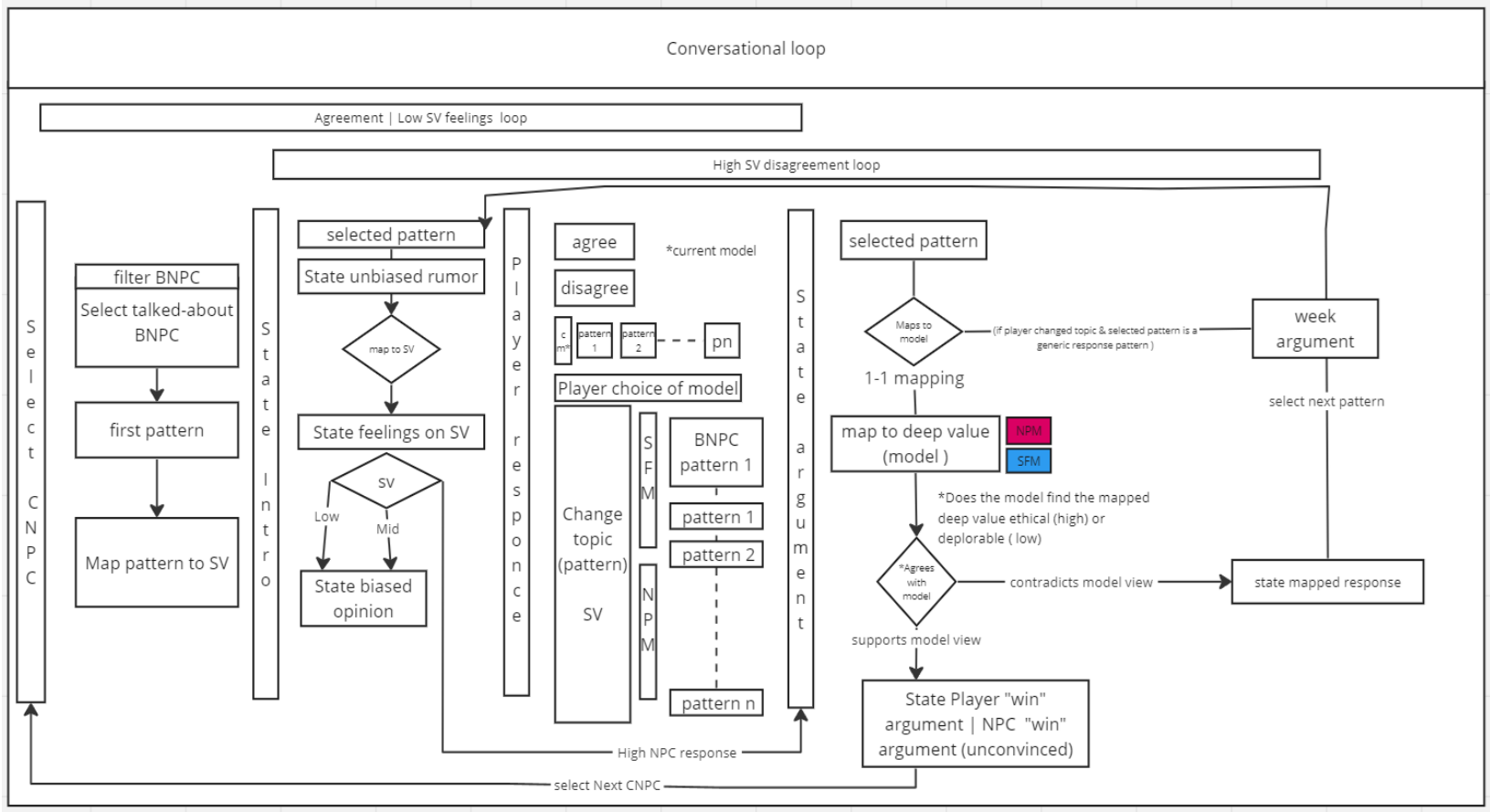

Figure 18: Argument Box: version one the conversational system diagram

5. Authoring AB

The original argument Box included text authored by our team's writer, Yasheng She. Unsurprisingly, authoring conversational flows at different levels was a taxing effort for our writer. As mentioned earlier, the game includes behavioral patterns whereby each behavioral pattern is translated into a surface value and eventually a deep value, each requiring separate strings of text depending on its state (e.g., agreement, disagreement, deep value, and con and pro stances).

We needed to create text for each of these values, including transitions such as unbiased rumors and introductions. We had twenty-nine surface values (including pro and con variations), authored at three surface-value levels, as well as introductions, agreements, focus areas[27], quick responses, and disagreement texts; this leads to about sixteen to nineteen strings per surface value (five hundred and fifty-one strings).

Additionally, our deep values (moral model) contained responses for fifty-eight behavior patterns, each mapped to a single deep value in a pro or con

[27] Focus areas included an expansion of this version of the system where CNPCs have specific issues that they care deeply about (beyond others). This focus area is based on Lakoff's work [148] in providing a variation for a moral model. To represent moral focus in our system, we included special loops that iterate around these specific patterns. which a CNPC finds highly important. However, these were dismissed as the system grew more complex.

As each behavior pattern could be mapped to several surface values, we needed to create responses for each pattern under each relevant surface value. We ended up with 387 per model, excluding generic responses (generic responses are added to avoid duplication). While the prototype focused on the SFM, the actual json files included both SFM and NPM variations.

In authoring the system, we created Google Sheets that exported the written text into JSON objects, which were read into the game. Table 4 shows an example SFM spreadsheet where we highlight surface values, a CNPC's last statement, the sub-value used to trigger the conversation, and finally, the morality level (high or low) referenced in the figure as the schema used by the NPC to justify their argument (deep value).

Surface Value	Npc's last disagreement	Sub value	Scheme Text	Schema	SFM Core SubV Text
LoverOfRisks	Boooo! Don't be a coward!	WillActOnLove	This shape is too weak to handle their emotions.	low_Strength	Acting on love that is not appropriate will hurt others and yourself.
Teetotaler	BOOHOO. you're a bumper. A little drink never hurts no-body.	socialLife	This shape makes good company.	high_MoralWholeness	Alcohol makes you a better company just like this shape. Look how many friends they have and I am sure drinks help.
Teetotaler	BOOHOO. you're a bumper. A little drink never hurts no-body.	WorksInAlcohol	This shape clearly understands what they want in life.	high_SelfInterest	Also, this shape does honest work and I don't think selling alcohol makes them a bad shape.
TeetotaslerAnti	That's sad. Shapes who love to drink are not happy, they're miserable.	WorksInAlcohol	This shape clearly understands what they want in life.	high_SelfInterest	Also, this shape does honest work and I don't think selling alcohol makes them a bad shape.

Table 4: Partial authoring table as a representative example of an AB SFM sheet.

Other authoring elements were structured similarly to the table above, with the caveat that each sheet includes data relevant to their place in the overall conversation (i.e., sheets for generic responses, player responses, surface values, and model sheets are structured differently). These tables were exported as JSON objects and read by the system, providing textual data where appropriate.

6. Lessons learned and Uncovered Issues

While our initial design contained interesting goals and features, our authoring and playtesting process revealed the following design issues.

6.1 Flow, Loop Design, and Control Issues

Giving players control of the conversational loop by allowing them to select any applicable pattern (seen by the system as a topic change) created transitional chaos. CNPCs saw pattern changes as topic changes and responded accordingly. Sadly, these transitions often led to confusing responses without additional context between each transition. Where the player might have been interpreting selecting a behavior pattern on the BNPC to talk about as a move in the current argument, instead this was changing the topic of the argument in mid-stream.

For example, consider instead the following scenario: A CNPC argues that a BNPC is moral for butchering animals for a living; the CNPC states this based on its surface value *AntiAnimalLover* and the BNPC's behavioral pattern *butcherRole*. The player then disagrees in the deep-model loop by changing topics. The player does this

by selecting the option associated with the pattern *InLoveWithAnothersSpouse*. Unfortunately, this pattern by happenstance is assigned to a social-based surface value, resulting in the following:

CNPC: We need to eat animals to survive.AND we're on the top of the food chain for a reason, it's our right to do what's best for us.

Player: WHAT?! That's so cruel and selfish of you!

CNPC (using MoralOrder): [BNPC] understands how the world works. They did what they have to do to make a living and no one should judge them for that.

Player (changing topics, selects [BNPC] is unfaithful, using *MoralBoundaries*): I would never trust this shape, this shape flirts with everyone even though they are committed, I don't think being social helps them.

CNPC (topic changed to social, using *MoralWholeness*): This shape is so open and friendly! Look how happy this shape is! Shapes who have a lot Friends must be good shapes.

As noticed from the conversation above, the transition between the morality of butchering animals to the BNPCs social-life is too jarring; the topics are too dissimilar and lack proper context to allow for a smooth transition.

Lastly, accessing the system's deep value model was hard to do given that the system randomly assigned topic surface values (low, mid, high) in its initialization phase.

6.2 Context and Authoring Burdens

As covered in an earlier section, the sheer number of CNPC utterances and player responses was extreme and yet, in many cases, failed to provide proper context for the narrative scenario.

Contextual problems arose because of the dynamic nature of our system. Responses, for instance, could be selected dynamically in many different orders, depending on the state of the moral argument transitions between responses, which were often difficult for the player to understand, and the authoring burden became intractable.

Additionally, due to the iterative nature of game design and the ongoing refinement of our system, we had to continuously strike and revise the actual dialogue lines mentioned above, increasing our authoring burden. We also needed to explain the system and modifications with each iteration, often resulting in unusable texts because of transitional issues, miscommunication, or logical bugs.

6.3 Forcing Patterns To Adhere to Strictly Defined Metaphors

As explained earlier, each behavior pattern was mapped to many surface values. Simultaneously, patterns were assigned a single deep value under their umbrella surface values; we believed this would help reduce the already extensive authoring required for the project. For instance, the pattern *DivorcedManyPeople* could be mapped to any one of these surface values: *BeTrueToYourHeart*, *LoveIsForFools*,

FamilyPerson, *AnAdventureWeSeek*, but under each surface value, a single deep value was assigned.

For example, the surface value *LoveIsForFools* judges *DivorcedManyPeople* as morally bad, using the deep value Moral Boundaries. However, it judges the pattern as morally good under the SelfInterest value (that belongs to a different surface value). As the player shifts conversations and topics, the patterns eventually become contradictory and confusing.

6.4 Player Feedback and Deep Value Impact

This version needed proper feedback elements. We included simple yes-no animations to respond to the player's dialog choices. However, Our CNPC lacked concrete reactions to the player's selected deep value response. Next, we list elements that were not problematic but posed design constraints and trade-offs.

6.4.1 Reusing a cast of characters.

At the start of the project, we chose to import our cast of characters from ToTT [130, 132]. Importing these characters helped us confine our world and eliminated the need for us to create our own simulation or structured character definitions from the ground up. While ToTT automatically handles aspects like relationships, names, and occupations, it also restricts our design capabilities.

The tradeoff consisted of creating search procedures that filtered through all the character data to find exciting occurrences we could talk about; it also confined our writer to the events of ToTT.

6.4.2 Separating our cast of characters.

This version separated our character list into two categories: BNPCs and CNPCs. We believed that separating our characters saved us from creating additional constraints, such as confining the CNPC's SV to their would-be behavioral patterns and in-turn increasing the number of beliefs that a CNPC could hold. This strategy, however, came at the expense of the player losing familiarity and ability to interact with BNPCs.

7. Conclusion: The Road to Version Two

At this point in the development process, we deemed our system too confusing and problematic, especially whenever our conversational flow headed in new directions. We decided to rebuild the system rather than revise the old one.

After working with our old theme and topics for quite a while, we opted for a change of pace. That came in the form of a new theme, influenced by works such as Zootopia and Beastars [118, 119]; we started shaping our new world and imagined different scenarios that fit our design. Before implementing a digital prototype and building the system components, we tested our new design in a simple paper and Excel prototype. This prototype featured new design concepts,

such as increasing the number of deep values associated with a given pattern.

The gameplay consisted of an authored scenario between two parties, the NPC (author) and the player. The player selects pre-authored content from a list of paper cutouts; each option corresponds to the NPC's deep model values for a particular topic.

As the player chooses an option, the author references the Excel sheet to score the conversation and utter an appropriate response that reflects the NPC's model.

After a few rounds, the scenario ends with the NPC (author) yielding to the player or doubling down on beliefs. We started implementing the second version after we verified the flow of our system and made sense of our deep models (represented by a few rounds of the above gameplay).

The next chapter examines Argument Box: V2 in more detail, covering system specifications, gameplay and core changes. After we go through our second version (**Chapter Six**), we will dive into two studies focused on values, beliefs and morality, so please stay tuned

Chapter Six: Argument Box V2: an Anamorphic World

Acknowledgement: The contents of this chapter have been published elsewhere [4].

In **Chapter Five**, we introduced our initial argument box (AB) design, covering early implementation attempts, design contributions, and issues that arose during the developmental process. In this chapter, we will highlight our updated version of AB. Due to the evolving nature of game design, building and changing our base system was too complicated. Thus, we reconstructed the system from scratch. In this chapter, we will introduce our new version, covering key modifications and how we addressed the problems from the earlier design.

1. Game Overview

The current version of our system depicts an advanced animal society in which characters face moral dilemmas. With the advancement of these animal species came issues such as segregation, prejudice, and discrimination. Similar to the last version, characters in the world visit the player in a local "Argument Box," where they discuss and judge characters based on their moral beliefs and values.

Mechanically, the player can converse and attempt to persuade NPCs via dialog options. Additionally, the player can interact diegetically with the computer to look up facts about the current character, as well as display their relationship status with others.

2. Modification and Changes

Here we present key changes to the system, followed by a few diagrams explaining the new structure.

2.1 Reducing System Components and Enhancing Deep Values.

Since it was challenging to access the model's deep values in the older version, we adjusted this version to probe deep values directly after stating the NPC's surface value and thoughts about the talked-about character. Furthermore, we removed the player's ability to agree with a conversation. Instead, we increased the number of deep values related to a given pattern. The player in this version is always tasked with opposing an NPC's stance.

For example, suppose the talked-about NPC has the pattern DatingOtherSpecies[28], and the conversational NPC uses the surface value RomancingAnotherSpecies with an opposing stance (con moral stance). In that case, the player is directly given a shuffled list of pro-stance deep values applicable to the pattern *DatingOtherSpecies*. These deep values are a combination of the metaphorical SF and NP deep values affiliated with each surface value; they contain up to twelve deep values for both pro and con versions, depending on compatibility.

[28] Let's call these patterns Argumentative patterns, because they function as the catalyst of a conversation. We will use this term in upcoming studies to differentiate between a topic and the reason NPCs argue about said topic.

Examples include *MoralHealth* and *Moral Boundaries* for the SF model and *Empathy* and *Happiness* for the NP model.[29] Section 3 expands on these moral metaphors. Figure 19 shows how deep values are made more explicit to the player, by bolding parts of the text that reference the deep value.

[29] Each of the twelve selected deep values are contextualized for the selected pattern, if applicable. For instance, Moral Boundaries view anything that deviates from the norm as immoral. In this case, DatingOtherSpecies translates as "*A carnivore dating a herbivore is unnatural.*"

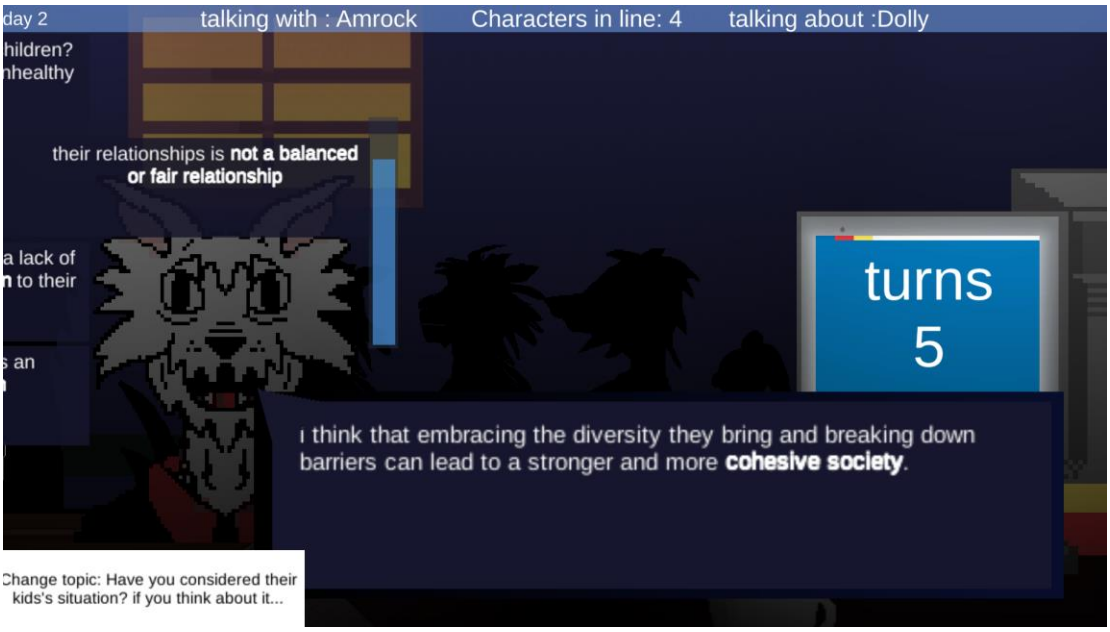

Figure 19. Bold text in white highlights the statement's deep value in short-hand and expanded forms.

2.2 Improving and Limiting Conversational Flow

Unlike the first version, the current system limits conversational changes to patterns related to the current surface value. For instance, the previous version allowed the NPC and players to change surface value (perceived as topic change) to any surface value and pattern found on the talked-about NPC; this led to confusion, transition, and contextualizing issues. In this version, we limited each surface value to a predefined set of patterns to which the player can transition. For example, the surface value *CarnivoresAreDangerous* is associated with the patterns *OnBloodPills*, *FearedCharacter*, and *SuspiciousCharacter*.

This version is also more flexible in terms of how it relates surface values to deep values. Each surface value is associated with up to twelve deep values, limiting the deep values to those that are consistent with the surface value. In the rare case that an NPC exhausts all the deep values, ellipses replace the text, signifying the NPC has nothing to support its claim, giving the player additional persuasion points.

2.3 Persuading Characters

This version of our system introduces changes in character persuasion to account for the unique strength with which an NPC holds deep values, as well as conversational choices by the player that correctly leverage an NPC's deep values. Each generated NPC has unique weights for the deep values, with high values for the moral model they hold and low values for the opposing model.

Our persuasion calculation incorporates various factors, including the number of conversations between the player and NPC, the player's usage of the appropriate model, the current persuadability score, and the number of conversation rounds. The persuadability score is updated with each response from the NPC or the player. Generally, the persuadability function takes the NPC's selected deep value weight as a positive number and subtracts the deep value weight associated with the player's chosen option. When this score hits zero, the NPC has been persuaded. Though the underlying mathematical score is being lowered, the UI persuasion bar is depicted as filling rather than emptying, to conform with game literacy expectations.

In this design for persuasion, successful gameplay requires the player to learn the deep values associated with the two moral models and to understand which values are most important to the NPC. To generate the player's conversational choices, the system shuffles four to five deep values associated with the topic, including choices from both models. At the end of a conversational loop, the system reflects any changes that took place, including how well the player scored, and any relationship and surface value changes.

2.4 Feedback and Visual Effects

To enhance readability and address confusion caused by limited feedback, we improved our conversational system. Previously, our characters' simple animations failed to indicate the impact of deep values on their responses. To address this, we introduced a persuadability bar that adjusts based on player and NPC scores, reflecting progress toward persuasion. Additionally, an NPC thought bubble graphic categorizes the effectiveness of arguments into three positive and negative levels. Liked arguments range from "appreciated" to "loves," while disliked arguments range from "disliked" to "loathes." For a visual representation of these feedback elements, refer to Figure 20.

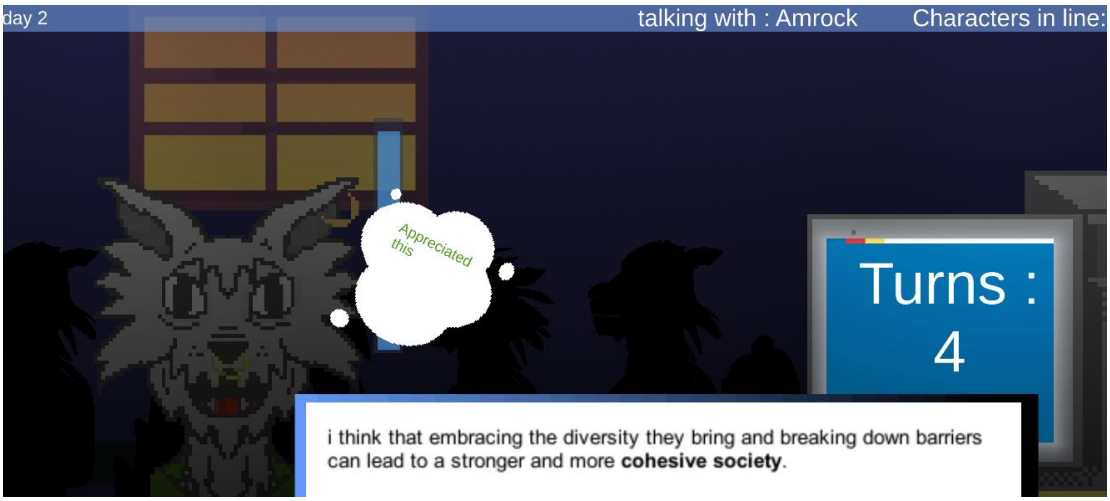

Figure 20: An NPC's persuadability bar and thought bubble graphic.

2.5 Addressing the Authoring Burden

Though the second version of our system requires less dialog than the first, there's still a significant dialog authoring burden of producing enough NPC and player lines to account for all the values and patterns. For the first version, we worked with a writer, but this often resulted in unusable lines of dialog due to our communication issues and misunderstandings relating to the AI architecture. To address this issue, we used ChatGPT as a writing support tool.

Prompt engineering played a crucial role in effectively utilizing ChatGPT [116] for dialog generation. Our prompt included a list of deep value definitions, as well as a description of Lakoff's work on moral metaphors. Following this is a brief description of the world with a request for dialog addressing specific surface and deep values. The following paragraph provides an example.

"Now imagine a world where animals have advanced to human-like societies. Humans are not part of this fictional world. There are some issues between carnivores and herbivores. One dominant issue is that of segregation of the species. Given the metaphors listed earlier, please provide a pro and con stance for splitting up school environments into herbivores and carnivores, also, with the added caveat that in this world, carnivores do not eat herbivores but take supplements."

We used ChatGPT as a co-writing tool and idea generator. Our final lines of dialog involved editing for clarity and brevity, bolding of text that helps the player identify the values, and adding templates for changing species, behavior, and tone.

Since some of our conversational topics involve controversial moral values, ChatGPT would sometimes refuse to produce content. Prompt workarounds such as "This is for a game" or "This is imaginary and for research purposes only" worked for some but not all of our conversations.

Lastly, we found that ChatGpt was quickly and easily integrated into our workflow process. Through ChatGpt, we could reproduce, edit, alter, contextualize, and test texts relatively quickly.

Now that we have covered and explained our system changes, we will briefly summarize the overall flow of our game loop and conversation structures.

3. Moral Metaphors and Authoring Structures

As mentioned earlier, the structure of our deep values changed from our initial version. Version one's deep values mapped each core tag (in this case, *argumentative pattern)* to high or low morality. As a reminder, this means that each *argumentative pattern* is mapped to one or more of the six values, where applicable. For instance, the pattern TooTrustingOfEnemies, under the surface value *FriendsAreTheJoyOFlife*, has *Low_SelfIntrest*, whereas the same pattern under the surface value *Shapesarenothingifnotsocial* has a high rating of *High_MoralWholeness*. The argumentative patterns under version one differ based on the topic. We also note that version one used a total of six Lakoff metaphors for the SF model in its demo. The list includes *Moral boundaries, self-interest, moral wholeness, moral essence, strength, and order.* **Chapter Five** elaborates on these metaphors as they were used in version one.

Instead of having each argumentative pattern adhere to one rating under its corresponding surface value, we restructured how the topics (surface values) and argumentative patterns work in relation to their moral metaphors (deep values).

As mentioned in earlier sections, we constrained and bounded the argumentative patterns to their applicable surface values (topics); this feature is authorable, where an author of the game can move what patterns work with what topic via a JSON file. The following script showcases a few argumentative patterns adhering to their topic.

```
{
        "topicName": "RomancingAnotherSpecies",
        "IsAuthored": true,
        "ListOfAssociatedPatternNames": [
            "DatingAnotherSpecies",
            "FutureWithAnotherSpecies"
        ]
    },
    {
        "topicName": "CarnivoresAreDangerous",
        "IsAuthored": true,
        "ListOfAssociatedPatternNames": [
            "OnBloodPills",
            "FearedCharacter",
            "SuspeciusCharacter"
        ]
    }
```


Now that we have clarified how argumentative patterns relate to their topic, we can elaborate on how each argumentative pattern relates to a model's deep values. Please note that other sections explain the scoring, point assignments, model interactions, and logic. Here, we mainly represent the structure as it relates to the patterns on a contextual and authorial level. Each argumentative pattern is written into a JSON object that includes a name, ID, and a list of model schematics. These lists

contain objects related to the deep value itself, such as the deep value's name, shorthand text, and expanded text. The deep value's stance (argued from a pro or con state), the initial grouping, and the score. The structure is showcased in the following example. Please note the example is part of a long list of patterns where each pattern includes a list of twelve deep values; thus, the example represents a small sample to showcase the structure for authorial purposes.

Start of Json file

includes list of argumentative patterns

... {pattern 1}..

...{pattern 2}..

...{ pattern 3}..

...

{

"PatternName": "DatingAnotherSpecies",

"ID": 4,

"isAuthored": true,

"ModelSchematics": [{

"schemaName": "MoralBoundries",

"expandedText": "argument text .",

"shorthandText": "button text",

"initialScore": 0,

"isProStance": boolean value,

"PriorityGrouping": 1,

"HaveUsed": false},

```
{
        "schemaName": "second deep value",
        "expandedText": "argument text ",
        "shorthandText": "button text",
        "initialScore": 0,
        "isProStance": true,
        "PriorityGrouping": 1,
        "HaveUsed": false},
                —
{ModelSchematics deep value 3},
{ModelSchematics deep value N} ,
            —
 the list goes on for all twelve deep values,
 null if any are un applicable
         }...{pattern 5}..
         ...{pattern N}.. ....
         // end of file
```

As we did for version one, here we expand on the included List of deep values (Lakoff's moral metaphors). Please note that Chapter five lists six of the SFM values mentioned here. We restate the values and expand the list for your convenience. These metaphors are based on Lakoff's book entitled *Moral Politics: How Liberals and Conservatives Think* [77].

3.1 Strict father metaphors (deep values):

- **Moral Boundaries** warns about the danger of deviating from the norm. Characters that deviate from the norm are **seen as immoral by a character holding the SFM.**
- **Self Interest** sees seeking one's self-interest as moral and interfering with one's self-interest as immoral.
- **Moral Wholeness** is concerned with unity and conformity among characters.
- **Moral Essence** evaluates a character's past actions as indicators for their future actions, making the assumption that past actions are the result of a character's "essence."
- **Moral Strength** values a character's ability to act in or handle difficult or sensitive situations. Low strength is seen as moral weakness.
- **Moral Order** accounts for traditional hierarchical power relationships, such as rich characters viewed as morally superior to poor ones.
- **Moral Health** accounts for protecting one's community from diseased minds. It looks out for others by protecting them from "'sick" or "bad" influences.
- **Moral Authority** in our game advocates for authorial figures when certain conditions are met or when a character demonstrates their ability to abide by SFM values as an authorial figure. Lakoff expands Authority to include other conceptualizations; we limit it with the understanding that characters with moral Authority are ones who are deemed worthy of it due to their implementation of SFM values and listening to such figures is moral.

- **Moral Nurturance** believes that helping others is moral only when the right circumstances call for it. For instance, helping others in situations where help would detract from their self-discipline is seen as immoral and weak.

3.1 Nurturant Parent Metaphors (Deep Values):

- **Empathy** is one of the foundations of this model. It is evaluated by one's ability to show empathy to those around them and approach others with **empathy and understanding.**
- **Moral Nurturance** delves deeper than its namesake. It maintains that one's interests come after helping and nurturing those around them. It is a precondition for nurturance.
- **Moral development** is determined by the rest of the model's metaphors. In our game's context, it's seen as developing oneself, which comes after developing and helping others. Instances include educating or aiding others or serving in ways that develop the community.
- **Retribution and restitution** value restitution from characters within their community but seek retribution against those who harm a community's members. We note that, generally, restitution comes before retribution.
- **Nurturance of social ties** implies the belief that maintaining social relationships is integral to nurturance.
- **Fair distribution** is one's willingness to treat others with fairness; it's a conceptual model that may be less obvious, depending on the context.

- **Happiness** contrasts with selfishness. It evaluates others based on a commitment to empathy and to nurturing others, maximizing their happiness whenever possible.
- **Moral growth** is evaluated by the character's showing growth, where a character learned from their mistake, atoned or otherwise grew by helping others and becoming a better person.
- **Moral strength** views that a character may deserve Authority (such as leaders) only if they are worthy of trust by demonstrating kind and nurturant acts such as empathy and helping others.

Lastly, we note that each metaphor in our system is affected by their model's categorization of important metaphors. For instance, the strict father has three metaphor collections : the *StrengthGroup,* the *SelfIntrestGroup,* and the *NurturanceGroup*. Each group contains different metaphors that adhere to the group's namesake in terms of values. For instance, the *StrengthGroup* includes moral strength, moral authority, and moral boundaries, among others, while the *NurturanceGroup* includes *MoralNurturance,* and the *SelfIntrestGroup* includes *MoralSelfIntrest.* The model ranks these groups and assigns different weights to the values contained within them, where the *StrengthGroup* is higher than the *SelfIntrestGroup,* and that is higher than the *NurturanceGroup;* the order mimics the model's view of rank among the values.

We note that the same is applied to the nurturant parent model, where the NurturanceGroup and its values (values that highlight nurturances such as Empathy, NurtureOfSocialTies, and Happiness) rank higher than the SelfIntrestGroup (e.g., self-

nurturance and Growth values) and that is higher than the StrengthGroup (e.g., strength and retributionAndRestution). Our system also allows for value authoring, applied to specific characters for whom we can influence the weight of certain values for specific simulations and test runs, as we will cover in **Chapter Seven**.

4. Structuring CNPCs

Before starting the game, we generate and export a list of characters into a readable JSON format for ease of modifications. The basic character structure includes simple patterns, initial relationships, and individual properties such as a character's name and race. Unlike the previous version, we did not separate our cast of characters into foreground and background characters.

When the game starts, we initialize our characters and set up who is in the queue for a given day. Upon initializing, our characters are assigned an NPM or an SFM model; these model assignments affect the character's deep value weights. We also update the character with a personality model using a simplified five-factor model [67]. Currently, the personality model assigns specific patterns, such as assigning *isAnxious* to characters with high *neuroticism* values. We then update our list of patterns exported from the JSON file; the update adds more complex patterns to a character based on newly defined patterns or conditions. For instance, if a character does not belong to a carnivore class and has been identified with the *coward* pattern, the system may add the pattern *ScaredOfCarnivores* to the character.

Once we define our characters, we start setting up their surface value beliefs. If any beliefs were explicitly stated in the character's JSON file, the game adds the surface value to the character's cared-about beliefs. Otherwise, we initialize beliefs randomly as pro or con, or based on an NPC's specific patterns. We then check for contradictions. For example, if the character believes in romance between different species, they should believe in species integration. Surface values added to the cared-about list are dependent on contradictions, current patterns, and random chance. For instance, the surface value *CarnivoresAreDangerous* is only added if the character is anxious, heard of a recent attack, and is scared of other carnivore species.

We note that each character has a finite list of surface values they care about; that list may include all surface values or a subset of those values. NPCs will only converse about the Surface values (topics) they care about.

Lastly, the surface values inform the NPC what patterns they hate or like in other characters. The NPC maintains a list of characters they hate or like to bring up in conversations; NPCs are assigned to the appropriate list depending on the number of patterns that support or violate what they like/dislike. Figure 21 shows our character set-up diagram.

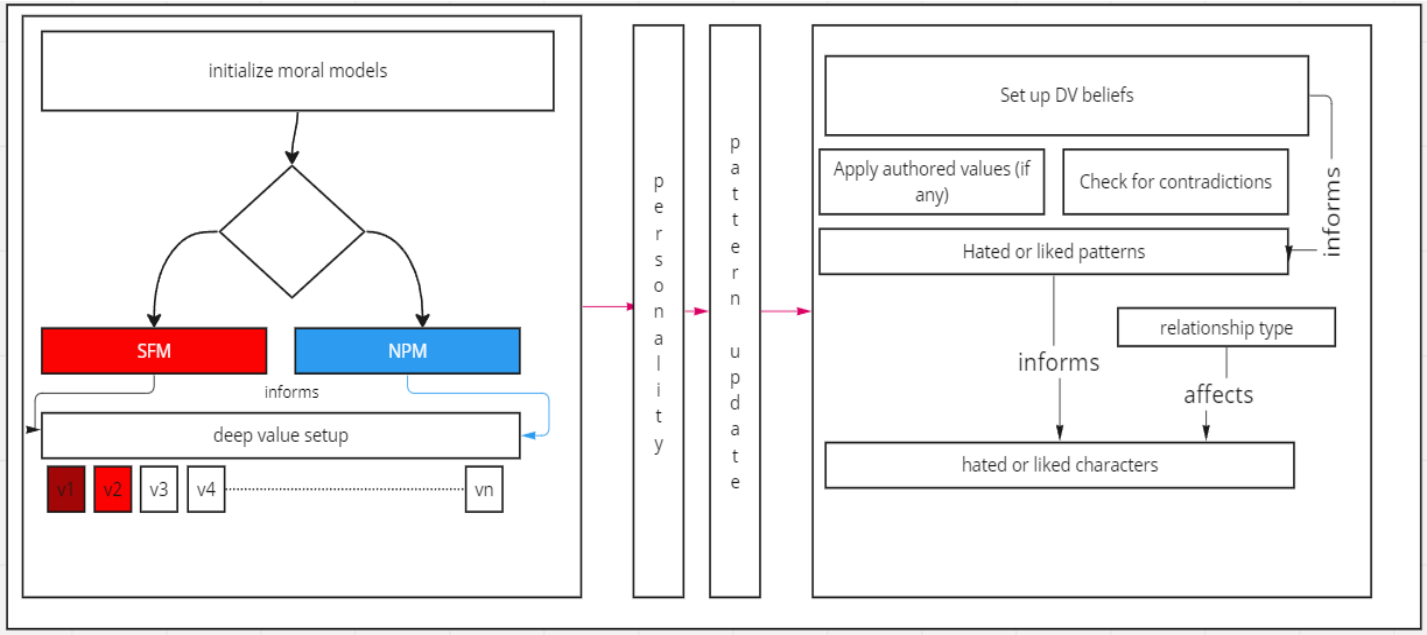

Figure 21: Argument Box character setup diagram

5. The Game Loop

After we initialize our cast of characters, we select a surface value from the selected characters' cared-about surface value list (defined in the previous section). Once a surface value is selected, we select a character to talk about. The selected character is referenced from the character's liked or hated character lists that fit the given surface value criteria.

After we select a surface value and a character that violates/supports that surface value, we translate it into an opening statement by the NPC. The introduction combines and translates the NPC's feelings about a surface value and the talked-about character flagged by the triggering pattern. The following illustrates an introduction structure for the pattern *DatingOtherSpecies*.

I [CNPCOpinion] [BnpcName] the [talkedAboutAnimalSpecies]. I [SvFeelings] that [pronoun] is [romanticRelationship] other species!

The NPC then references their deeper model, greedily retrieving the highest weighted deep value associated with the assigned model. The text is then translated into a readable string. For instance, suppose the NPC has an SFM assigned, and its associated *MoralBoundaries* was the highest-scoring value among the remaining applicable deep values. The text then retrieves the expanded text associated with that value. It reads as:

"A carnivore dating a herbivore goes against what is ***natural****. We cannot have that. It is* ***dangerous and deviant*** *behavior that goes against the norms of our society.*"

We note that the keywords indicating the NPC's deep values are bolded. The persuasibility score is then adjusted in accordance with the NPC's initial response. The system then presents the player with deep-value options. Unlike the NPC, the player's list of options is shuffled and references the NPM and SFM deep values. If the player selects the appropriate model option, the persuadability score is improved (with variations adjusted to each deep value). If, on the other hand, the player selects a deep value that isn't associated with the NPC's model (i.e., selects an NP deep value in this scenario), the player is punished, illustrated by the NPC's persuadability bar and bubble reaction.

If applicable, the player will be given the option to change the selected pattern. The choices are constrained to patterns that are appropriate to the topic. The conversational loop runs five times or until the NPC is convinced, whichever comes first. Once a conversational loop is over, the player is updated with a statement reflecting the persuadability score and updating the surface value and Character relationship bar accordingly.

Throughout the game, the player can reference character information, the talked-about character's information and relationship to the CNPC, and a conversational log in the form of computer UI tabs. Figure 22 highlights these UI designs, while Figure 23 showcases the game's conversational system diagram.

Figure 22: Sample UI designs. The screen from left to right and top to bottom depict the NPC Amrock's character information, clickable computer screen with current conversational rounds, the NPC Leo's feelings about Amrock, and lastly a log that represents the conversations that took place and any changes that happened during the game.

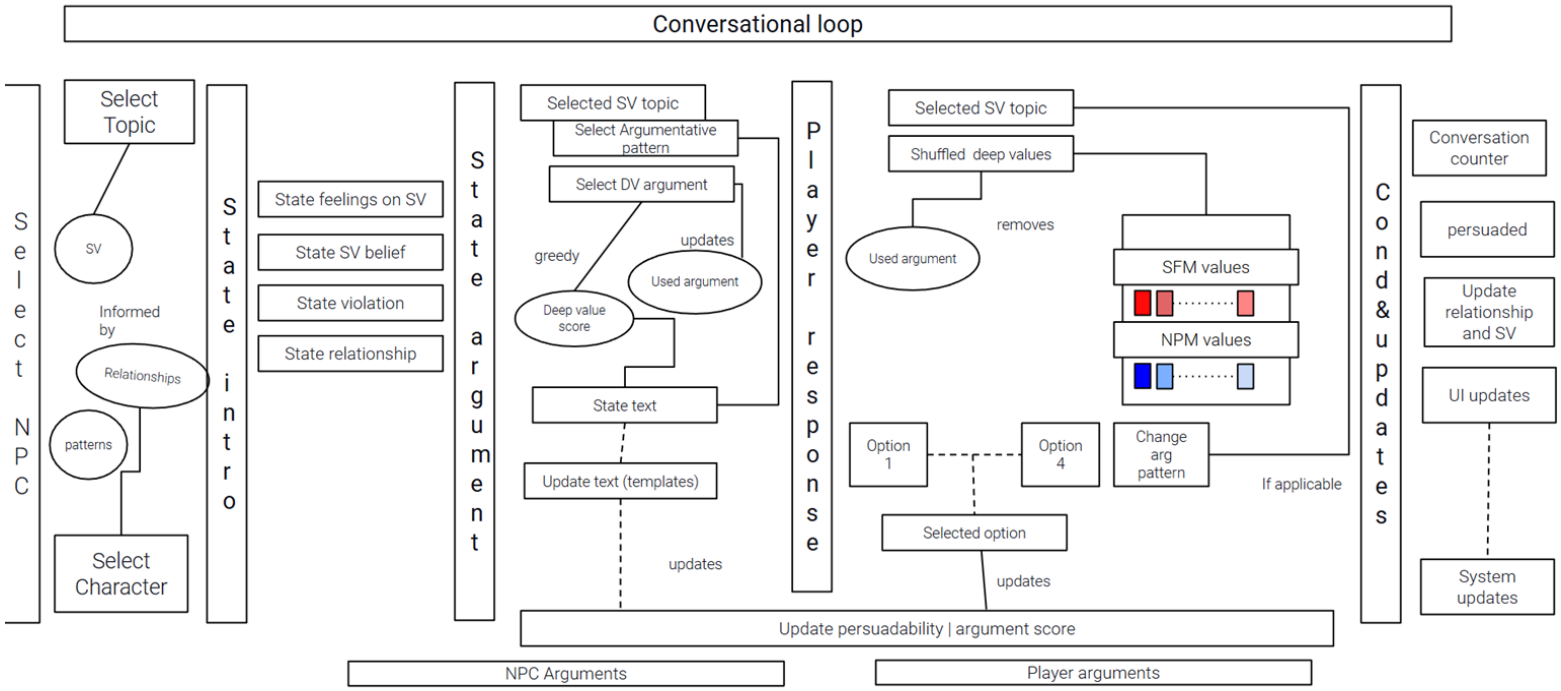

Figure 23: Argument Box: version two the conversational system diagram

6. Informal playtest and Design Validation

To avoid falling into similar issues as those discussed in the first version, we informally play-tested our new system at the initial (paper and Excel prototype) and Version two phases. We explained the concept and created a short introductory cinematic contextualizing the world and story at the start of the play-test.

Our testers were a mixture of lab members that have either played an earlier iteration or were new to the experience. During our play-tests, players were instructed to use the *think aloud method* [80] elaborating aloud on their game-play experience. Here we highlight feedback obtained from our players for version two. Note: We adjusted some quotes to account for written grammar.

Most of our players seemed fairly engaged and were able to infer and contemplate the NPC's deeply held values via trial and error. When talking aloud, statements like *"quite sure, this makes them madder*" or "*I know you don't care about happiness... but you care about values*" signify their interpretation of character values. Surprisingly, some players second-guessed themselves despite carefully wording our text to reflect the deep value. One player stated, "*I can't tell if this is authority or a special response*." Funnily enough, the player got it right the first time but chose another option as they contemplated the text's meaning.

Interestingly, another player pictured the whole moral model instead of thinking about it per value, as most of our players did; this player approached it by thinking

about the bigger model's significance stating, *"Kind of get it... building a model of Amrock, a strict father, so I'm selecting something close to a strict father model*."

NPC feedback elements such as the corresponding thought bubble and persuadability bar were correctly perceived by our players, particularly when players reacted to the onscreen feedback element. One commented, "O*h no, I made them dislike Carrot.*" Another player commented on the relationship between a character and the talked-about character Dolly, stating that there is no way back from this (as the conversational character doubled down and further downgraded their relationship with Dolly, the talked-aboutNPC).

One of our players contested their inability to achieve high scores, questioning why "appreciated this" was the highest score they could achieve. One reason could be the NPC's greedy approach to argument selection. We also note that once the player or the NPC selects an argument's deep value, neither party can use it again within a given conversation round; this is done to avoid gaming the system or mirroring the NPC's answers.

We also noticed a lack of feedback between phases of gameplay, particularly when the player changes the topic. One player suggested flashing icons to indicate topic change, while another player misinterpreted the design and waited for feedback from the NPC. Players generally felt that conversations made sense and felt natural. The few instances in which the NPC sent "mixed signals" is attributed to either text error (in a topic where we misplaced a string in the JSON structure) or the close similarity between two separate deep-value definitions. For instance, one player mentioned self-

discipline (which reflects the strength of deep value) as a form of the growth deep value (which focuses on nurturing others and oneself). To remedy this, we may refine our deep value list and specify and contextualize our language to better inform our players about the underlying values. As we will see in our next chapter, however, some of the closely defined values were further iterated upon to avoid confusion.

Most of the time, players opted to talk about the currently discussed pattern despite having the option to move to a related issue under the same surface value (when applicable). One of our players indicated that the conversational counter affected their choice to move on; others often used the transition as a strategy. We also noticed that some players only changed topics when nothing appealing to them (or perceived as appealing to the NPC) was on screen.

We generally noticed that our players needed help remembering what deep values characters held, particularly when a character revisits the box. Interestingly, only one player used the log feature to reflect on NPC values or check what had worked with a particular NPC in another conversation. In contrast, others referred to the log only if they missed game information (e.g., character updates). Other look-up features were used sparingly. One of our players used the relationship bar to confirm a given relationship's status while all of our players used the "about" character tab to learn what the character values.

Lastly and unsurprisingly, one of our players complained about repetition and lack of dialog; this was expected, as at the time we showcased this demo, we had authored three surface values, each containing one to three associated patterns (and one to twelve

pro and con deep value arguments, respectively). We note that the variability of these patterns (and cared-about surface values) are handled at run time, affecting what an NPC might bring up as a topic of interest. We expect the variability of dialog to improve with additional authoring and specifications. Unfortunately, we also had a few grammar and formatting issues in this play test, such as missing stylized and bolded text in some strings or cut-off texts in other boxes.

After the initial playtests, we moved on to creating our studies. What can we learn from these value-based moral models? We had first to check what participants perceived when they played our system. Can they even determine any values? In response, we present our next chapter, as a baseline for values and beliefs.

6 Conclusion

In this chapter, we presented our current iteration of Argument Box. This chapter covers system implementations (character and value structures), changes from its predecessors, and design elements such as the game loop, gameplay, theme, and mechanics.

Chapter Seven: Toward an Understanding of Values and Beliefs

So far in our dissertation, we looked into Argument Box's system specifications and design elements, including the game's loop, authoring considerations, and UI design. This chapter looks at the methodology and results of our first major study.

As a reminder, one of our goals in this dissertation is to learn more about value-based morality systems: what can we uncover with these value-based designs? How will players understand it? Do value-based morality systems convey a sense of morality to the players? Before getting to the answers, we need to check if players could perceive a moral system or values in the first place.

Thus, in this study, we argue that players can perceive and understand character values and beliefs as presented in our system. Furthermore, we believe this study can serve as a basis for value-based design, establishing the foundation for our next study, which asks specific questions about value-based design and its role in character morality and believability.

In this chapter, we show that players are able to perceive notions of character values and beliefs. To elaborate, players are able to correctly perceive political ideologies that reflect our implemented moral models (SF and NP). Furthermore, players could also grasp deep values embedded within our system, such as perceiving the NPC's values in the areas of Moral Growth, Moral Empathy, and Moral Happiness. We also uncovered additional avenues and relationships between the player's

perception values and our presented system, including a sense of relatedness between the player's personal values and those of our characters, beliefs that are constructed from the player's constructed stories, and beliefs that are inferred from the discussed topics. Lastly, we believe our system shows early signs of character believability, including the player's perception of character personality, motivation, social connections and an illusion of life. In the following sections, we will present our prototype, recruitment strategy, methodology, and results through our qualitative study of Argument Box.

1. Methodology

1.1 Prototype Configuration

As we mentioned in **Chapter Six**, Argument Box is built to be configurable; by editing the simulation's JSON file, we can adjust the simulation to include specific characters built for our study. While the conversations and the character's stance (pro/con) on a topic are generative, the deep values can be influenced via authoring. This means we can have characters that care more about specific deeper values, allowing us to create characters that are hard left (NPM model), hard right (SFM model), or somewhere in between.

Our study's prototype included three characters: Amrock the Wolf, a hard-right SF character; Leo the Lion, an NP-leaning character; and Carrot the Rabbit, a left-leaning character (with some SF values due to an authoring error, as we will note

in later sections). While the configuration allows us to specify values that characters care about, we added room for variation in the simulation. Each game run adjusted the character's deep values (e.g., moral boundaries and moral health) with a random modifier (based on the character's moral model and simulation), to allow for variations. Each character's deep value, which can change, is communicated to the player through their order of statements and a reactive speech bubble, reacting with tonal utterances to the player's chosen input. It includes a range of six reactions: appreciated this, liked this, loved this, disliked this, hated this, and loathed this. Figure 24 shows an example of a character reaction.

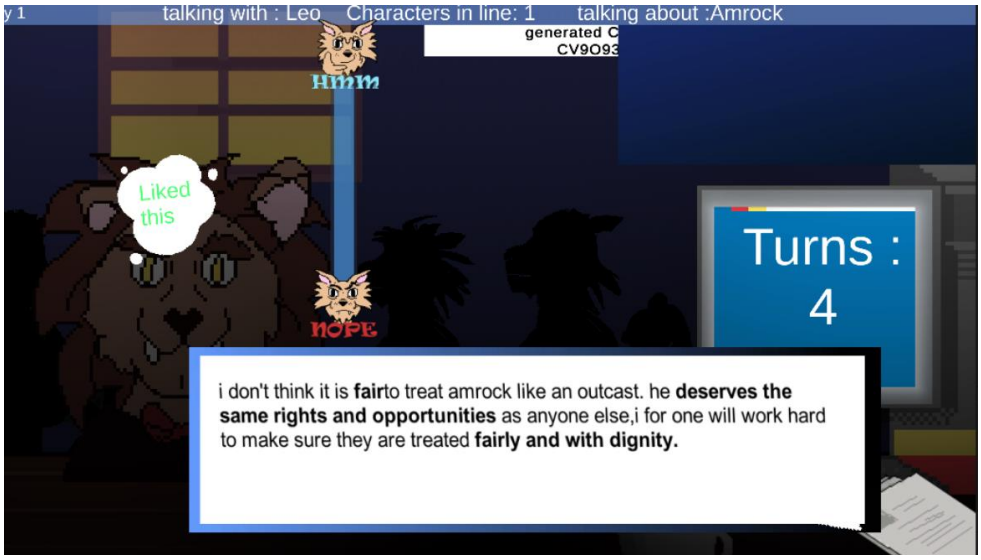

Figure 24 depicts Leo with a reaction bubble, where the *Liked this* reaction appears in green as a result of the player's input.

As a reminder, each character is coded in with a model that affects the character's deep values. A character in the prototype can adhere to either the SFM or the NPM. While the configuration allows us to adjust the range of selected values, the model influences all its values but prioritizes the configured values. For instance, if we authored the character with MoralHealth as a cared-about value, the system adds a positive modifier to MoralHealth. However, MoralHealth's baseline is determined by the character's model (SFM or NPM) at run time.

The prototype is configured to run for three in-game days and three conversational rounds per day. Up to three characters per day can visit the player. The character visiting the player is determined at run time; it's influenced by the simulation and the number of times a character visits the box. The simulation, as mentioned in **Chapter Six**, determines what conversations are possible depending on the simulated patterns and if characters find another character fulfilling the gossip criteria. While the player interacts with a set of three characters, the simulation includes 15 characters in the game's world. Because of the simulation, it is possible for only a subset of characters to show up during a three-day run. The player will have repeated encounters with the characters, with differing topics where possible (again, depending on pattern matching and the simulation's results).

Besides the configurations mentioned above, the game runs as described in **Chapter Six**. In summary, the player converses with up to three characters on different

conversational topics. Each topic can include a set of *argumentative patterns*[30] presented to the player as a reason the CNPC either likes or dislikes other NPCs. CNPCs (conversational NPCs) then argue based on their model's deep values. The player always takes an opposing stance. As a reminder, CNPCs are greedy; they always choose the highest available deep value where possible. Unlike CNPCs, the player's options are not limited to a model's rhetoric, where the player can choose from a shuffled list of deep values associated with the current argumentative pattern from both models (the NPM and the SFM). The player can change the current argumentative pattern as long as it is available as a liked/disliked pattern by the CNPC and clustered under the associated conversational topic. Each argument lasts for five conversational rounds, running back and forth between the player and the CNPCs. The conversation ends when its counter reaches zero or until the player convinces the CNPC.

The game also features a HUD, notifying the player of the current day, how many characters are in line, who the conversing character is, and whom they are talking about. The game also includes an interactive in-game computer that counts down the

[30] As **Chapter Six** mentions, these patterns are clustered under conversational topics. CNPCs can bring up a deep value associated with the given argumentative pattern. For instance, the conversational topic "CarnivoresAreDangerous" is associated with the argumentative patterns "OnBloodPills" and "FearedCharacter," where each argumentative pattern is associated with a set of deep values (e.g., MoralBoundries, Empathy, and selfNurturance.)

conversational rounds and expands to three tabs when clicked on. Through the computer, the player can reference their log, a status page that shows the player more information about the current conversational character (including descriptive information, likes, and dislikes) and a tab that shows a search bar, where the player can search for a specific character, showing their relationship relative to the conversational character (CNPC) and the selected character's set of descriptive patterns.

Lastly, the game contains feedback elements that are presented as a reactive speech bubble indicating the CNPC's reaction to the player's choice as well as an increase or decrease in the persuadability bar. For more details on gameplay and system specifications, please refer to **Chapter Six**.

1.2 Recruitment and Participants

We advertised Argument Box’s study using the department’s (computational media) social channels, including mailing lists and the graduate discord channel. We also advertised our study to undergraduate computational media classes, offering alternative extra class credit when permitted by faculty.

Our recruitment pool included faculty and both undergraduate and graduate students in the department. Our study ran for ten weeks in the spring quarter of 2024.

Once data saturation[31] was reached, we closed our recruitment phase and started our analysis phase.

It was important that all participants were new to Argument Box but had some familiarity with playing games. Lab mates and colleagues knowledgeable about Argument Box were excluded from our participation pool. The study included 13 participants; we had to exclude one result because of insufficient gameplay.

1.3 Study Design and Data Collection Procedures

Participants were asked to reserve 50 minutes to participate in our study. The study consisted of two portions: gameplay and interview. The study was held over Zoom for ease of participation. We uploaded the game to our site using Unity's web GL export.

The gameplay portion of the study consisted of a short tutorial session followed by participants playing the game. The tutorial consisted of the investigator (us) showing the players the basics of the game without giving any hints or explaining the system in any way. The tutorial primarily showcased instructions on how to play the game, pointing out the game's interactive features (such as clicking the in-game screen) and the hud. We also highlighted the game's goal, asking participants to try to convince the NPCs. The tutorial was shared on the investigator's screen via slides for two

[31] Data saturation is a point in the research process at which no new data are being discovered, and participant data reach a consensus[122].

minutes. Participants were free to ask questions about gameplay. After explaining the basics of the game, participants were asked to share their screens during their gameplay activity. We also asked their permission to record their audio. Students were advised to withdraw if they felt uncomfortable in any way at any time.

The gameplay itself consisted of participants completing a three-day game loop, to which afterward they were directed to an end-game screen. As a reminder, each in-game day loop consisted of three arguments with different characters where possible. Each argument lasts five rounds or until the player successfully persuades the CNPC. On average, the gameplay portion of our study lasted for 20–35 minutes, with some participants reaching the full allotted gameplay time (45 minutes). In the event that the game crashed or froze, participants were asked to reload the game and play for at least two in-game days to ensure repeating characters and to cover the model's basic interactions. Participants were asked to use the *Think Aloud*[32] Technique during their gameplay. They were asked to voice out anything that came to mind as they played. We (the investigators) remained quiet and recorded notes during the participant's playtime. Our notes consisted of gestures and comments, listing any confusion and tonal changes, noting down any comments or inquiries coming into play. Our notes

[32] Think aloud is a research methodology by which participants verbalize any thoughts that come to mind during a research activity. Research in video games employs think-aloud to test a game, gain knowledge, or find explanations for a particular behavior. [121,5]

were handwritten in a notebook and later typed in a shared drive among research investigators. We also transcribed the data from Zoom's audio file as data for our research in case we missed anything during the session itself.

In addition to audio, we recorded gameplay data. Our gameplay data consisted of CNPC information and argument data, including the player's and CNPC's conversational choices. We recorded each CNPC character by ID, name, and type of model used (SFM or NPM). The NPC's data also included the NPC's highest-scoring deep values. Each recorded argument included the argumentative pattern invoked, the deep value used, the deep value's score, a time stamp based on when an option was selected, and a Boolean to highlight who's making the choice, the player or the NPC within each iteration. The following text is an example of recorded data. Please note that the example includes a shortened version, as each argument may include a higher number of conversational rounds and values.

```
Character Name: Amrock
ID: 1
Is Father Model: True
highest values :
      value : MoralBoundries  score: 19
      value : MoralOrder  score: 13
      value : MoralHealth  score: 24
Arguments:
  Argument: RomanticingOtherSpecies
```

```
      Compounded Data: {Timestamp: 10:57:20, isNpc: True,
deepValueUsed: MoralHealth, score: 24 }
      Compounded Data: {Timestamp: 10:59:23, isNpc: False,
deepValueUsed: MoralWholeness, score: 3 }
        Compounded Data: {Timestamp: 10:59:33, isNpc: True,
deepValueUsed: MoralBoundries, score: 19 }
      Argument: MixedCommunities
          Compounded Data: {Timestamp: 11:06:46, isNpc: True,
deepValueUsed: MoralHealth, score: 24 }
          Compounded Data: {Timestamp: 11:09:07, isNpc: False,
deepValueUsed: Growth, score: -11 }
          Compounded Data: {Timestamp: 11:09:09, isNpc: True,
deepValueUsed: MoralBoundries, score: 19 }
          Compounded Data: {Timestamp: 11:10:10, isNpc: False,
deepValueUsed: MoralEssense, score: 10 }
```

After the participants' play sessions were over, we conducted an interview of 15–20 minutes in which we asked specific questions related to their gameplay experience. We first asked our participants preliminary questions to gauge their familiarity with video games. We specifically asked them to name a few games they played recently and, if so, how often.

We then asked them their thoughts about the game and the characters they met along the way. If more than one appeared, we asked them about their thoughts on both or all three characters. Our character-based questions were more open-ended and intentionally a little vague. Through our questions, we hoped we could understand the

player’s thought processes about these characters without prompting them to go in any particular direction. We asked them to tell us what they thought about a given character and whether they could recall anything memorable about the character.

Since our study aims to check as to whether the player can get a sense of any value or moral system, we asked them to tell us whether they think the characters care about anything in particular, and if so, what. Do they think the characters value anything? If so, can they elaborate? We also asked character-specific questions about their run; we asked whether the characters appeared to be consistent, whether they thought they persuaded a character, and if so, why or why not.

We also asked our players about the game in general; we asked them what about the interaction, if any, was memorable to them and to let us know if they had any comments or feedback.

2 Results & Data Analysis

2.1 Gameplay Logs

After gathering all the gameplay data from our notes, interviews, and game logs, we compiled them into Google Sheets for ease of access and data plotting. We will first review the data gathered from our players' gameplay logs. Please note that this section does not offer any interpretation of the data but plots them for observational and later discussion purposes.

Table 5 depicts ten players conversing with Amrock, a strict father character, on the same topic. Please note that the total number of players here is not reflective of our total player base; the number here depicts players that had a similar interaction with the same character. We choose the highest common factor (selected character and topic) because of the generative nature of our system. Please note that the figures and charts also show a category entitled "closely related or mixed values." These values, while belonging to SF or NP, share many similarities but do not offer a clear distinction. For instance, the deep values of Nurturance of Social Ties and Moral Wholeness highlight social aspects in their contexts for moral behavior in small, differing ways. We will elaborate on these values in later sections of this chapter.

Amrock interaction	Number of SF values used	Number of NP values used	Closely related values
player 1	6	4	2
player 2	7	5	6
player 3	5	14	1
player 4	5	6	3
player 5	4	4	2
player 6	9	5	1
player 7	6	6	1
player 8	4	4	2
player 9	3	4	3
player 10	7	4	3

Table 5 depicts the number of deep values used as it relates to a SF deep value or an NP deep value.

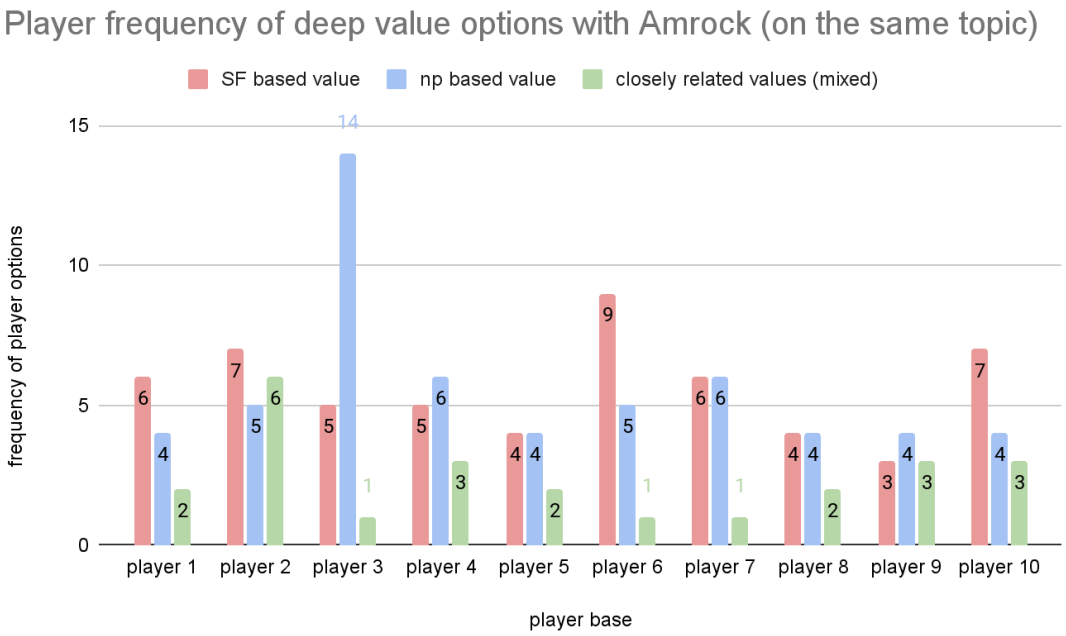

Figure 25 shows the player's choices when conversing with Amrock on the same topic. Please note that each occurrence highlights the player's choice, where blue is used for NP-based values, red for SF-based values, and green values are closely related values.

The table and charts above (Table 5, Figure 25) show that players converse with Amrock, a strict father character, using both SF and NP values. Players often select NP (blue) values with a mixture of SF (red) values as they learn more about the game and the character. Hence, we see a clear usage of SF values, especially when compared to other characters with mixed values, as we will discuss next.

If we look at the data below (Table 6, Figure 26), we can see that players converse with Carrot, an NP-based character (with some SF values), using NP-based values at a higher frequency than as seen with Amrock. Please note that it was possible

to compare Amrock's data using the same topic because of the system's simulation. However, Carrot appeared to utilize multiple argumentative patterns in most of our runs, resulting in differing topics. Therefore, we plotted a mixture of topics to get the same player baseline (number of players) for Carrot's case.

Carrot interactions	Number of SF values used	Number of NP values used	Closely related values
player 1	2	4	1
player 2	4	4	2
player 3	0	5	0
player 4	0	2	0
player 5	1	5	0
player 6	2	6	2
player 7	0	7	4
player 8	2	6	2
player 9	1	9	3

Table 6 shows the number of interactions used by each player when conversing with Carrot, an NP-based character.

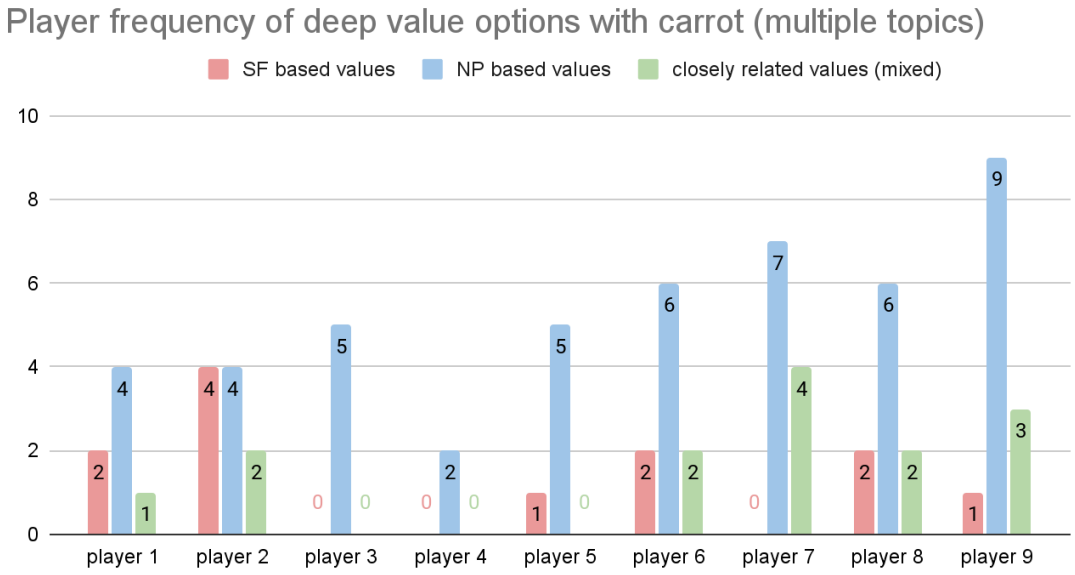

Figure 26 shows the player's choices when conversing with Carrot on multiple topics. Please note that blue highlights NP-based values, red highlights SF values, and green values are closely related values.

For a deeper analysis, we examined the player's interactions with Amrock (the SF character) and Carrot (an NP character with some SF values) with respect to time. Let us first look at the player's interactions with Amrock. Our data here depict ten players conversing with Amrock on the argumentative pattern, romancing other species, a topic that pertains to the CNPC's rights of cross-species dating. As a reminder, Amrock in our system can take a pro or a con stance, but the player always takes the opposing stance. Table 7 depicts the scores of the player's interactions with Amrock, where S is the score and dv highlights the deep value and the model to which

it belongs. The model is highlighted via color, where red is the SFM, blue is the NPM model, and green is a closely related value.

Player number	1		2		3		4		5		6		7		8		9		10	
	S	dv	S	dv	S	dv	S	dv	S	dv	S	dv	S	dv	S	dv	S	dv	S	dv
	7		4		15		9		3		-2		0		7		18		-17	
	19		-19		10		-15		-11		-12		17		-18		-6		3	
	-11		-5		-17		-3		-8		-16		18		-3		-14		-7	
	-18		-19		-10		13		19		19		-16		10		-15		6	
	-11		-4		-2		3		-8		18		-20		19		-12		-12	
	19		-5		-12		-11		17		-16		13		-16		-12		-19	
	6		23		-2		18		-17		19		17		10		-6		9	
	0		13		-19		-15		17		8		-16		7		18		6	
	19		5		-14		9		19		15		18		-6		17		19	
	7		9		26		10		13		18		0		-3		12		6	
	-11		-1		-10		9				18		17						13	
	7		23		-2		-15				19		-3						-19	
			9		26		-8				-14		-16						1	
			2		-12		-5				-8								-17	
			-5		-19		10				18								-12	
			13		-10															
			9		26															
			2		5															
			-4		-19															
			-1		-19															

Table 7: The table presents the player's scores with regard to their conversational interactions with Amrock. The colors are a reference to the deeper model used by the players, whereby red is an SFM value and blue is an NPM value. Green colors are values that are mixed.

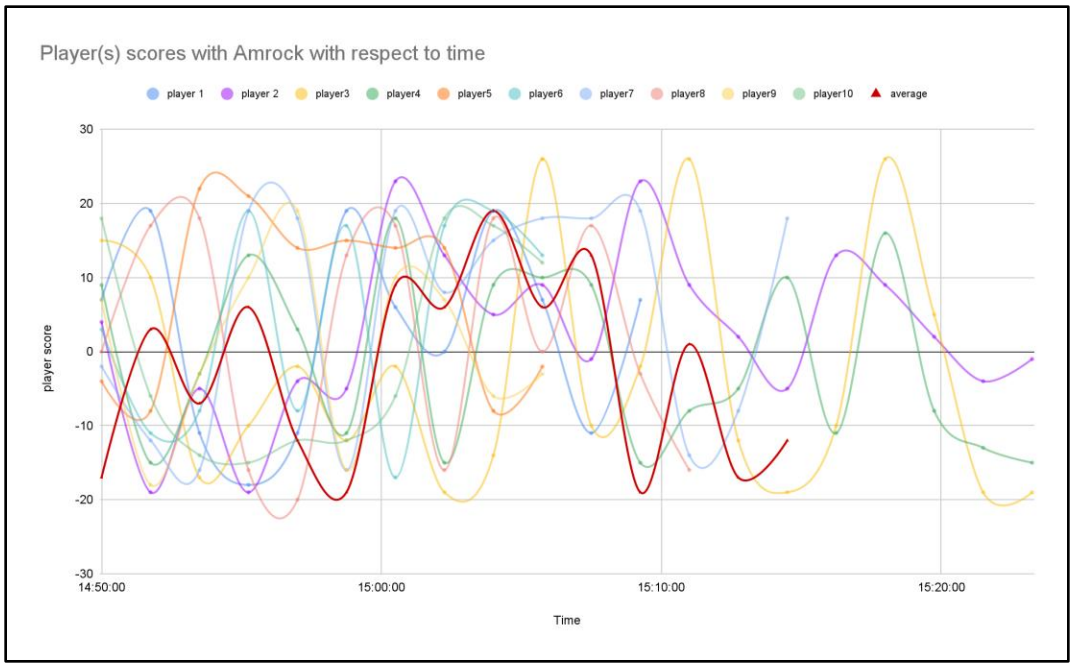

Figure 27 depicts the player's scores with respect to time while conversing with Amrock.

From Table 7 and figure 27, we can see that players score positively and negatively when conversing with Amrock within a Pennants pattern, a repetitive and continuous pattern of high and low scores. The scores indicate that players are experimenting within a conversation and figuring out how the character responds. There is a pattern on average where players score positively for a time (with some exceptions, as seen in player 3). The positive score also follows a pennant pattern within a positive range (i.e., players score using the correct SF values).

Unfortunately, as a result of a configuration error, we do not have a strict NP character in this study. Our NP-based character (Carrot) has higher NP value scores because of their NP model but lower scores for some of their SF values (yet some are still positive); this means that Carrot will more likely converse using NP-based values

(since they score higher and Carrot is greedy) yet Carrot might accept some SF values with positive reactions. As a result of this, the baseline for scores is positive instead of zero.

Here, we highlight the player's scores with Carrot using multiple topics. As mentioned earlier, we used multiple topics because of the generative nature of the program. We believe that having more player data matters since we are looking at the underlying values and not the content of the sentences. Having said that, all of our sentences are constructed using the same underlying moral metaphors; the topic itself should not interfere with the data since they are structured the same way.

Player number	1		2		3		4		5		6		7		8		9		10	
	S	dv	S	dv	S	dv	S	dv	S	dv	S	dv	S	dv	S	dv	S	dv	S	dv
	23		10		28		10		21		18		31		9		8		22	
	34		3		21		12		23		31		32		18		19		3	
	23		6		41		8				-3		32		5		32		12	
	2		13		21		34				19		-1		30		26		28	
	9		6		26		17				31		32		12		12		23	
	7		34								33		31		30		15		22	
	33		21										26		27		19		28	
			38										13		26		30		12	
			4										-1		2		4		3	
			18										27		13		10		23	
													20		16		26		36	
																	15		23	
																	30		22	
																			15	
																			24	
																			15	
																			28	
																			3	

Table 8: The table presents the player's scores with their conversational interactions with Carrot. The colors are a reference to the deeper model used by the players, where red is an SFM value and blue is an NPM value. Green colors are values that are closely related.

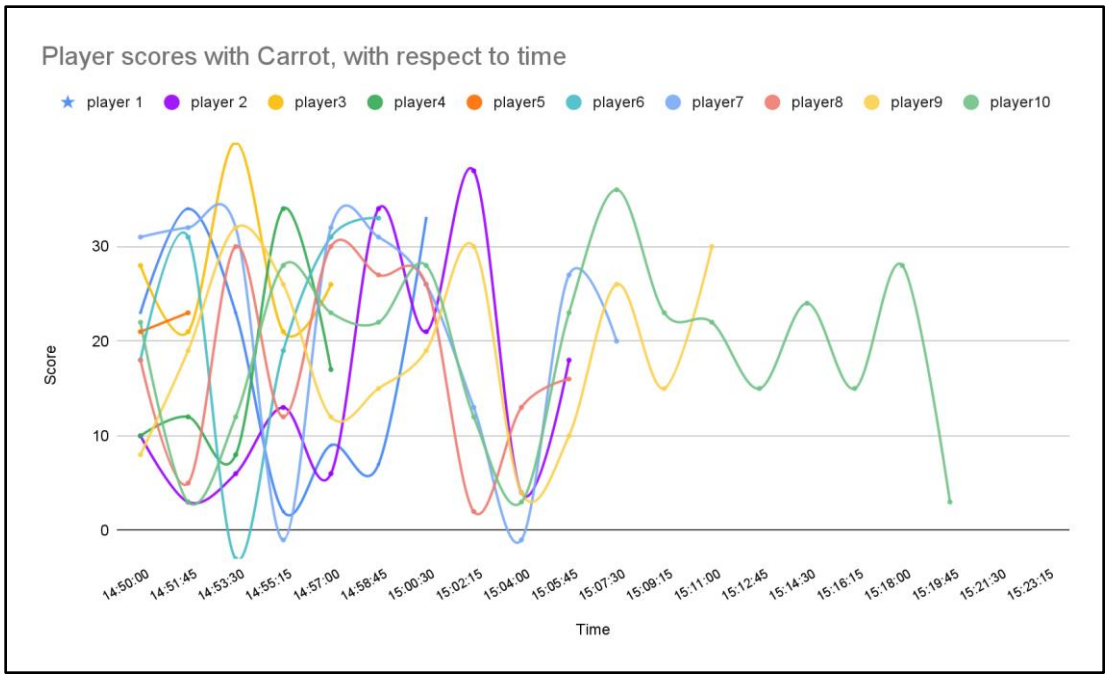

Figure 28 depicts the player's scores with respect to time while conversing with Carrot.

Unlike with Amrock, we noticed that the players utilized the correct moral model in most cases while conversing with Carrot. As we can see from Table 8 and Figure 28, the player's frequency of using NP-based values is much higher than SF values. We also noticed a shorter response time within Carrot's conversational rounds. Please note that figures 27 and 28 have been formatted to use a normalized time for all player inputs. Each timestamp is relevant to the conversation taking place.

Now that we have covered our log data, let us examine our methodology with our interview data. As mentioned earlier, each interview lasted approximately fifteen minutes, during which time we inquired about the player's experience within our system and asked players their thoughts on the CNPCs they encountered along the way.

After conducting our interviews, we transcribed our audio sessions (combined with our notes where applicable) into PDF documents detailing all player inputs per study session. We then extracted all of our data per question in different Google Sheets tabs. Once we had all our data organized, we examined the text and coded all player inputs into our initial coding, looking at the data line by line. We also subcategorized our data (where applicable) into subcategories depending on the CNPC mentioned during the interview.

We then examined our data thoroughly, noting the different themes that emerged. Generally, the player's data focused on two broad categories: those involving the characters themselves and those involving the gameplay elements presented by the system. Please note that this section details our general findings; in later sections, we will discuss these findings in greater detail. When examining all data relevant to the CNPCs, we noticed that the player's data contain elements relating to the CNPC's descriptive information, personality assumptions, tactics involving the CNPC (thoughts on how the CNPC operates), the player's real-world assumptions (and their association to the CNPC), and constructing stories about the CNPCs they met. The data also included other elements about the CNPC and the questions asked during the interview. The list of CNPC data includes the player's assumed CNPC values, what CNPCs were memorable, the persuadability of the characters involved, and whether the characters were consistent. Here, we list the data in greater detail:

- Eight of the 12 players mentioned values similar or related to Lakoff's areas.

        - For instance, some players mentioned political information, such as seeing the CNPC as holding traditional or progressive values, while others named core values noted by Lakoff's metaphors, such as empathy, growth, and strength.
- Three of our players mentioned areas relating to values but expressed differently. For instance, some players thought that our CNPCs had values such as freedom of association and freedom of self-expression; we note that these values are similar to the topics involved. Some players inferred these values based on the topics discussed (e.g., dating other species).
- Ten of our players made personality assumptions based on the presented values.
    - e.g., Amrock is a supremacist.
- Seven players constructed stories and background information about the characters they met.
- Six players related the game's data to real-world events and artifacts.
    - For instance, one player saw the in-game item blood pills as a metaphor for hormone replacement drugs in real life.
- Eight players mentioned memorable characters based on a mixture of conversations, stances, and character features.

- The majority of our player base felt that Amrock, the SF character was hard to persuade, others noted an easier time with Leo and Carrot, the NP-based characters (with some SF values).
- Seven players felt they were able to persuade the NPC.
- Ten players felt the characters were consistent.

The gameplay themes that emerging from the system included the player's role-playing experience, the challenging nature of the system, tactical gameplay information, world-building through additional UI features, and gameplay issues or confusing elements that arose during play. Players could generally figure out the gameplay through trial and error, building an NPC mental model, and playing the game as a puzzle. Two of our players felt uncomfortable while playing the role of a character whose beliefs did not line up with their personal beliefs.

Some of our players mentioned difficulty within the game's character supplementary features (mini computer interaction involving the log, character search feature and character information), while others liked it for world-building purposes.

This section discussed our general methodology and results involving gameplay logs, player utterances, and interview accounts. In the next section, we will discuss these results in greater detail, adding in interpretation.

3. Discussion

In this section, we argue that our system and data point to early evidence of character believability; we argue that players are able to perceive notions of

believability, including a character's personality, motivation, and sociability. We note that while this study is exploratory in nature, finding early notions of believability is promising. Our next study focuses on believability elements in greater detail.

We also argue that players can perceive values and beliefs in many ways, including political ideologies (which our system models via NP and SF modeling), deep values, and the player's relatedness to our NPCs. We believe this study establishes the foundation needed for our subsequent study, which further explores the implications of values-based design and its role in character morality and believability.

Before we begin our discussion, please note that player quotes in this section have been corrected for purposes of grammar and context.

3.1 Character and Personality

Unsurprisingly, many players made personality assumptions based on the character's values. While our game did include personality modeling, it did not highlight it in the CNPC's utterances (text responses) but was mainly used in generating the patterns and beliefs a character would hold (**Chapter Six**).

As a reminder, our game is generative; every run may place the characters that the players meet in a pro or a con state; however, as a result of our configuration file, we can solidify the character’s main model, influencing their deeper values. Having said that, Amrock, our SF character, according to some players, seems to be progressive when taking a positive stance on a political issue, while other characters, when taking a con stance, seem to be prejudiced despite their model. One player, for instance, felt

that Amrock was the most progressive and friendly of the lot, while Carrot seemed to be a bit bigoted.

Some players felt Amrock the Wolf was a stubborn character. We believe this is because of the player's inability to use SF values to convince Amrock instead of using NP values. Players made some personality assumptions based on the character's perceived level of stubbornness; for instance, one player said: "At some point, I felt like I was talking with grandma" when speaking to Amrock, while others mentioned personality factors about the difficulty they had with the character. One player, for instance, mentioned, "*The wolf was really stubborn and did not share what my personal view would be in this situation. The rabbit had this roundness and softness. I remember the wolf had sort of a spikiness to them. They felt sharp."* While others called out the character as they saw it, for instance, one player said, "*Amrock is a supremacis*t." While others felt that the character's personality lies in their values, one player said, "*I think they make strong arguments in the defense of their love."*

In general, we noticed a link between the character's values and the assumed personality attributed to the character by the player. We think having a positive or a negative stance on one's value shifts the character's personality from the player's perspective. We can see this clearly when viewing characters like Amrock, evidenced by some players viewing Amrock as friendly while others perceived him as stubborn and mean. We note the character has the same model in each run, with a variation on the depths of each value (i.e., Amrock is still an SF character with different variations of the SF values for each run).

3.2 Characters and Constructed Stories

One of the more interesting outcomes of this study is how players constructed stories based on the values and the characteristics of the characters involved. While the game does little regarding world-building and character definitions (aside from players optionally searching for character patterns), players still construct stories from the character's values. For instance, one player commented, “I feel like Carrot believes what is on their Facebook feed, which seems to be fear-mongering.” In contrast, another player justifies the CNPC's beliefs on personality factors, commenting, "The bunny brings the past and has more fear and anxiety and maybe pulls a few fallacies to justify their beliefs."

Another player imagined jealousy as a motivation for one of our characters, saying, “*What he (Leo) said about relationships… Maybe Leo feels like they wanted a soulmate, and these people are with their solemates.*”Another player tied the CNPC's relationship as a motivation for defending mixed species integrations as a topic by saying, “*I think Amrock was in love with a rabbit who lived in a mixed species community. Maybe Amrock cares about defending those choices and making the case that mixed species communities were healthy*.”

By reading character patterns (traits), one player pictured the CNPC as self-loathing and going against their agenda, saying: “*The sense of loathing and projection found in Leo... from my reading, traits, and beliefs (topics), I think highlight what is going on her*e."

Other players tied the game to real-world events in both characters and artifacts. Some players, for instance, felt that characters emulated life-like qualities of political society right now. One player said "*I found them parroting general arguments in the US right now; maybe that's the thing I would come away from,*" while another said, "*It reminds me of Zootopia in the movie. I guess it kinda reflects some little aspects of our own society right now*."

In general, we found players constructing stories and assumptions based on the character's values, patterns, and relationship statuses. We believe that a link exists between the values seen and the players' assumptions about the characters. These assumptions arise from a combination of in-game elements (assumed character personality, history, and motivation) and real-life elements (personal experience or personal perspective on world events).

3.3 Characters, Values and Beliefs

Our study shows that most of our players attached some value structure to the CNPCs. Please note that we did not compare the actual depth per CNPC in this study, as we aimed to learn what the players perceived instead. However, we keep in mind the general areas of beliefs (SFM or NPM). We categorize player constructed beliefs as follows:

- **Beliefs that emulate political values and views.** Some players attributed character beliefs to real-world political views, highlighting traditional/conservative, and progressive values in their statements. For

example, one player mentioned, "*Carrot has a more progressive tone,*" and another stated, "*Carrot likes more and more progressive values.*" In contrast, another player stated," I think Amrock probably holds very (true) to the rules, he talks about traditionalism" and another stated, "*Amrock cares about tradition*" and lastly, "*if I remember correctly, Amrock likes strength, courage, and responsibility, which maps to my understanding of real life. Makes me think of right-wing politics*."

- **Beliefs that highlight deep values**. Some of our player base highlighted beliefs similarly worded or mirroring the ones mentioned in Lakoff's metaphors. For example, one player mentioned *"health and safety, community and personal freedom. Strength, around larger animals and I guess dedication to the ones they love,*" where the former's health, community, and strength are similar to Lakoff's Moral Health, Moral Wholeness, and Moral Strength (please see **Chapter Six** for more definitions on specific deep values). Another example includes a player mentioning "*(on Leo) He likes growth and society but not cohesive society,*" and another stated, "*I think the lion cares about personal empathy, their personal lives and how happy they are*" and "*I think Leo cares about pride and maybe supporting the community. When I chose things like growth or diversity and community, he seemed more appreciative… Leo is more about the overall community, like supporting each other,*" where the last three examples include elements similar to Lakoff's Moral Growth, *Empathy*,

Nurturance of social ties, and *Happiness*, although Lakoff's metaphors highlight happiness and well-being for others as well as oneself.

- **Beliefs that are constructed from stories.** Some players constructed beliefs based on their constructed stories. Examples of this include the earlier mentioned Facebook story where one player stated, "*I feel like Carrot believes what is on their Facebook feed, which seems to be fear-mongering,*" or a character believing in freedom, stating, "*I think Amrock was dating the rabbit. He would be a proponent of freedom of association*" and another stated, "*Leo's values are strongly (made of) freedom of expression."*
- **Beliefs that are inferred.** Some players inferred beliefs based on the discussed topics. As a reminder, the discussed topics for this demo included interspecies dating, mixed communities, and purchasing forbidden goods (for more details on these topics, please refer to Chapter X). For instance, one player inferred change as a value by stating, "*The bunny believed in having change happening (being positive) is a good thing."*Another stated, "*Both Carrot and Leo were talking about how predators are dangerous. They value safety in their own way.*"
- **Beliefs that belong to the player character.** Some players attributed a belief belonging to themselves or the player character. One player, for instance, stated that toward the end, they felt like Amrock hated them. Others felt their personal beliefs affected how they interacted with characters, making the game challenging, while others believed their beliefs affected how they talked with a

character. Some players thought the CNPCs wanted to trick or convince the player, stating, for instance, that, "*the lion wanted to make you (the player) believe that carnivores are not dangerous."*

Another interesting aspect that arose from our belief inquiry lies in the assumed strength of a belief; one player, for instance, believed that the NP-based characters felt more soft in their beliefs, while the SF character felt strong (held their beliefs tightly). Another echoed that statement, stating that the SF character felt more stubborn. Another stated, "*I felt like Amrock the Wolf was so strange; they are so set in their ways, and they are so positive! They are hard to convinc*e."

3.3.1 The differences between SF and NP characters.

On that note, let us discuss the differences between the SF and NP-based character(s). Unsurprisingly and based on the data above (charts in section X), it showed that players generally had a harder time conversing with Amrock, the SF character. According to our player data, this is mainly attributed to a perceived personality trait (e.g., Amrock seen as stubborn) and the character's misalignment with the player's personal beliefs. Players also noted that the SF character was hard to persuade. These elements are highlighted in charts 25 and 26, where the former (Amrock's case) is more chaotic, suggesting trial and error. At the same time, the latter is clustered and shortened, suggesting an easier understanding of these characters. The minority of our players (three) felt that Amrock was progressive or friendly, while others (the majority) felt Amrock was more hateful, stubborn, or traditional, quoting traditional values like bravery and courage. Having said that, we believe some players

may have experienced Amrock taking a positive stance on a topic but expressing it under the SF model.

In contrast to Amrock, some players felt that Carrot and Leo were more persuadable, troubled, and easier to convince. One participant felt annoyed that Carrot and Leo were softer, while Amrock felt stern in their beliefs.

When we asked the players about any memorable moments with the characters, six of our players remembered Amrock through their conversations, difficult gameplay, personality traits, descriptions, and values. On the other hand, Carrot was dismissed, and one of our players noted that they did not care much about Carrot but mentioned recalling them dating, while another mentioned a feeling of softness attributed to the rabbit, and lastly, one player recalled Carrot talking about progressive values. Leo (the lion), another NP character, on the other hand, was remembered because of conflicting scenarios from the player's assumption and the character's actual portrayal in the game. For instance, one player expected to like Leo but disliked him as a result of his portrayed values/topics. Another player recalled that Leo has a sense of "self-loathing," projecting a mismatch between their patterns (descriptive information such as the lion having scary features) and their textual statements (e.g., believing in herbivore rights). Two others recalled him because of his perceptibility and aligned values.

4. Gameplay

4.1 Gameplay Issues

While most of our players enjoyed playing the game, we noticed that two players expressed discomfort in situations where the player character took an opposing stand against the player's own beliefs (since the pro/con stance to a topic is determined at run time and the player always opposes the NPC).

One player expressed confusion about the game's design and flow. That player seemed to correlate their gameplay experience with the quality of arguments; to them, the NPCs seemed to lack logical responses (expressed by the player as lacking organizational logic).

Two players felt that the UI elements (that depicted NPC patterns) conflicted with or contradicted other UI elements (other patterns on NPCs). We believe these situations could transpire from a logical error or may have been intentionally designed and authored with these patterns in mind (while the simulation has rules in place for strictly contradicting patterns, some exceptions, edge cases, and authored-in patterns via JSON are possible).

These situations are highlighted when an NPC has patterns that seemingly do not fit the topic or the character. For instance, a carnivore character can hate those with scary features while having scary features themselves (going against their agenda) or a character that hates interspecies relationships yet works with them.

We consider these contradicting patterns, which if taken as a design decision, can become a double-edged sword. On the one hand, players may find it confusing, but on the other hand, players may interpret these patterns as part of the character's story and world-building (examples can be seen in section 3.2). We believe that interpretation and player perception can have a wondrous effect on the outcome of a game. After all, perception is fundamental to evaluating an NPC's believability!

Other than the above, some participants commented on the location of UI elements, UI designs, and repetitive text (in some cases). While we adjusted the former in our next iteration of the study, the latter flows from the generative nature of the game, pattern availability, and lack of extensive authoring scenarios; some repetition toward the end is likely to happen after a few conversation rounds (days/characters), especially if the characters hold the same models/values or if characters remain unconvinced. We note the authored scenarios for this demo included three topics, each with one to three subtopics (argumentative patterns) with up to 18 deep values (string argument responses) at both pro and con stances.

Lastly, we noticed that some players were confused by similarly structured deep values, especially when they were part of the two contradicting models (NPM and SFM). For instance, Nurturance of Social Ties, an NP value, and Moral Wholeness, an SF value, generally care about social elements. The latter was narratively framed as valuing the community and the former as maintaining social relationships. Players were often confused when faced with these deep values, as they did not meet their

expectations (when one value would give a positive response and the other a negative one, which is anticipated, given the model's structures).

4.2 Other Experiences and Feedback

When players were prompted about their most memorable interactions or experiences within Argument Box, they responded with themes relating to the characters, NPC values, their strategy, UI features, and the argument format.

Some players liked having information about the characters, such as their relationships, likes, and dislikes; they also liked figuring out the NPC's responses and values. Another liked the character design, expressing some humorous moments in a character's return to the Box. Others liked the game's structure (i.e., debate format or expressing opinions that contradict the player's feelings on a topic). While expressing the above components, one player felt the conversation was abrupt but memorable. Another suggested UI change.

5. Conclusion

In this chapter, we argued that players can understand and perceive character values as presented in our system. This study serves as a basis for our next study, which asks specific questions about value-based design and its role in character morality and believability.

Chapter Eight: Values, Believability, and Morality

Chapter Seven demonstrated that players were able to perceive moral values and beliefs within our system. Knowing that a preconception of values exists, we can now ask specific questions about value-based design and the role it plays in character believability and morality. Can values and beliefs create interesting characters?

Through this chapter, we will argue that not only can players discern an NPC's beliefs but also establish connections between values and notions of character believability. Specifically, we believe that values can influence the player's perception of an NPC's personality, motivation, morality, and overall illusion of life.

We believe that our system offers promising results and could serve as a basis for morality and value-based design. In the following sections, we will discuss system iterations that resulted from **Chapter Seven**'s study. We will review our demo, study design, and findings as we learn more about the relationship between value-based design, morality, and character believability.

1. AB Changes

As in Chapter Seven, Argument Box's demo is configured to include up to three characters visiting the player in three in-game days. While the general gameplay and (most) configuration details are the same as in the first study (please see **Chapters Six and Seven** for more details), we modified the system to account for balance and character moral model variety. We also consolidated specific deep values that caused

confusion in our first study. This section aims to clarify the modifications made to AB's configuration and moral model scripts.

1.1 Character Configuration

Similarly to our last study, the player plays the game with the goal of dissuading NPCs of their opinions. As a reminder, in AB, the player typically converses with two to three characters per in-game day, where each conversation lasts for either five rounds or until the player successfully dissuades the NPC. The characters in this demo include Amrock the Wolf, Leo the Lion, and Carrot the Rabbit. Let us first examine the revised character configurations used for this study.

Amrock the wolf, as in our last study, is a strict father character; this character is influenced by the game's Strict Father model to like values associated with the strict father (e.g, Moral Health and moral Wholeness) and to dislike values associated with the Nurturant Parent model (e..g., Empathy and Moral Nurturance). In addition to the game's base simulation, we configured Amrock to deeply care about these SF core values: Moral Order, Moral Boundaries, Moral Strength, and Moral Health. This configuration ensures that the character at least likes or loves arguments based on the mentioned deep values, creating a heavily strict father character in the process.

Unlike the last study, we adjusted Carrot the Rabbit as a pure nurturant parent character. Like Amrock, Carrot's baseline values are determined by the simulation; however, because of her NPM model, Carrot will only approve NP-based arguments. Lastly, Leo the Lion is modeled as an NP-based character with some SF tendencies.

We configured Leo's custom deep values to include Moral Health as an SF-based value. We note that Leo also cares about NP values such as empathy. We believe configuring characters this way creates a good range of characters with diverse moral values. Now that we have covered the characters' main configurations, let us discuss our system adjustments.

1.2 Model Adjustments

We adjusted the game's deep models (SFM and NPM) and character scripts for in-game balance reasons, where we adjusted the range of deep-value point allocation (Loathed to loved) to account for a better point variation between the six ranges (**Chapter Six** includes more details on the implementation side). We believe that adjusting these deep values creates a more balanced gameplay experience, whereby players are more likely to receive upper and lower ends of the model scales (loathed and loved), as opposed to earlier iterations where *appreciated this/disliked this* was more likely.

We also revised the total number of deep values belonging to each model. As mentioned in **Chapter Seven**, some similarly structured deep values confused our players as a result of the likeness of the argument structure and its use in opposing models. With that in mind, we removed all instances of Moral Wholeness (i.e., the importance of community) and left instances of Nurturance of Social Ties (i.e., the importance of maintaining social relationships) in the game. We also removed some deep values that contradict their moral model if taken at face value; this includes Moral

Nurturance from the Nurturance group of the SFM model and MoralStrength from the NPM model, where the former highlights nurturance from an SF perspective, and the latter highlights strength from the NP perspective[33].

1.3 A New Note-keeping System

We created an optional note-keeping system since this study relies on participants remembering many values and beliefs. The first note-taking feature incorporates a drag-and-drop interface where participants can drag value icons and assign them positions on a hate/like scale according to their interpretation; this GUI is available for all three characters. We also added a blank text field where participants can enter any notes that come to mind during gameplay sessions. Figure 29 depicts a sample note-keeping system.

After completing the game, players are taken to an end-game screen where they can reference their notes and the conversation log between them and the characters.

[33] We note that these values highlight elements that oppose their moral models; this is a lower priority in each model. The model(s) justify their usage only when certain conditions have been met (e.g., nurturing others in SF scenarios only happens if the person proves themselves worthy of that help among fulfilling other SF criteria). In a perfect scenario, users could link the talked-about character's compiled traits and relate their usage due to their past actions or deeds; we note, however, that doing so would be challenging in our game due to the minor world-building component.

We note that note-keeping in-game is entirely optional. Players were free to use other methods of note-keeping, such as notebooks or scraps of paper if needed when filling out our survey.

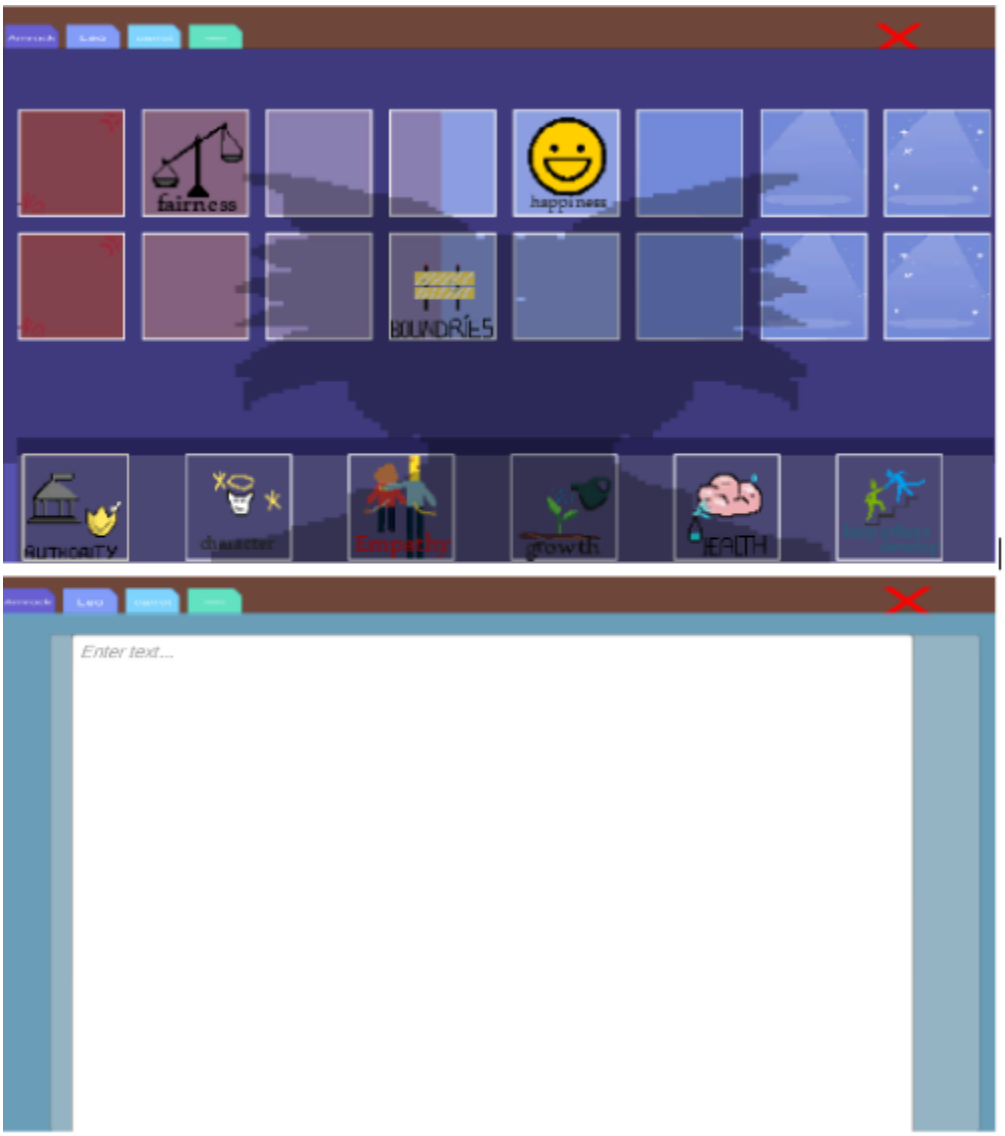

Figure 29: From top to bottom, a note-keeping system in the game depicts the drag-and-drop interface available to each character via tabs and an empty scrollable text field where players can type notes during gameplay.

2 Study

2.1 Recruitment and Participation

We advertised our second study in the summer of 2024, during the months of June, July, and August. The study was advertised within the computational media department and internationally in our home country of Saudi Arabia. In the USs, the study was offered as an optional part of a summer class, where interested students could sign up to participate for class credit. We advertised the study via word of mouth, email lists, and social media channels. All participant data are to be anonymized. We note that participants cannot participate if they played an earlier iteration of the game or if they took part in the previous study. By the end of August, we had a total of 18 participants that both played the game and completed the survey. Participants that only played the game and did not participate in the survey were removed from our analysis.

2.2 Overview

The study consisted of two parts: playing an online game and completing a post-gameplay survey. The game included an in-game code linking the player's results with their gameplay data. Before we dive into the gameplay and survey portions of our study design, we wanted to remind our readers about our study goals and broader research agenda.

In our last study (**Chapter Seven**), we asked if players could perceive values. In short, the answer to that was yes. Since identifying a value structure was

accomplished in our previous study, we can now dive deeper into values and their relationship to character design and morality. We approached this study to learn more about values and their relationship to characters. What can we learn about values? Can players correctly identify a character's beliefs? Do players sense any moral standings within our characters? Does a relationship exist between values, morality, and believability? Considering all these questions, our survey focused on values, believability, and morality. Before we dive into our findings and results, let us look at how we structured our study and gameplay.

2.2.1 Gameplay overview.

The gameplay design is similar to that of **Chapter Seven**, with minor system adjustments and the addition of the note-keeping system discussed earlier (section 1.3). In summary, players were asked to converse and dissuade NPCs in three in-game days. Each game day consisted of three conversations and up to three character interactions. We note that each conversation lasts five rounds or until the player successfully persuades the NPC on a given topic. After finishing three days, players are redirected to an in-game screen that displays the player's conversation logs and their note-keeping tabs.

Similarly to our previous study, gameplay data are saved as a text log. The log conveys the NPCs present during the player's run (including their deeply held beliefs and model affiliations), conversation data, topics, argumentative patterns, and the conversation iterations between the player and the NPC. Please refer to **Chapter Seven** for more information.

Players played the game for about 20–45 minutes, after which they completed a post-gameplay survey.

2.3 Study Design

Since we would like to learn as much as possible about values and beliefs and their role in character believability and morality, we structured a survey that focused on open-ended questions. In addition, the survey also contained a myriad of other question types (e.g., linear scale, short answers, long answers, checkboxes, multiple-choice) related to character believability, morality, and beliefs.

We will next explore these questions and the motivations behind them. By this chapter of our dissertation, readers know we value character believability as a research area. Naturally, a lot of our questions will focus on characters. As a reminder, research (**Chapter Two**) has established frameworks and methodologies used in evaluating believability notions. These methods asked participants to rate characters based on their believability criteria (e.g., personality, change, and social relationships).

Thus, our questions asked players what they thought about various believability criteria through linear scales and multiple-choice questions. We note that these questions were followed by an open-ended response box through which participants can elaborate on the specific criteria in question. We believe this is important, as believability is gained by player perception; gaining insight into why a criterion failed or succeeded is important in establishing links between our characters and their believability metric. We also argue through these methods; we can assess a character's

values as a link to their perceived believability because of the system's architecture that emphasizes values and beliefs as a considerable part of our character designs. We noted that our questions were accompanied by character images in case our participants forgot character names.

Without further ado, our believability-based questions focused on the following:

- **The NPC's awareness level** is used to gauge their awareness of the world around them.
- **The NPC's understandability level** inquired as to whether the player could build a mental model of the NPCs, framed as the player's capability in understanding the NPC's thought processes.
- **The NPC's personality** asked players whether they felt an NPC had a personality and, if so, to describe it.
- **The NPC's predictability** was assessed by asking whether players found the NPC's behavior predictable.
- **The NPC's consistency** asked whether players thought the NPCs were consistent.
- **The NPC's sociability** asked whether players can picture social links between characters.
- **The NPC's interestingness** level asked whether players found the NPC to be interesting or engaging.

We note that these questions were presented to the players (with minor terminology explanations) as multiple-choice or linear-scale questions between two extremes. Each question asked players to expand upon their responses in their relevant text fields.

To understand these value-based characters further, we wanted to see how people perceived our NPCs. We wanted to capture their thoughts before asking them specific value-based or morality-based questions. Therefore, our next set of questions asked the players their general thoughts about the characters, if they perceived any morality within them, and if they found the characters to be lovable, hateful, or memorable. Each question mentioned earlier contained an extra text field where players elaborate on their thoughts. We believe that asking open-ended and somewhat vague questions can add depth to the player's responses and aid us in discovering other elements without influencing the player's responses directly.

After players had a chance to express their opinions unprompted, the next set of questions dived deeper into specific morality and value-based questions as used by our system. Since this study builds on top of our last study (**Chapter Seven**), we assume players understand that NPCs in our game have values and beliefs. We can bring the Lakoff values to the forefront and ask players questions about those values.

Accordingly, the next set of questions focused on value identification and alignment. We presented each of our NPCs with a private table in which each row contained a Lakoff value used in our system, and each column contained the depths at which the NPC holds that value. Each Lakoff value from both NP and SF models has

been translated in layman's terms, making them more understandable. For instance, the value of *Moral Strength* was translated to "*strength and willpower are important.*" We then included a range of beliefs where players can select whatever range they think is an appropriate response for each value on the column side. The column headers included: *is repulsed by this value, hates this value, is indifferent about this value, holds this value*, and *deeply holds this value*. We also included edge cases such as *I think this value did not show up* and *NA, the character did not show up* as possible options. The players were then instructed to fill out the table with all applicable responses about the character in question and rank the depths of their values. Figure 30 illustrates how this question was presented to players in Google Forms. We note that all possible Lakoff values are expressed in the table, with Figure 30 highlighting a subsection because of formatting limitations.

What values do you think the character **Amrock** cares about? *

	is repulsed by this value	hates this value	is indifferent about this value	holds to this value	deeply holds this value	I think this value did not show up	NA (this character did not show up)
preserving the natural order	☐	☐	☑	☐	☐	☐	☐
being normal is important	☐	☑	☐	☐	☐	☐	☐
scared of spreading influence and ideologies	☐	☑	☐	☐	☐	☐	☐
a sense of character and constitution is important	☐	☐	☐	☑	☐	☐	☐
strength and will power is important	☐	☐	☐	☐	☑	☐	☐
abides and believes in authority	☐	☐	☐	☐	☐	☐	☑

Figure 30: A subsection of our value identification and alignment table as it pertains to Amrock, the Wolf.

We then included additional questions to capture the player's thoughts on the game's morality structure at a deeper level. One way is by asking players to compare what they experienced with other games that highlight moral systems in their designs. We included some examples of games and descriptions as a reference. We also explored the relationships between morality (through values) and believability, by asking players whether their experience made the game more or less engaging or believable.

We ended our survey by including a couple of open-ended questions meant to gain further insight into our characters and the overall player experience. We asked our players whether any values resonated with them, who their most and least favorite characters are (and why), whether anything struck them as odd or surprising, and lastly whether they had any other feedback or comments. This section provides an overview of our survey question; please refer to Appendix A for more details.

2.4 Study Results and Discussion

From our earlier chapters, it is evident that our system emphasizes morality through character values. We believe that a systemic value-based design can enhance the overall character, making characters more *believable*. However, we understand that limitations, such as an extended game world and insufficient authoring content and context, are hard to accomplish in a small demo. Nonetheless, we believe we can gain valuable insight regarding believability, character values, and, in turn, morality.

This section will first divide and examine the results through three primary lenses: believability, value design, and morality. Please note that while the following sections present quantitative data, our focus will highlight qualitative data. We examine qualitative data more closely for three reasons.

- 1) Qualitative data help us understand our players and gain insight into their thought processes.
- 2) Qualitative data have been established as a valid standard for understanding certain topics, such as character believability (Chapter Two).
- 3) Some questions may be misunderstood by our players, resulting in somewhat skewed quantitative data. In addition, the nature of a survey is such that there is not an opportunity to clarify or otherwise answer questions for the participants. Participants may interpret the questions or terminologies according to their understanding. For instance, some of our players interpreted morality as righteousness. Please note that the following sections highlight these interpretations and situations where appropriate.

2.4.1 On believability

As mentioned in earlier chapters, *character believability* is an often ambiguous and misunderstood topic. For a character to be believable, they must demonstrate a combination of believability criteria that enhance the overall character. Chapters One and Two discuss these believability criteria in greater detail. Before we begin, we

would like to review NPC believability. NPC believability pertains to the NPC's ability to mimic life, convey emotion, express personality, have social relationships, and act autonomously with intention; NPCs should also portray a role and act convincingly, among other criteria. The list of criteria is indeed extensive. It is difficult for an NPC to hold these criteria, let alone maintain them through extended gameplay. So how can we build believable characters, let alone test them? Earlier chapters presented a deeper understanding of character believability; here, we summarize and argue the following:

- **Our evaluation has to inquire about the believability criteria themselves.** In order to determine whether our NPCs are believable, we need to find out what about our NPCs is believable. Examining elements from the list of believability criteria is a good way of measuring the overall believability of a character and has been established in the existing literature [115].
- **A combination of criteria is sufficient to express character believability.** While adhering to more criteria creates a better overall character, a combination of some criteria may be sufficient depending on the game and artifact in mind. We used characters like the Valentine butcher from RDR2 in previous chapters to argue this point. **Chapter Two** provides a deeper analysis and case studies showcasing how some believability characteristics appeared in games.
- **Believability is evaluated based on the player's perception.** Player perception plays a crucial role in identifying and maintaining believability criteria. We saw examples in Lim where a literal block (NPC) appears to have a rough personality and a sense of autonomy (**Chapter Two**). We also saw

instances of perception in **Chapter Seven's** study where players created character stories from in-game artifacts.

- **Qualitative and exploratory measures are needed to understand believability.** As expressed in Chapters One and Two, character believability is often evaluated based on qualitative data. While some quantitative data can be helpful in identifying patterns and giving an overview glance of results, we must follow up with qualitative-based questions to ensure a thorough understanding; after all, believability is based on perception.

Unfortunately, we understand that maintaining character believability will eventually fail because of limitations and authoring constraints[34]. In many games, this comes in the form of repeated interactions or players gaming the system or gaining extensive knowledge of a character or artifact to break the illusion. Here, we expect repeated dialogue in exhaustive cases (in our situation, this depends on the simulation, NPC stance, and available patterns) to eventually affect our characters' believability.

Before we begin, we would like to remind our readers that the demo used in this study includes three characters. A pure SF character (Amrock), a pure NP character

[34] Additionally, according to Wardrip-Fruin and Murray [149,123], the system will eventually lose believability as a result of the Eliza Effect because of the loss of collaboration between the player and the system; if either party stops collaborating, believability will eventually fail. (Examples include the player pushing past system limits or the system running out of proper responses.)

(Carrot), and an NP-based character with some SF tendencies (Leo). As a reminder, each character's arguments and value depth are determined at run time (influenced by our demo configuration; see section one). We believe that a strong relationship exists between value-based design and most of the believability criteria (as a result of system designs and limitations). As our system highlights value-based design at its core, we believe values can affect the play's perception of the characters involved, making them believable. Without further ado, let us first examine our results as they pertain to believable criteria.

2.4.2 On Personality

We start by reviewing our character's personality results. In a straight yes, no, or NA answer, most of our players believed that Amrock held a personality, while Leo and Carrot had mixed results. The three charts (Figure 31) and table 9 highlight the characters' personalities at an overview level.

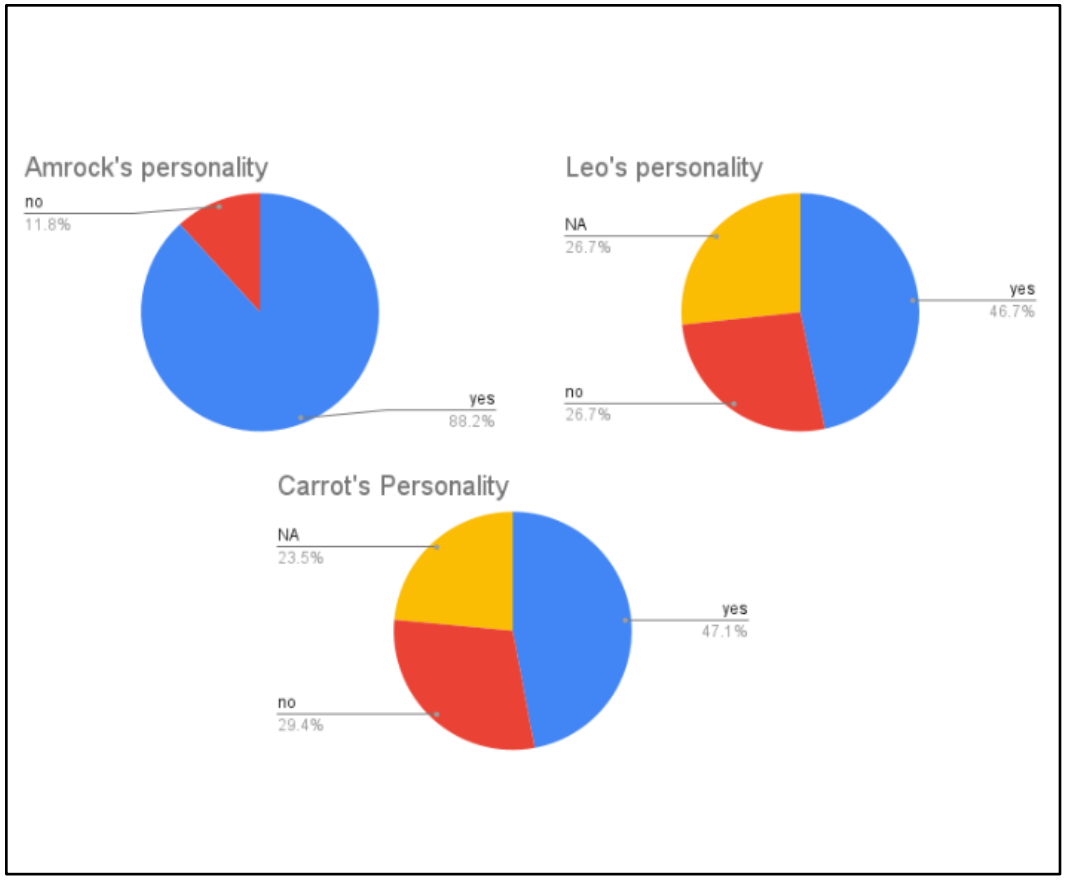

Figure 31: Figure from left to right, top to bottom: Amrock's, Leo's, and Carrot's personality charts where the blue color indicates a personality presence, yellow indicates an inapplicable status, and red represents no personality.

Character	Amrock	Leo	Carrot
Personality	15	7	7
No personality	2	4	4

Table 9: The player's personality responses for all three characters.

Upon examining the personality descriptions given by our players, we noticed three themes emerging from the characters.

- **Personality and emotion-based descriptions.** As the title suggests, some players described common personality traits canonically found in character descriptions. For instance, participants would describe characters as *hopeful, strange, angry, stubborn, happy, mean, shallow, and self-centered*, among other things.
- **Wide model descriptions.** Some players described characters based on their model orientation. As a reminder, Lakoff uses the SF and NP models as metaphors for conservatives and liberals. Some players used terms associated with either camp in describing the characters. For instance, phrases like “Conservative that resists change”c were used to describe Amrock. Other phrases hold similar definitions describing the models as a whole. Examples include the players describing characters as *oldhead, closed-minded, old-fashioned, republican, liberal*, and *progressive*.
- **Values as personality descriptions.** Other players brought up specific values associated with the characters as personality descriptions. For instance, one player described Amrock according to Amrock's beliefs: "*He (Amrock)* ***believes*** *that everyone should stay within their* ***own species,*** *and it goes against the* ***natural order*** *if they start to diverse.*" Other examples include "Amrock is a person who values being considered **normal** and **in the herd**. He believes in **character and authority.**" The same descriptions can be found in Leo and

Carrot. For instance, Leo was described by one player in these terms: "*Leo values **inclusiveness** and **helping others***," "*He (Carrot) was incredibly **open to mixed communities***" and "*Will accept others' words and slowly **change***."

Interestingly, some participants used comparisons with other NPCs to highlight a character's personality. Others used an NPC's social standing as a description. For instance, the phrase, " I think Leo was a **lot more** understanding" or describing Leo as "**more** tolerant of others" and "…childlike and focused on his dislike of **one individual.**"

According to some players, some characters appeared to lack personality because of iterative conversations, encounter rate, the conversation itself, and conforming to norms (phrased as "*doesn't have their own opinion*," or *"doesn't want to hurt others"*). We note that the "NO" responses came from describing the NP-based characters. While some results are expected because of the nature of the game and limitations (encounters, limited conversations), others seem to hint at a lower personality due to the difficulty level where "easier' characters" (in this case, the NP-based characters) have less personality than do harder ones. We note that SF characters are generally described as more difficult (**Chapter Seven** established this, and upcoming sections highlight this similarity, as we will soon discuss.)

We argue that some statements about the agreeable nature of NP characters are, in fact, personality descriptors. While some players mentioned "NO" in their responses, they elaborated by mentioning descriptions that are docile or agreeable in nature. For instance, one player mentioned that Carrot did not have an opinion due to the nature of

her model. The player stated, "*When you don't have an opinion and just don't want to hurt another person's feelings, you won't be honest about your feelings or beliefs and just give the opinion that would please the other.*" Upon synthesizing this argument, we believe these responses are meant to highlight the strength of a personality and not its nature. In the former example, the player literally described a softer personality, one that is overly docile and friendly.

The player's perception indeed plays a role in determining the overall personality. We noticed a key relationship between personality and value descriptions, where values impacted the player's perception of a character's personality. We also noticed that the phrasing used to describe the characters is somewhat unique, occurring differently in some of the responses; this is illustrated as a value that stuck with the player via interaction or a general descriptor such as “easily irritated.” At the same time, other descriptions appeared often, such as Amrock being described synonymously as stubborn or closed-minded.

2.4.3 On character motivation.

Motivation can come in many forms. In many media representations, characters are usually authored to have a character arc and storyline that drives them toward their goals and conclusions. We also saw some examples in **Chapters Two** and **Three**, where characters in games and social systems act upon their goals via interactions informed by their autonomous systems, authoring content (storyline-based games), volition (e.g., social systems), and planners (e.g., GOAP). Indeed, many systems present motivations in different ways. However, as stated earlier, motivation, like other

believability criteria, is based on our player's perceptions. So, how can we properly evaluate motivation?

One way to gauge character motivation is to review and analyze our players' responses. We need to understand whether players can understand our characters and their thought processes. Do they see the characters as predictable? Let's now examine our players' responses and analyze them in terms of understandability and predictability. Following that, we will look at the change characteristic by measuring the NPC's consistency.

Upon inquiring if players understood what our characters were thinking about, most responded that the characters were understandable. On average, our characters had a 3.2 score, where a score of one is seen as an extremely negative score (no understandability at all), and five is an extremely positive score (indicating high understandability). Table 10 and Figure 32 highlight our players' understandability scores. We note that we care more about the players' responses than the actual statistical score. Thus, in our analysis, we take anything above three as an explanation for an understandability score and anything below as an indication of a low understanding of our player's perspective.

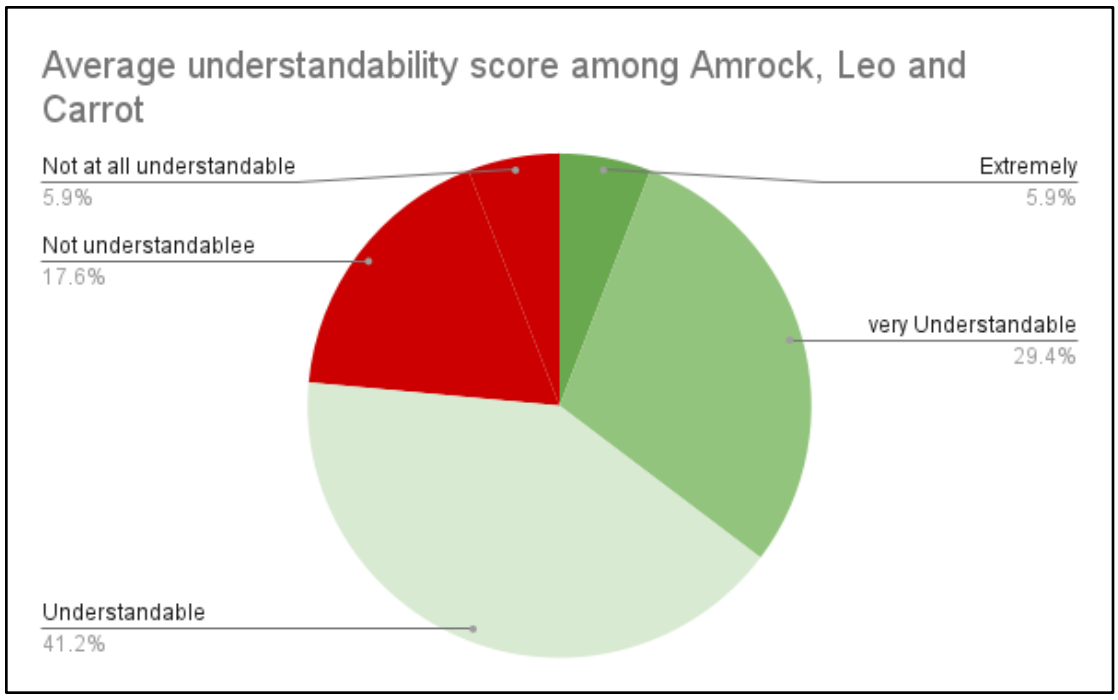

Figure 32: presents the average understandability score among all three characters.

NPC	Not at all understandable (1)	Score of 2	Score of 3	Score of 4	Extremely understandable (5)	NA | no result indicated
Amrock	1	3	7	5	1	0
Leo	1	2	4	5	3	2
Carrot	0	3	3	4	2	5

Table 10 indicates the scores submitted by our players in response to the question, Is it easy to understand what the character [X] is thinking about (where X is the name of the NPC)?

Before presenting our results, we would like to note that Some NA/No results indicated by players in our survey have expanded text responses. The responses we analyzed here contain data explaining the reasoning behind the player's experience. We note that responses that highlighted the score without explanations were dismissed from our analysis.

Upon examining the players' responses, we noticed a trend whereby students equated understanding an NPC with comprehending the NPC's values and beliefs, the NPC's thought processes, the level of difficulty when conversing with an NPC, the player's perception of the NPC's personality, and finally the NPC's responses.

Examples include players stating that the NPC was "*reactive"* and "*opinionated."* Other examples hinted at the NPC's general beliefs. For example, when writing about Amrock, a player stated, "[He] is *set on old beliefs*." Another player reflected on both personality and belief traits, commenting, "This character seems very reflective of a typical racist, so it's easy to draw parallels between the two and understand Amrock better."

Other players expressed their understanding of some characters, like Leo, based on the ease of the conversation itself. Others utilized values and traits to express understandability. One player, for instance, said it was easy to understand Leo due to the character being more *sympathetic*. As noted earlier, some players linked understandability with understanding the NPC's general model. For instance, one player believed that the NPC's values aligned with their character (and expected responses), providing a high understandability score. Another mentioned the

character's positive feedback and attitude as an indication of understanding the mental model.

On the other hand, low understandability scores (indicated as three or below, depending on the player's feedback) were equated with the difficulty level of persuading characters, the character's stance on a topic (mostly seen as a negative stance), and an NPC making contradictory statements. Examples include statements such as: “It was difficult to determine the character's exact value set and therefore difficult to deduce which dialogue options would be persuasive,” ' “*Sometimes he agrees with you, and sometimes he argues*” and, “([It is] difficult to find effective dialogue options.”

Some players related the understandability of an NPC to their topic stance within a game (pro or con). For example, one player stated that they could not understand an NPC due to their negative stance, commenting, “*I don't think I can understand because he always wants to believe that it is better if species are separated, and with that thinking, society won't grow*.” Contrastingly, some players suggested that an NPC's positive stance presents a higher understandability score. For example, one player stated: “*Carrot tries to understand people, which is good*.” We note that these cases reflect the player's understanding as related to the NPC rather than the player's understanding of how the NPC thinks. While it's not reflective of understandability, it still speaks of relatedness.

In general, we found our NPCs somewhat understandable. The NPC's beliefs and stance affect the player's perception and mental model of the NPCs. We note that

the difficulty level or ease of the interaction warrants extensive research, but we believe a link exists between the difficulty of interaction and the player's understanding of an NPC. Lastly, we saw elements in which an NPC's values and model reflect the player's understanding of an NPC's thought processes. We note that values will be discussed in greater depth in a later section.

2.4.4 On predictability and consistency.

According to our data, our NPCs were somewhat predictable and consistent. We believe the predictability and consistency were circumstantially dependent on a combination of the NPC's model, interaction rounds, and arguments. Figure 33 depicts an overview of our predictability player data.

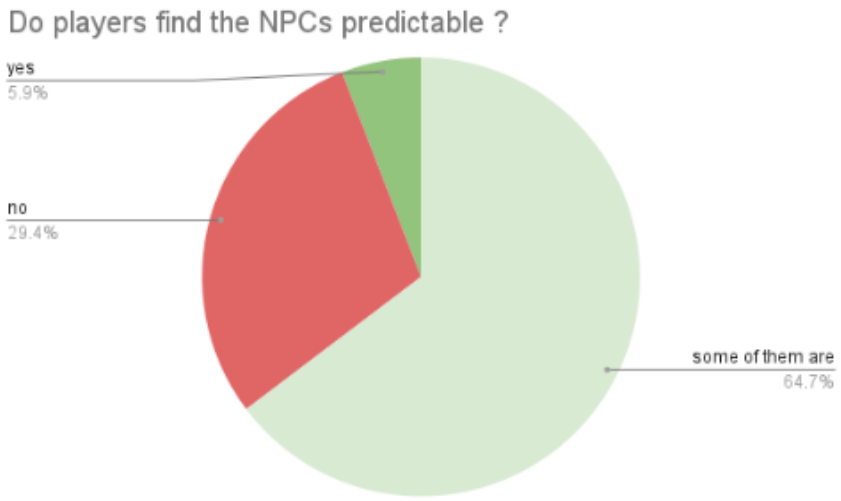

Figure 33 highlights general findings with regard to an NPC’s predictability.

Players often tied predictable behaviors to the character's repeated arguments and positive persuasibility results (NPC's persuaded reactions). It seems that repetition is the key factor in making the characters more predictable. We also note that the NP-based characters seemed more predictable than the SF-based characters. Example player statements highlighting predictability include: "*Once I understood how they think, I say the right thing to gain points,"* "*Carrot was more predictable than Amrock,"* and "*Leo is predictable because you can very easily tell what he thinks and what he responds well to*." Note: Leo and Carrot are NP-based characters.

As for unpredictability, players cited personality traits, having difficulty in persuading characters, and the NPC's counterarguments as reasons for unpredictability. Example statements include: "*I felt that when I was trying to persuade them, I would think that they would answer in a certain way, but they would act in a completely different way*;" *"It is difficult to predict how they will respond to dialogue options;*" "*I was at first taken back by Amrock's personality,"* and lastly, "*The best example is Amrock, although many good reasons were given since the first day, he accepted only a few good facts, but even when he was starting to accept the new order, he began to rethink the idea that this new order is wrong*."

Consistency by our interpretation is tied to the change characteristic. Consistency, according to our players, was tied to the NPC's perceived personality traits, presented values, and persuadability level. Figure 34 depicts an overview of the NPC's consistency results according to our players.

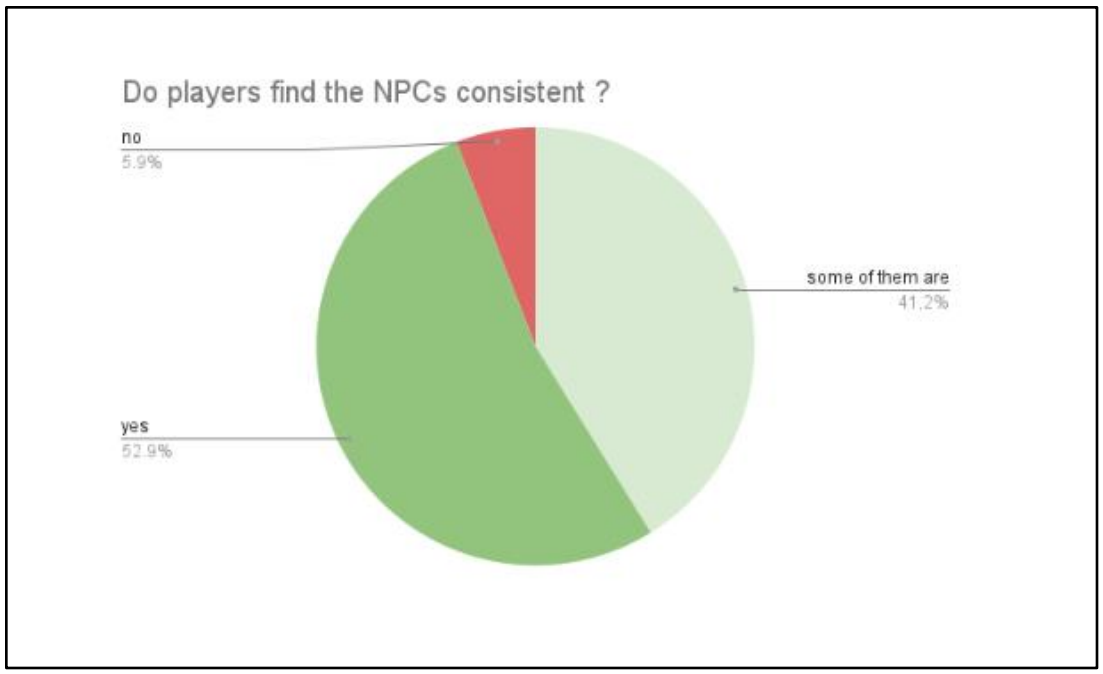

Figure 34 highlights general findings with regard to an NPC's consistency.

As a reminder, while our system implements change, it communicates it at a superficial level. Change is simply displayed in the play through intro and outro statements. NPCs reflect their feelings about an NPC on five levels, but the actual arguments used in the majority of the game remain unchanged, as they are tied to the values themselves. Without multiple-flavor text added to each value-based argument, it would be difficult to thoroughly understand change as a standalone criterion in this project.

Having said that, some players felt that NPCs changed their personalities with newer iterations, while others felt that the NPCs were consistent, mainly due to their beliefs and values. Examples of statements on consistency include: "*Carrot was consistent with their thoughts and judgment toward her argument topic";*' "*I feel like*

their personality changed when they came back another time;" "Carrot was consistent with their thoughts and judgment toward her argument topic' and '*Carrot's personality, especially, seemed consistent. Amrock's was somewhat consistent.*' Like predictability, SF-based characters seem less consistent than NP-based characters. We think that the player's own political alignment (most likely left due to results) may influence how they perceive and understand these characters.

We generally consider our results positive. We argue that in evaluating an NPC's predictability, we should balance predictability and unpredictability. We believe that designing NPCs that are too predictable can create monotonous and boring interactions. Moreover, creating entirely predictable characters can create situations where players game the system and thus break the illusion of life, an essential facet of character believability.

On the other hand, creating unpredictable characters can lead to misunderstandings and confusion. Balance is needed to maintain a good predictability measure, which we believe our games support (at least until all interactions are exhausted).

As for change, ideally, we would like our NPCs to change and grow. According to our findings (**Chapters Two and Ten**), NPCs should grow and develop but maintain consistency. Ideally, NPCs should gradually change over time. We believe gradual change is a limitation that our current system does not sufficiently address due to the nature of the demo and the game

2.4.5 On social interactions.

While social elements play a part in generating an NPC's conversations (i.e., NPC starts conversations by gossiping about someone), it does poorly in displaying that information to the player. As a reminder, the social aspects of our system are displayed to the player through a relationship bar (presented as an optional side menu) or referenced by the NPC at the start and end of a conversation. The introductory statements are filled from a template, for example: "I [CNPCOpinion] [BnpcName] the [talkedAboutAnimalSpecies]."

$

I [CnpcFeelings] that [pronoun] is … (argumentative pattern text); could be translated to the NPC saying "I [hate] [Carrot] the [Rabbit]. I [dislike] that [she] is [dating different species]." Similarly, conclusions pop up on the player's screen, showcasing the effect of a conversation, where the player is made aware of a relationship change in status. While the template variables change depending on the NPC's state, we believe it is minimal in showcasing or highlighting relationship statuses or effects.

Consequently, our system's social aspects were somewhat present according to our players. Five participants said they believed that all of our characters were social, seven believed it varied depending on the character, and five believed the characters were not at all social. We note that the quantitative measure is skewed due to how our players interpreted a character as *social*. As such, the following sections focus more on

the qualitative interpretation as provided by our players, proving the validity of social components to some extent.

According to most of our players, social relationships were defined based on the NPC's active or passive role in a conversation. To elaborate, some players correlated social actions with an NPC's ability to judge or observe others. One player said, "They *seemed to have strong opinions on each other, which makes me think they know each other well*." Another remarked that, "*Some [characters] clearly interact with others, while others seem like they just observe*." Yet another player commented, "*There are characters that say Leo is dating someone of another species, and he also talks about his family, so he seems pretty social*."

Others viewed social actions as the character's ability to participate positively in an interaction, while negative interactions were stated as reasons against a character being social, taken literally as a sociable trait. Examples include: "*Amrock seems to be against interspecies relations, and the lion hesitates. Carrot may be more willing, seeing as they supported carnivores in illegal meat buying." "The characters talk badly about each other and do not seem to like each other much." "They seem so disconnected and even hostile to each other that I don't see it happening.*"These mostly describe the characters as exhibiting antisocial behaviors rather than in terms of being social agents.

Let us next examine values and morality, after which we will discuss the effects of value-based design.

2.4.6 On character, values, and morality.

Now that we have discussed believability in some depth, let us turn our attention to values and beliefs. So far, we have seen values linked to multiple believability factors, including an NPC's personality, consistency, motivation, and understandability. Before we examine how our players categorized and identified NPC values according to their moral models (and Lakoff-associated definitions), we wanted to understand what they thought about our characters and their morality, as well as values without giving them explicit definitions of our values and moral system. In the next paragraphs, we highlight how players established and perceived morality through value-based design.

As mentioned in section 1.3, we added some obscure and open-ended questions about the characters involved, asking players about their general thoughts about said characters and if any were memorable, interesting, lovable, or hateful. We then asked if players perceived any moral values and inquired about their general thoughts on the character's morality. Our goal with these types of questions is to elicit more data about value-based design (since this system highlights values). What can we uncover about our characters? We note that this section accounts for players that provided responses in our survey, as blank responses were omitted from our analysis.

According to our players, Amrock was the most memorable character (11 votes) in our game. Amrock was followed by Carrot and Leo, where five players marked them as memorable. Among the responses, players felt they were memorable due to the game's setting (for instance, one player commented on its similarity to a show called

Beastars), the difficulty or ease of arguments, order of appearance, and, in Amrock's case, specifically, his *"stubbornness*."

Similarly, in asking if players found the character interesting, Amrock took the lead, with Carrot and Leo equally falling behind with half the votes each. Players seemed to find Amrock interesting due to his perceived personality, difficulty, and the nature of his retorts (SF arguments).

Strangely enough, almost all of our players unanimously chose Amrock, the SF-based character, as the most hated character in the game. Carrot had three votes, whereas Leo had none. In contrast to this statement, Leo and Carrot were found to be lovable (8 and 7 votes, respectively), whereas Amrock received one lovable vote. The characters were found to be lovable due to their perceived values and agreeableness. For instance, some players mentioned liking NPCs because they were understanding, empathetic, and honest.

Certain themes emerged after asking players about their general thoughts about the characters. Some players felt that the NPCs reflected real-world phenomena; this is expected, since these characters mimic conservative and liberal rhetoric, according to Lakoff's models and metaphors. Another theme centered around morality, where a few players perceived varied morality scales as noticed in characters or as elements in conversations. Other players recalled personality traits, character viewpoints, and difficulty levels as memorable experiences. We next asked our players their thoughts on moral values and the character's general morality as presented in our system.

Generally, our players felt that our characters had moral values but expressed morality differently. We note that some players misunderstood the concept of moral values in our question, as their responses associated values with ethics or righteousness (i.e., if they thought the characters were morally good). For instance, one player said: "*All (characters) except Carrot. Carrot was shady and mean."* Another reflected on Carrot's positive stance: "*Yes, I think Carrot had morals in the sense that she seemed the nicest and most open-minded.*" Another player quoted a lack of morals, stating, "... *especially Amrock. He has a discriminating mind, though not expressed...*"

Others mentioned statements closer to the general metaphors of Lakoff's moral families (SFM and NPM as conservative and liberal metaphors). For example, one player said: "*The wolf guy held views that matched up with more* ***conservative things****, while the other guy...*" Another player stated: "...represents a social group in today's society, i.e., **republican, liberal, democratic** ideologies and in turn values."

Others referenced or hinted at specific moral values in their statements. For instance, one player commented about Leo saying: "... *he* ***understands*** *the point the other person is trying to make,*" emphasizing *empathy*, while another said: "*I believe Leo and Carrot valued* ***helping others****, whereas Amrock valued* ***themself*** *more,*" highlighting *nurturance* and *development* metaphors from the NP model and *self-interest* from the SF model.

Other players related morality to the discussed topics or issues the characters brought up, citing the themes that emerged, like buying meat on the black market or falling in love with other species. Others commented on the diversity of the implied

moral systems; for example, one player said: "I *think both have their own morals in their own ways*," and another said: "*...They do but differently. They all see the world and what's right and wrong differently*."

We followed up with another open-ended question, asking players their thoughts (more generally) on the characters' morality. Similar to moral values, some players thought morality equated to righteousness, where players mentioned the NPC's stance or discussed topics. Others noted the diversity of the moral models. For example, one player said: "*It is surprising to see that people living in the same place have such drastically different viewpoints*." Another mentioned," *None of the characters seem to be entirely in the right, but the morality varies between them*." Other players mentioned their feelings on the character's morality; these statements expressed elements of their gameplay or interactions. For instance, one player thought the character's morality was *interesting* though *predictable*, while others stated it was *hidden* or *realistic*. Two players indicated feeling confused, while another disliked how the moral system was structured.

So far, we have discovered that players associate moral values with good ethics, specific values and general political stances; we also note that some players perceive different levels within our moral systems, adding a sense of uniqueness to characters. We note that some of these results echo statements from our previous study (**Chapter Seven**).

2.4.7 Moral value alignments and identification.

Now let us look at how our players identified and categorized our NPCs' moral values according to Lakoff's specified definitions as referenced in **Chapter Five**.

Before we begin, we would like to remind our readers that our cast of characters includes:

- A pure SF-based character (Amrock the Wolf).
- A pure NP-based character (Carrot the Rabbit).
- An NP-based character with an SF-value (Moral Health) (Leo the Lion).

We note that the depth at which a character hates or likes a value ranges at runtime and depends on the character's moral model. However, we note that it is influenced by our configuration file (please refer to section X for additional details.)

Let us first look at Amrock, the wolf. We generally found that players could correctly identify Amrock's core values and beliefs. The collection of charts in Figure 35 gives us an overview of the players' value identification and alignment with regard to Amrock.

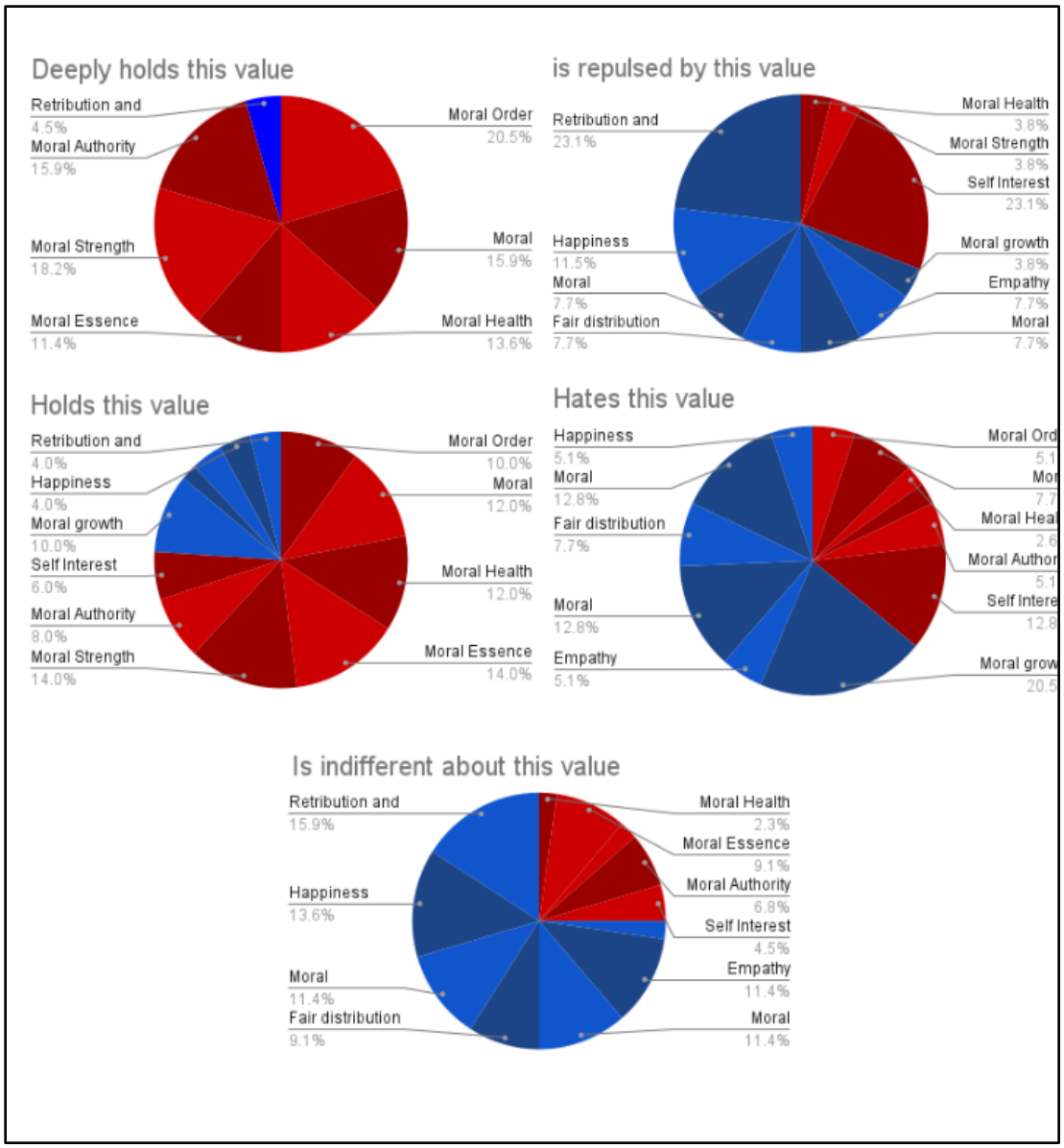

Figure 35 presents Amrock's values and beliefs according to our players. The ranges include deeply holds this value, holds this value, is indifferent about this value, hates this value, and is repulsed by this value.

Please note: we color-coded pie chart slices with alternating colors, with red associated with values belonging to the SF model and blue associated with values

belonging to the NP model. From our charts, we can see that our players mostly identified Amrock as an SF-based character. Most of our players thought Amrock deeply cares about the values of Moral Health, Moral Essence, Moral Strength, Moral Order, Moral Authority, and Moral Boundaries, all of which are SF values. We also noticed a similar trend in values the player thought Amrock held, where most values are red. In contrast, most players also identified that Amrock mostly hates and deeply hates NP-based values; this suggests a general understanding of values and their moral models. Let us look at Carrot's data next.

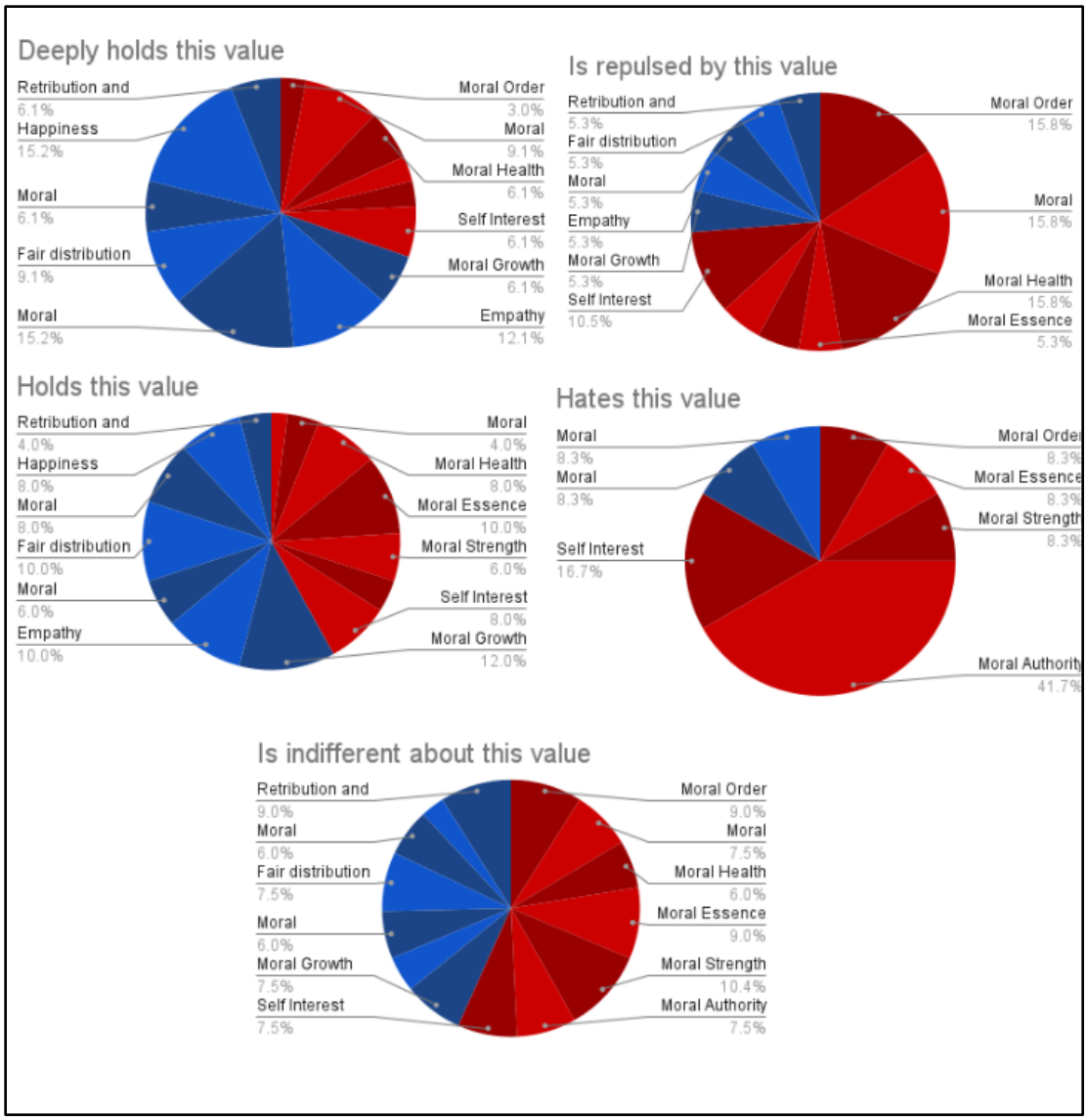

Figure 36 presents Carrot's values and beliefs according to our players.

We find the opposite transpiring when we examine the players' responses regarding Carrot, the NP character. The majority of our players believed that Carrot cares and deeply cares about NP-based values (values in blue). In contrast with

Amrock, most of our players also associated SF-based values with values Carrot hates or deeply hates. We also note that a larger player error exists in NP-based characters than in SF-based characters. Before discussing these results, let us look at Leo's data. Leo's data results were similar to Carrot's. Figures 37 and 38 highlight the differences between Carrot's and Leo's data, focusing on their liked and hated values.

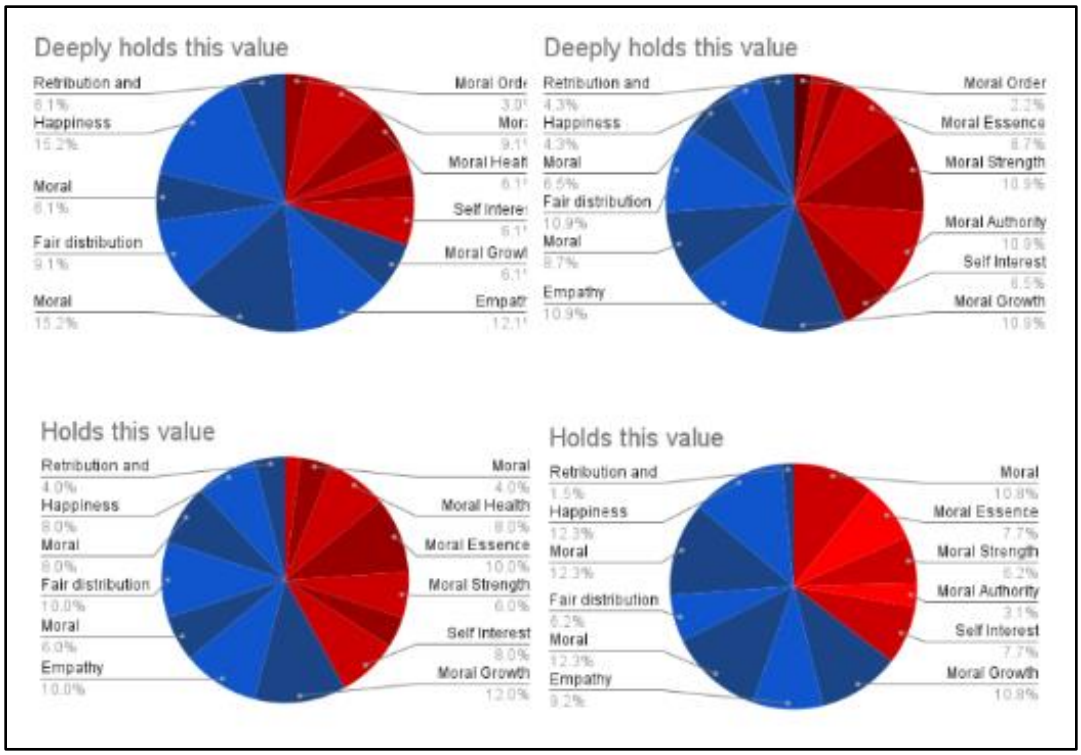

Figure 37 presents Leo and Carrot's data side by side. Left to right presents Carrots and Leo's held beliefs accordingly.

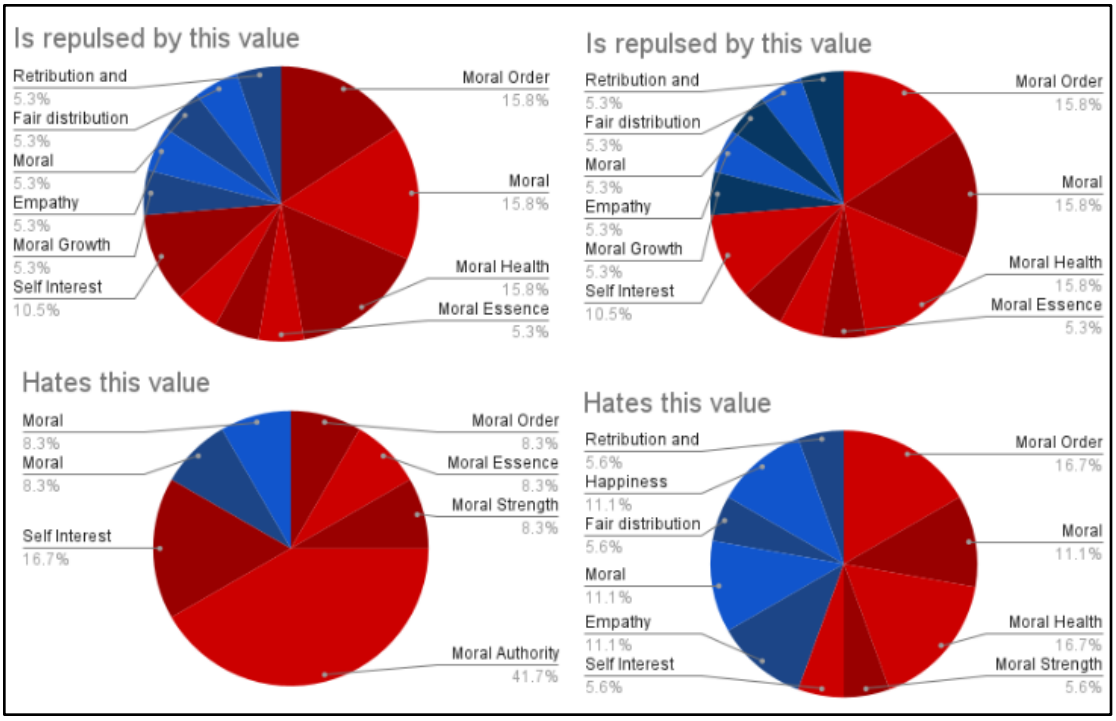

Figure 38 presents Leo's and Carrot's data side by side. Left to right presents Carrot's and Leo's hated values accordingly.

Generally, players believe that Leo is mainly an NP-based character, which is established on their value assignments (as reflected via pie chart colors). We note that these two characters (Carrot and Leo) portray a similar model structure at the extreme ends of their liked and hated scales with minor differences in (secondary) ranges where most players thought Carrot hated SF-based values more than Leo, although not with our configured value.

Strangely enough, some players thought Leo both hated and was repulsed by our configured SF-value *Moral Health*; around the same number of players thought Leo was indifferent about this value, and the minority thought Leo cared about this value; we believe this suggests some assumed correspondence between values that

belong to the same model. We also believe that several factors may have contributed to the increased error rates with our NP-based characters. If we examine the data reported earlier, we can see instances when players cited that NP-based characters were easier to persuade, while SF characters were more difficult; this suggests a shorter interaction loop with NP-based characters as the character moves on once convinced. Players also reported that the NP-based character is more predictable than the SF-based one.

In general, we found that moral values are correlated with some believability criteria, such as a character's personality and motivation. Other elements had poor quantitative results, but player feedback indicated otherwise due to the player's misinterpretation of the question. For instance, relationships were present according to our players' comments, though misinterpreted as negative relationships, earning the character(s) negative social scores. Players also mentioned a sense of uniqueness or variety in a character's moral stance. We also note that unpromoted players linked morality to specific deep values or grouped them into groups according to political labels in the states (liberal, conservative, democratic, etc.).

The difficulty level of persuading a character seemed to make the character and gameplay more interesting to our players, although some confusion was noted. Further testing is needed to determine what role difficulty plays in character believability.

Upon displaying the NPC's deep values to players in layman's terms, we found that players could identify and link values to their moral models (SF and NP). We

believe this indicates that players can successfully paint a mental model of NPC values or beliefs with specific value misconceptions (e.g., moral health in Leo's case).

We then asked our players to compare our game to other moral systems. We included a range of examples to highlight diverse moral systems, such as those discussed in Chapter Three. We note that players who did not recall or played moral-based games were instructed to write NA as their response.

Some players compared our system's characters to explicitly authored characters. For instance, one player said: "*Amrock is reminiscent of racist characters like Kenny from The Walking Dead telltale series*." and "*The wolf guy reminded me of the leader of the StormCloaks.*" We believe this adds a new avenue for authoring belief-based characters. The implications of adding personalities that are generative, seem unique, or indicative of moral reasoning in background characters can play a major role in creating a more immersive world. We do think that applying the concepts in this system in a larger game setting is warranted to assess this claim further.

Some players also compared value-based morality to choice-based morality as a structure, mentioning games like Telltale's and Mass Effect or other RPGs like Far Cry Primal. One player felt that the characters were too friendly in comparison to specific game characters. Lastly, another player mentioned experiencing a larger avenue of morality, though not seeing characters act on it.

We then asked our players if they thought morality, as presented in our system, makes the characters more or less believable and more or less engaging. We note that, at this point, we presented our players with a believability definition and created two

measures highlighting the extreme ends of a scale; we categorize three as believable or somewhat believable. The full list of questions and scales can be seen in Appendix A.

We note that the majority of our players felt that value-based morality made our characters more engaging (58% scored a four or above, ~6% below a score of 3) and more believable (68.8 % scored above 3, 18.8 below a score of 3).

3 Conclusion

In conclusion, our report indicates that our characters are generally believable or at least somewhat believable. We uncovered multiple avenues in which values play a role in NPC believability and morality. Specifically, we discovered a strong association between values and a character's perceived personality. We have also seen multiple instances when, according to our players, a character's values impacted their motivation, change, and the overall illusion of life (e.g., agents seen mimicking life, reactive.)

Players could also identify an NPC's specific values and categorize those values into broader camps or definitions. Values also play a role in how players perceive morality within our characters. Some players described morality according to broader definitions, while others noted specific definitions as moral behavior. We note that while some players confused morality with righteousness, their descriptions pertain to the concept of morality. They were able to perceive characters as morally good or bad. Some players also perceived a character's sense of morality as being multifaceted,

expressing different levels of morality, which we believe adds a sense of uniqueness to characters.

Overall, players felt that value-based morality is similar to choice-based morality as a structure (as described in **Chapter Three**), though an additional study would be needed to verify the results.

3.1 Limitations and Implications

While our study offers promising results and could serve as a basis for value-based design, we believe that further research into specific avenues is warranted to gain deeper insights. For instance, we specifically saw instances when difficulty played a role in the player's perception of the characters, but the nature of our question did not aid in uncovering gameplay difficulty as a topic. The presented study is also somewhat small. Players could converse with a limited set of characters during a period of 20–45 minutes. Although these characters are generative in nature, the overall models (characters appearing as SF or NP) were authored to appear as such in the demo; we believe we can utilize the generative nature of the system to create more morally diverse NPCs in the future. Players also experienced a limited set of possible conversations due to the nature of the system (based on the allocated patterns at runtime and their stance) and available authored content at the time of the study. While further research is needed to gain deeper insights, we believe our study provides promising results and could serve as a basis for moral value-based design.

Chapter Nine: Factions—Morality, Values, and Beliefs

Acknowledgment(s): The contents of this chapter have been previously published elsewhere [7].

1 Introduction

In **Chapter Seven**, we found that the generative nature of the system and the multitude of values helped shape our characters, ones with seemingly diverse moral standards. We believe that value-based designs could aid and further improve background characters in games. We also note that our last system (**Chapters Five and Six**) did not, unfortunately, highlight change, an aspect of character believability that is often missing from many academic systems and games.

So how can we incorporate change while introducing similar concepts to our previous iterations to background characters? One general area that often includes morality, highlights change, and features many characters is *factions*. In this chapter, we will examine morality through a faction lens. We start by identifying and analyzing popular factions in media (films and shows) and games. Since our research focuses on games, we created a faction taxonomy whereby we highlight moral systems, use cases, common elements, and differences. We end our chapter by showcasing our current

progress in faction design, one that incorporates character change and growth as a main mechanic.

2. Faction Design

Factions play a significant role in many game genres, from RPGs to RTS and simulation games. So what are factions, what do they consist of, and how are they designed? As we will see in greater detail in later sections, factions in games and other media share these basic features:

- **Factions contain groups of characters.** A faction contains a group of characters coming together for a goal or a cause. Different characters can have different goals and agendas but believe in the faction's stance.
- **Factions stem from conflicts.** Factions are usually formed as a result of conflict between different opposing groups, usually arising from a violation of character or societal beliefs and values.
- **Factions deal with political and moral situations**. Characters often join or abide by factions based on their shared beliefs or perspectives on morally ambiguous situations.
- **Factions reward committed members.** Factions often reward members by providing security or increasing monetary, social, or personal values.

- **Factions are hierarchical by nature.** Faction members enjoy different roles, from security to leadership. Members are often part of a hierarchical structure with certain tasks dedicated to specific members (where appropriate).[35]

While factions across media generally share many of the outlined features mentioned earlier, the details and implementations differ in games. In game design, *factions* as a term is used to describe numerous group dynamics that appear differently, depending on the genre, the game's world, and the designer's intent.

This chapter will identify different types of factions in games and how they appear and work. Using case studies of exemplary games in the field (such as those that won awards or are distinguished by a large player base), we will provide a taxonomy for factions in games.

Please note that the examples used in our taxonomy are meant to illustrate different faction systems, which we believe can be applied to other games of the same nature, but our review is far from exhaustive.

[35] Even factions with equal status among members (e.g., no named leader) tend to have marked differences in terms of power (e.g., a character whose views hold greater weight than others).

3. Faction Systems in Media

We chose movies and TV shows as non-game faction representations for two purposes: One, to explore how factions and faction characters appear in popular media. Two, to serve as a secondary point of analysis in creating believable faction characters.

This section covers exemplary factions in popular movies and TV, illustrating how characters appear and discussing their overall design. First, we examine popular media showcasing two primary factions at war. One strong example is the movie Avatar [26]. It tells the story of humans colonizing Pandora to acquire Unobtanium, a rich mineral that can substitute for the earth's scarce energy resources.

Immediately, we notice two sides emerging, each with its own goals and agenda. The human side wants to save their planet by obtaining Unobtanium, and the Na'vi side fights to protect their planet and resources. As we will see in later sections, these conflicting goals align with game faction design. Factions start over a conflict, and in this particular case, it is a resource.

Let us take a deeper look at some of the characters involved in each faction. Jake Sully, a former marine, embodies an avatar (a substitute body to blend in with the Na'vi) with the goal of spying on the Na'vi. Jake initially takes this mission in exchange for medical treatment from his commanding officer, Colonel Miles Quaritch. Eventually, Jake meets Neytiri, a female Na'vi who rescued him from alien dog-like creatures. So far, we note two themes emerging: hierarchy and representative characters (Jake and Miles for the humans and Neytiri for the Na'vi).

One aspect of note is the characters' ability to grow and possibly change sides. Jake, for instance, eventually changes missions and sides from spying on the Na'vi to protecting their home world. However, Jake's change in allegiance does not happen abruptly; it takes time. We witness Jake's avatar-self welcomed into Na'vi society; he mingles and lives among them for months, building relationships. We also notice some events that push Jake further toward changing sides, like Jake falling in love with Neytiri, giving him a more empathetic view of certain events. We also note that these events are not limited to main characters; these events affect other humans and Na'vi characters, as well. The Na'vi become more accepting of Jake, and some humans also grow more empathetic—while others, like the main antagonist, Colonel Miles Quaritch, grow suspicious.

Lastly, we notice events happening to the characters that cement their stance or make the characters effectively change sides. These events drive the plot forward. On one occasion, we witness Miles bribing Jake (a paraplegic) by fixing his legs in exchange for him spying on the Na'vi. This highlights a personal need/want for Jake. While Jake initially accepts the deal, he turns sides and discloses the truth. However, this goal affected the character's growth and was eventually attained in avatar form.

We also notice other significant events that elicit a moment-of-decision response from the protagonist. For Jake, it is the invasion of the Hometree, a sacred Na'vi spiritual tree. Jake had to choose whether to fight the Na'vi or turn against the humans.

For a second case, let us look briefly at the MCU character Nebula [62]. She is the adopted daughter of Thanos, a genocidal warmonger bent on eradicating half the galaxy's population for the prosperity of the remaining half. Through the movies, particularly the Guardians of the Galaxy series and the Avengers Endgame saga [128], [129], we see Nebula go from being subservient to Thanos to betraying him and fighting alongside the Avengers. Nebula's transformation is emphasized when Nebula time-travels from the past to aid Thanos against the present-time Nebula, who fights against him. Similarly to Jake, we attribute Nebula's development and change in allegiances to her exposure to the opposing side (e.g., "captured" by the Guardians), completing her personal goal (besting her sister in a duel), and fulfilling a missing need (fostering a positive sisterly bond with Gamora). Nebula interestingly developed a goal to seek revenge on Thanos, aligning with the Avengers and the Guardians' goals—establishing common ground. In this second analysis, we notice a theme of change emerging yet again.

We also notice that change took place over time and with the fulfillment of character needs, albeit not in the expected form. Before concluding this analysis, let us look at an example of what happens after faction change—from the Walking Dead TV show [37], [38]. We chose this show because it incorporates many factions and betrayals in a post-apocalyptic zombie-filled world. While there are many characters to choose from, we decided to examine the character Negan—the vile, evil, but oddly charismatic leader of the Saviors faction.

In the show, we witness how Negan eventually alters his ways once he loses and is captured by the leading protagonist group, Alexandria. While Negan did not accomplish his goal of defeating Alexandria, he changes due to his sentence exposing him to Alexandria's way of living and forming a father-like relationship with Rick's child, Judith. While we can see some common elements that align with the previously mentioned examples (such as relationships, goals, and needs), what is interesting here is the after effect of Negan joining Alexandria.

While the character admits to liking Alexandria's way of living, he holds some of his former beliefs, such as saving children above others; he exhibits this by running after Judith in a blizzard, protecting her even if it would cost him his life. We also note that Negan's personality factors do not change much (witty and impulsive) despite the character facing a long sentence. One of the "big" events showcasing Negan's growth occurs when a large group of the dead invade Alexandra under the control of a faction known as the whisperers, a cult-like group that lives among the dead and believes that death is inevitable. The existence of this group establishes a common ground between Negan and Alexandra. He fights for Alexandra as a spy, putting himself in harm's way. Despite this significant event, only some characters changed their opinions of Negan while others were wary of him; we attribute this to character personalities and history.

The Walking Dead provides a gradual reveal for what happens after a character joins an opposing side, something difficult to establish in the earlier examples due to the running time of movies.

Our examples here show themes emerging in portrayals of movie/TV factions and their characters:

- Factions are formed based on common ground and/or a shared agenda.
- A faction's characters are often goal driven. Character goals are motivated by many factors, including relationships, ambition, and character desires.
- Characters are dynamic and ever-changing, altering their relationship with their faction(s). Characters can leave factions, join new ones, gain new goals and form new relationships.
- Characters also react as factions shift. They can change, grow, and love or hate new members.
- Events play a significant role in a character's development. Events in media often give characters a moment of self-reflection and the opportunity to change, such as switching factions, presented with often significant consequences.

As our keen reader observes, change is a significant component of faction-based characters, one that is often lacking in game-based factions—as we will soon see.

4. Faction Systems in Games: A Taxonomy

Factions in games are the primary focus of our paper. Here we taxonomize the first two main faction designs as explicit and implicit faction groups.

4.1 Explicit Factions

Explicit factions are those in which faction goals, rules, and agendas are communicated unambiguously to the player. Furthermore, faction members are unambiguous and are aligned with their designated faction (unless specific characters are authored otherwise). Faction systems in these categories follow traditional design standards of their genres, such as using units in RTS games.

We identify five types of faction systems that fall under the explicit faction category. Based on our surveyed games, it is possible to find multiple implementations of different faction systems within a single game. The following subcategories highlight designs by using exemplary games as case studies.

The end of each of these faction system analyses will summarize the key extrapolated features, including the system's moral spectrum, the player's main choices, the player's interactions with the factions involved, the faction's use of hierarchical structures, the faction goals, the differences between representative and background NPCs, and the unfolding story structure. These are the fundamental dimensions found to be common among the defined categories.

4.1.1 Absolute factions.

Absolute factions are the most common grouping of NPC characters found across game genres and particularly in adventure and RPG games. These factions usually adopt a role that is antagonistic to the player. Popular examples include *The*

Legend of Zelda series and the Super Mario series [111], [112]. Let us take the former as an example and extrapolate key design features.

In the most recent installment of *The Legend of Zelda* (*Tears of the Kingdom*) [113], the player takes on the role of Link, the hero of Hyrule. Link is once again tasked with the main objectives of finding Princess Zelda and defeating Ganandorf. As the game progresses, the player notices two main types of NPC groupings, those that aid the player in some way and those that directly oppose the player.

On the antagonistic side, NPCs are usually gated by level, story, or player experiences. The NPCs (usually) have multiple types, are based on some kind of hierarchy, and change their difficulty based on progression. The core design has a single kind of interaction with the player, that is to attack. On the protagonistic side, we have NPCs that aid the player in some way, either as questmasters, teachers, or storytellers. They only help or get help from the player. The interaction between these two-faction systems and the player is clear-cut. The player can either interact negatively with opposing factions or positively with friendly factions. For example, it is impossible for the player to talk with the opposing faction and attacking a friendly NPC does no damage (which holds true for friendly NPC counterattacks as well).

Games with these kinds of systems progress fairly linearly. Both main groupings of factions follow their own unified goal, the destruction of the opposing faction. Although sub-groupings exist, such as the Zora tribe in Zelda, they still carry the same roles and agenda mentioned earlier. Members in sub-factions (on both the "good" and "bad" sides) are usually represented by their leaders or specifically via

authored characters and scenarios. Taking as an example the Yiga clan, which has its leader, Master Kohga. All members of the Yiga clan consequently share the same goal as Kohga: to exterminate Link. The members in this case have little to no control over their own desires. We will see this pattern repeated in other faction-based systems. The same pattern can be seen on the "good" side across different clans, though some side/background NPCs have extra personality portrayed through scenarios or quests.

Figure 39: Figure depicts a group of Bokoblin following a Boss Bokoblin, an example of absolute faction hierarchy found in Zelda: Tears of the Kingdom

In summary, these faction systems share the following characteristics:

- Does not allow the player to align with multiple sides.
- Features a limited set of player interactions per faction.
- Faction members are placed within a hierarchical order.
- Faction members share a single goal without dissenting voices.
- Limited or no differences among background or side NPCs.
- Follows a linear story structure and progression.

The value of a faction is either black or white, with little-to-no shades of gray (players and NPCs assume the roles of heroes or villains).

4.1.2 Role-based absolute factions.

A subcategory of the absolute faction system is the role-based faction system. These factions are relatively common in many game genres, from adventure to horror games. Members of an absolute role-based faction portray a single role, follow the rules, and cannot be interacted with beyond a permissible set of ways. Examples include cops in GTA, law enforcement officers in Red Dead Redemption 2, and guards in Assassins Creed [126, 134, 155].

This faction system shares the same fundamentals mentioned in the absolute faction system. Unlike fully integrated absolute faction systems, these role-based factions are a small component of a larger outer world. One key difference is that these role-based factions commonly share a trigger condition tied to the player's state in the world (if applicable)

Role-based NPCs are reactive to the player and have little to no interaction with other NPCs within the same faction. Characters fulfilling these roles are entirely morally black or white, making negotiations impossible. These NPCs serve as a means of leveling up, adding difficulty to a game, or gatekeeping player progression.

Figure 40: Figure depicts a collection of role-based factions from multiple games. From left to right, Rangers in RDR2, Police in GTA, and Guards in Assassins creed.

4.1.3 Allegiance and sState-based factions.

Unlike absolute factions, allegiance-based factions engage the player's perspective and their choice of moral alignment. Allegiance-based factions are common in RPG and MMORPG games. They are commonly black and white, with shades of gray. Famous examples include games such as Fallout 4, World of Warcraft, and the Dragon Age series [69, 79, 76].

Under the hood, player actions can move their character closer to (or further away from) different factions. Our survey identified two common types of allegiance systems. The first is a reputation system, described next. The second is a system of absolute or accumulated choices that change the state of a given faction. Games can utilize both systems at the same time with potentially frustrating and confusing results.

Reputation is typically seen as the currency the player character (PC) acquires by taking actions in the world; these actions can increase or decrease the PC's status with a faction or change how different NPCs perceive the player in the game's world. Reputations can be linear or multidimensional. They can be globally or locally (character-based) scaled. We further examine reputation systems by using Fallout 4 and World of Warcraft as popular examples.

In Fallout 4, the player is an apocalypse survivor traversing the wasteland. Players eventually meet factions like the Institute (wishing to protect and advance technology) and the Brotherhood of Steel (wanting to advance but regulate the world's technology). Upon interacting with members of each faction, the player discovers that each of these factions has its own goal and global reputation bars.

The player can complete quests to increase their standing with their chosen factions and further develop the story. Moral choice is defined by the player's personal preference. They may join whatever faction their character deems moral. Doing favored actions can increase their standing, while unliked actions decrease their standing. Players can join many factions simultaneously despite their different agendas (this is relatively common in games such as Skyrim, Fallout, and Dragon Age). However, some gated progression elements may be blocked based on PC reputation or choices.

Each of these factions has representative characters that (could) aid the player in the world, as was the case in Fallout 4. These representative characters can observe the player and like or dislike their actions on a local reputation (like/dislike) scale. In Fallout's case, these local reputation systems are further tied to the global faction

reputation system. We note that faction reputations and questlines are also intermingled. For instance, completing the *Brotherhood of Steel's* quest line makes the player an enemy of the *Railroad* faction; keeping all of this overhead in mind might be too daunting for particular play styles.

Fallout also incorporates absolute choices into their faction designs. For example, Preston Garvey, one of the leaders of the *Minutemen* faction, can become angry if the player becomes a boss of Nuka-World, a lawless city of raiders. If the player refuses to repent (presented to the player as an absolute choice), they will become exiled from the Minutemen. This particular choice effectively changes the player's stance with the factions involved, regardless of any reputation accumulated thus far.

Taking World of Warcraft as another example, players can belong to one of two opposing factions: the Horde or the Alliance. Reputation in WOW is a currency (which players "grind" to get) used to increase their standing within a faction. There are sub-factions in the game, some of which adhere to the main game's two factions (e.g., allied factions), while others provide lore, remain neutral, and similarly reward the player with quests or unlockable items in exchange for faction points. WoW utilizes the allegiance-based structure in both absolute choices (portrayed through the player's choice of a given race) and a reputation structure. With that said, the overall structure of WOW feels binary, with each PC and NPC belonging entirely to one particular faction. It mixes the above mentioned absolute structure with the allegiance structure for their subfraction designs.

Like Fallout, WOW utilizes representative characters in each faction, such as Thrall for the Horde and Wrynn for the alliance. These characters, among other representative characters, are featured in specific game scenarios, major questlines, and raids. They showcase the faction's ideals and what it stands for. While these heavily scripted characters are unique and quite enjoyable, background characters in these factions play little to no role aside from simple reactions.

Taking guards, quest givers, or vendors as examples, we can see NPCs react to the player or enemy NPCs but pay little to no attention to other NPCs within the same faction. This type of reaction stems from superficial reasoning capabilities. A common example is NPCs attacking opposite faction members merely based on the character's allegiance. The lack of intelligent NPC-NPC interaction in an open-world scenario detracts from the overall believability of the game. It creates characters that are mere copies of each other with seemingly simple goals and little to no personality traits, which, as we stated in **Chapter Two**, are essential characteristics of believability.

As with absolute factions, NPCs in allegiance-based factions follow a hierarchy. Faction members play different roles, from leaders to guards and grunts, each acting according to their own individual tasks. While characters performing these roles are considered contextually believable [165], this approach offers no route to greater believability or larger roles for these NPCs.

Lastly, allegiance-based factions often incorporate branching storytelling structures dependent on the player's choice of allegiance. The choices the player makes alter their standings with the interacting faction. In some cases, such as *Fallout 4*, key

moments can alter the player's path, blocking specific routes (conflicting faction stories) and resulting in minor alternative endings.

In summary, the common characteristics of these factions are:

- Factions have different moral alignments depending on a given faction's stance. Faction stances can be black, white, or shades of gray.
- Players' moral choices either align with factions or against them. Some choices are inaccessible, depending on players' choices with the factions involved.
- Greater player interactions are possible. However, one type of core interaction per faction remains true (faction as hostile or friendly). Factions are more dynamic and can change state depending on players' actions.
- Allegiance-based factions share role-based hierarchical systems (e.g., leader, guard, or vendor).
- Faction characters share one end goal with additional specific goals authored for specific characters.
- Factions have some highly scripted characters that represent faction members. Background and side characters are basic and somewhat limited, with little to no characterization.
- A PC's progression and story follows a branching structure, determined by the player's choice of allegiance and the state of the world.

4.1.4 Unit-based factions.

As the name suggests, this category of faction systems utilizes characters as simple superficial units that fulfill or hinder the player's goals. They are common in the RTS genre. Examples include the Civilization series, Age of Empires games, and Human Kind [8, 50, 52].

In these games, the player typically assumes control of their chosen faction to conquer the world or otherwise win against other factions within the world. The player takes control of characters (units) at a macro level; there is little to no characterization of units within the game. Characters simply fulfill their roles (e.g., army units attacking or defending the player's base). While these characters lack the believability characteristics mentioned in **Chapter Two**, they are still seen as contextually believable since they fulfill their roles at a macro level. While incorporating believability characteristics on units could improve some games, it could also detract from the player's focus at the macro level, which is key to this type of game.

Let us examine the Civilization series to identify more design features. In Civilization [136], players assume control of a faction, aiming to expand their empire through diplomacy and/or warfare. Players can take control of other factions and territories through gameplay, making them puppet-like states. As with absolute factions, the unit-based factions either support the player or play a role antagonistic to the player.

Figure 41: Overview of the game Civilization 3. Figure shows unit icons on a map structure.

We also notice that faction members follow the player's goal or the system's assigned faction goal. The player's choice of factions is superficial at best; these games focus on strategy rather than *characterhood*.[36] For instance, the player's choice of an empire provides them with either cosmetic changes (e.g., different architectural

[36] *Characterhood* is the NPC's ability to act in an assigned role convincingly. The term is borrowed from Dennett's notion of personhood, emphasizing the importance of an NPC's ability to act rationally (perceived as a rational being) and intentionally (actions with perceived intentions) to qualify as a basic NPC [49,48].

buildings or unit costumes per faction choice) or tactical advantages (e.g., different status points for each faction). As far as we know, there are few-to-no representative characters; if a character represents a faction, it merely provides a meta-commentary or aids the player in a tutorial-like fashion. Later Civilization games (5 and up) introduced the concept of "heroes." These heroes provide tactical benefits to the player, like increasing specific status points or unlocking special abilities, but do little to add to the characterhood of the units involved. Lastly, units are differentiated by their roles. For instance, civilians in Civilization aid in building and expanding cities while army units attack or defend the cities.

Surprisingly, the game lacks a transparent hierarchical system compared to other faction systems. We believe this is due to the lack of representative characters. For instance, NPCs follow the player's commands on the player's side, and there is little-to-no difference among units in rank. Though, as mentioned earlier, some units are special (heroes), they still act the same and have little-to-no impact on other units. Some RTS games feature special units like leaders or kings on enemy sides, but they are superficial units; they lack substance compared to other faction games. For instance, unit-based RTS games lack any causal relationships between characters in leadership roles and other faction members, especially when compared to other faction systems.

The relationship between hierarchical units becomes more apparent compared to other absolute or allegiance-based faction systems. For example, in Halo [25], if the player takes out a leader NPC, the grunts scatter—clearly expressing a causal

relationship between NPCs in hierarchical roles, which is missing from unit-based faction systems.

Unit-based factions drastically differ from the previously mentioned faction systems since there is little to no characterization of units and little to no agency in response. Player choices are limited to cosmetic or tactical choices.

In summary, these faction systems share these characteristics:

- Morality is undefined narratively. Units play a supportive or antagonistic role to the player.
- Player choices in unit-based RTS games are always in support of tactical actions. There are few-to-no moral choices available in these games.
- The player's interactions with factions are limited to tactical actions.
- The faction's use of hierarchical structures serves flavor and narrative purposes only. There is no actual causal relationship between member ranks.
- All units share the same faction goal, either to serve the player's goals and agenda or to oppose them.
- Unit-based factions lack representative characters; there is little-to-no difference between NPCs in unit-based RTS games.
- Unit-based factions can include a semilinear story in campaign modes ending with different narrative flavors, depending on the chosen faction.

4.1.5 Player-based reflection factions.

Rather than using factions in the traditional sense, some games utilize faction design to portray the consequences of player actions in the world. Similar to other factions mentioned in this taxonomy, these factions typically have a representative character and a hierarchy-like structure. We will use the game Vampyre [139] as an example of this design and extrapolate key features.

In Vampye, the player takes on the role of a doctor turned vampire in 20th-century London. Throughout the game, the player can act either as a monstrous entity draining blood from victims or as a healer, helping communities recover from the plague afflicting London. If the player takes the more sinister route, farming blood, leveling up, and gaining experience become easy. On the other hand, if the player takes a morally righteous path, the game becomes more challenging, as there is less experience gained from victims.

The faction design in this game takes on the appearance of community zones and locations. Each zone has a collection of NPCs representing their citizens and a few characters at the core, acting as “pillars” of that community. During critical moments in the game, the player can kill, help, or mesmerize these pillars. Depending on their chosen action, the community (faction zone) may suffer; this is shown to the player via a community (zone) health bar. The community’s health bar acts like a reputation bar in other faction games. In addition to pillar interactions, doing positive or negative deeds in a zone can positively or negatively affect that community’s overall health.

The use of faction systems in this manner is interesting and creative, giving the players a sense of dread and consequence to their actions. However, members of each faction feel rudimentary; there is little characterization besides what is scripted for some of the citizens. The main interactions consist of semi-optional side quests or visiting NPC houses and choosing to drain or not to drain their blood.

Unlike other factions in this taxonomy, characters do not share faction goals or act to realize a common agenda. The factions are used to portray consequences and enhance the game's overall story. The hierarchy exists not in a traditional sense, but in the NPCs' impact on the health status of their corresponding zones. For instance, if the player kills a pillar of the community, the community suffers and, in some cases, beyond repair; it is heavily choice-dependent.

Other games utilize similar faction structures as part of the overall game to show the consequences of player actions. In Dragon Age: Origins, for instance, zones may become inaccessible due to the player's inaction, the NPCs in these zones act as citizens and have little to no characterization. Reflection factions differ from other mentioned faction systems in the faction members' reactivity to the player. In Fallout, for instance, members can react in a hostile way toward the player, depending on their actions. Reflection-only factions show the consequences of the player's actions without much reactivity on the individual NPC level.

In summary, these faction systems share the following characteristics:

- Morality is consequential, depending on the player's actions and the world's state.

- Factions reflect the player's choices and possible consequences.
- Various player interactions are possible; however, they are dependent on the game's features.
- Hierarchical structures are consequential. Factions reflect the player's interactions with key figures in the world.
- Representative NPCs are highly consequential to the faction's state (if they exist). Background and side NPCs have little-to-no characterization.
- The faction's state adds flavor to the overarching story, usually reflected in branching storytelling systems.

Figure 42: Figure displays a faction zone's status in Vampyre

4.2 Implicit Factions

Unlike explicit factions, implicit factions are usually ambiguous, unclear, or left to the player's interpretation. This section will explore common design elements found in implicit faction systems.

4.2.1 Implied allegiance-based factions.

Implied allegiance-based factions, like explicit allegiance-based-factions, suggest a relationship between a character and its faction. However, in the implied case, there is no system-level connection between the two elements. Dragon Age: Origins [21], for instance, utilizes both explicit and implied alliances in the characters' designs.

Here we focus on the implied aspects of the game. Let us take the characters Oghren, a dwarven warrior, and Wynne, a human mage, as primary examples. Upon

meeting Oghern, the player realizes their race—in this case, a dwarf—and connects this character to the dwarf race faction (citizens of an Orzammar). Other dwarfs exist in the game and usually it is assumed (and revealed through lore) that they hail from the only remaining dwarven city surviving the Darkspear (the main antagonistic force the player faces).

Unlike explicit allegiance-based factions, characters here have little to no influence over (or from) the factions to which they are assumed to belong. Characters like Oghern and Wynne seemingly belong to their corresponding class or race factions (Ozrimar in Oghern's case and the magic circle in Wynn's) but do not share a global reputation system with the faction. In fact, the main deciding reputation factors for both factions are driven by player choices, quests, and encounters in faction areas.

These elements are separated from the representative characters. Instead, the representative characters have their own individual friendship/romance bar. The game does present the player with opportunities to increase this friendship bar while visiting each of their respective faction locations—played through a few cinematics with branching dialogue or possible gift items the player can collect. We note that representative characters can react in their factions' locations, but this is limited to flavor text and the characters' approval of the player. For instance, while in the magic circle's tower, the player's choices impact Wynns's approval rating. She can leave the player's party, attack, or stay with the player, but the magic circle's resolution depends on the player's unique choices.

In summary, these faction systems share these characteristics:

- Morality is based on the player's perspective and choice of alignment. Implied allegiances typically include shades of gray whereby players' choices account for their moral standing.
- The player's main choices alter personal character relationships rather than faction relationships (with the exception of authored scenarios).
- The player's interactions with representative characters have no real consequences on the implied faction unless explicitly authored.
- No actual hierarchical structure exists between representative characters and their implied factions.
- Representative characters are highly scripted and authored; background characters have little to no characterization.
- Implied factions are typically showcased through branching storytelling structures, and faction (representative character) choices could affect the overarching story, particularly in the player's personal relation to the characters involved.

4.2.2 World-based factions.

These factions are based on player exploration or revealed when certain conditions are met. World-based factions usually share one goal, are unchanging, and do not impact the story in any significant way. They exist to deepen the world and provide players with additional context and enjoyment. These faction systems are usually found in games with an open-world setting, such as RDR2 and Skyrim [17, 126].

Let us take RDR2 as our primary example. As the player progresses through RDR2, they may run into different groups of NPCs with a shared agenda; this usually plays out as a scenario. For example, as players explore the land, they may notice lit torches moving in a pattern somewhere in the distance. If the player follows the light source, they will encounter a group of KKK wizards performing a ritual. The ritual plays out with NPCs lighting a cross on fire.

The fire eventually gets out of hand and spreads to some members. Since RDR2 supports an open-world system, players can interact however they wish. However, these particular instances will always follow their scripted scenarios; no matter what the player does, it ends via a confrontation and the imminent death of some members.

Because the game supports a fame/infamy reputation system, the player can act with or without "honor." The player can speed up their deaths by adding more fuel to the fire or let them die as the scenario plays out. The player can also choose to capture or kill the remaining NPC.

In world-based factions, the NPC's moral alignment is only revealed if the player encounters a particular scenario. Based on context, the player presumes all present NPCs are part of a given faction. No additional information is revealed. We cannot, for instance, know who these characters are or gain additional knowledge about any scenario characters involved unless otherwise scripted.

In summary, these faction systems share the following characteristics:

- Moral alignments depend on the presented scenario. World scenarios can range from black and white to shades of gray.

- Player choices are usually limited and defined by the unfolding scenario.
- The given scenario constrains the player's interactions with the factions involved.
- Hierarchy is circumstantial and dependent on the scenario.
- Members are role-based, reacting to the player in whatever way their roles call for. There are little to no differences between faction members.
- Factions do not contribute to the main story but enhance the game world.

4.2.3 Procedurally generated factions.

Games have used procedurally generated content since the release of Rogue [152]. Games have now evolved to the point of generating and simulating large-scale societies and worlds filled with characters. Like-unit-based faction games, most games with procedurally generated factions treat the player as an overseer of the world rather than a character in the world. Famous examples include Black and White, WorldBox, Caves of Qud, Dwarf Fortress, and Ultima Ratio Regum [15, 24, 71, 72, 90].

Dwarf Fortress simulates a colony of dwarfs generated within a rich world. Each dwarf has unique personality traits and history. As we play and observe characters in the world (or look up the history of a civilization or specific character), we find ourselves coming up with emergent stories. We may see characters fighting or defending a fortress and taking different sides. In this case, the faction system is semi-perceptual; players assign values and stories aided by a log of events and observable

confrontations. Character behaviors in these simulation-like games are more dynamic, unexpected, and susceptible to interpretation.

In contrast to other faction systems, background characters are much more dynamic; each member is as valuable as any other faction member. Though a hierarchy exists, it depends on the design of the game. Dwarf Fortress, for instance, includes kings, queens, and player-assigned ranks (e.g., noble assignments). On the other hand, Black and White uses unique creatures (deities) that mimic the player, held at a higher position, whereas all other NPCs share the same position.

Lastly, as mentioned at the start of this taxonomy, one game can house multiple faction systems within its design. Let us consider the game Crusader Kings 3 [120] as an example. This is a strategy game in which the player takes on the role of a selected character as a part of a developing dynasty.

In Crusader Kings 3, factions can behave similarly to unit-based and procedural factions, depending on the game's current phase. If we look at Crusader Kings 3's combat phases, we note that units act as a means to an end, a resource the player uses for their tactical and military actions. They serve an agenda and follow the patterns we identified in our unit-based faction systems. However, examining the game outwardly, we realize that it employs features commonly associated with procedurally generated factions.

Characters have small, but differing, personalities in the grand schema of the simulation. In order to take action, the player reads NPCs as data points (that have

certain traits, perks, or titles) rather than living vibrant characters. Interpreting and acting on NPC data can lead to players perceiving and creating emergent stories.

For instance, a player may wish to claim someone's throne (as part of what the game calls a "hostile scheming" action) as part of their dynasty. In order to claim the throne, the player begins a scheme against a target NPC; the NPC needs to be flagged with specific traits and perks, such as a target NPC tagged with "liege." The game then calculates the odds of an action succeeding based on a variety of data points, including ones found on the player and NPC's "spymaster" characters.

In summary, these faction systems share the following characteristics:

- Morality is based on the player's perception. Players can perceive different actions as morally good, bad, or ambiguous.
- Player choices are usually limited and defined by the simulation.
- Interactions with given factions are limited to predetermined actions afforded by the game, such as labor task assignments in Dwarf Fortress.
- The faction's use of hierarchical structures depends on the simulation; in some cases, like Black and White, NPCs have higher standing based on ranks, while others like WorldBox have little-to-no differential ranks among NPCs.
- Characters are unscripted; there is little-to-no difference among characters in the world.
- Stories are constructed by the players based on their perception of characters and the world's state.

5. Toward Growth Values and Beliefs

From our overall taxonomy, we find that most faction systems utilize a representative character that acts in place of their chosen faction. These characters are heavily scripted and play unique roles. While these representative characters in most games are a great source of character believability, they still lack some of the properties that could further improve their character and faction design.

Additionally, we observe that elements like character development or autonomy are rarely found in in-game factions. Instead, characters simply react to the players and their state within the world. This lack of character development and autonomy is further exemplified in background NPCs. Background NPCs comprise most of a faction's population, yet they play little-to-no role in their faction and fulfill minor-to-no believability criteria. While fulfilling a specified role is considered contextually believable for unit-based factions, it presents a missed opportunity for games in which the player directly interfaces with these NPCs.

We believe we can use factions in media (section 4) and NPC believability studies to bridge the gap in NPC faction design. Section 4 provided examples showcasing how characters often change in factions, a missing property in most in-game faction systems (unless a character was explicitly authored as a betrayer).

Here, we highlight our analysis of character growth and change, as we saw from our earlier media examples.

- **Change can happen to multiple characters, not just the protagonists.** Change does not only happen to the main characters but

to other faction characters (e.g., Na'vi becoming more accepting of Jake, Alexandra members accepting Negan, and shifting alliances throughout the arcs of The Walking Dead)

- **Change is often gradual.** Our close examination showed that characters can gradually change if exposed to the opposing side's ideals or environment—or if a character shares a common value with an opposing character.
- **Change can be triggered in multiple ways.** Change can happen by choice, by happenstance, or by force.
- **Change is often beneficial to the character or faction.** Characters changing sides or ideals often provide the other faction or themselves with a need or want. This can be many things: objects, relationships, or something of emotional significance.
- **Relationships can be a key motivating factor for characters to change factions.** We note that relationships do not have to be reciprocal. People associated with a character (may) further like/dislike them based on their personality and stance.
- **Personality factors can make characters susceptible to change and affect how a character changes.** We note that a character can grow and change their personality to some extent. However, in most comparisons, a character's core personality remains the same (such as the character being witty in Negan's case).

- **Events and story beats can push a character to change.** Stories and plots can have events that push characters further toward their chosen or opposing stance. In some cases, we witness significant events that present a character with a "moment of choice" to choose a side. Such a moment is often presented under difficult circumstances.
- **Change is not all-encompassing—some older beliefs and values could remain unchanged.** Once a character changes their faction/stance, they do not necessarily change all their beliefs and values. Their ideology may change, but they can still hold some prior convictions at different levels, depending on the value.

By using the concept of character change, we can create more believable game characters. This begs a question: How can we create dynamic change, especially in a large-scale faction system?

As mentioned earlier, factions stem from conflict, and conflict originates from a violation of a given character's beliefs and values. If we take any of the faction systems scenarios mentioned earlier, we can trace the (representative) characters' alignment to a common belief.

Unfortunately, most of these systems use character beliefs and values as a narrative device rather than a systemic one. To illustrate, here is an excerpt from World of Warcraft's Horde and Alliance faction selection. The Horde's faction description reads:

"The Horde fights for honor and freedom, surviving in a world that

questions their unity and strength."

While the Alliance reads:

"The Alliance fights for justice and duty, fervently devoted to those who fight under their banner."

Given the factions' definitions, we can immediately extrapolate the values in which the Horde and Alliance believe. The Horde believes in freedom, unity, honor, and strength, while the Alliance believes in justice, duty, and community. Now, imagine a scenario in which these values are systemic instead of only being narratively defined.

Imagine, for instance, a Horde guard running into an Alliance member who overpowered a strong monster (not part of the two factions). Would their value of strength overpower their value of honor and protecting their home? What if an Alliance member noticed a noble Horde act? Would they condemn the Horde member or aid them? Can they change their mind about that horde character?

We can immediately imagine more dynamic gameplay scenarios. A character's beliefs and values can influence that character's growth. As established earlier, characters can grow when they are exposed to common elements (e.g., shared character values), relationships, or personality factors—or driven by certain events). Most of the elements above are native to many games today, though rarely integrated into faction designs, particularly for side or background NPCs.

By integrating values and beliefs at a systemic level, we can create characters that grow and evolve, a commonly missing and vital element in in-game factions, especially when compared to faction designs in other media.

6 A Prototype of an Evolving Faction

Acknowledgments: A special thank you to Shi Johnson-Bey for joining me on this project. Shi has contributed valuable feedback, development, and design insights for our faction system.

Figure 43: The game world features a prototype of two faction zones and resource zones.

As mentioned earlier, change is an important and often underdeveloped facet of character believability, more so in in-game faction characters in comparison with their media counterparts. This section presents our prototype for developing a faction system

by which characters grow and change due to various encounters. We note that this system is still in development and serves as part of our future direction and work.

Inspired by games such as World of Warcraft [23], our setting features two races that belong to two opposing factions, often fighting over shared resources, a relatively common scenario in faction disputes. What we hope differentiates our system (and what we aim to test in future work) is the characters' ability to change, adapt, and possibly switch sides. We note that at the time of writing, this game is currently at a testing and simulation level. This section will elaborate on our current system specifications and design considerations as we embark on this new adventure.

As we mentioned earlier, change is often gradual, something that could happen to any character, and is affected by multiple factors, such as a character's personality, values, and current circumstances. How, then, can we make a system that highlights these aspects? How can we apply it to a multitude of characters? In the following sections, we will cover our characters, the driving force behind our system, and possible future gameplay scenarios.

6.1 System Specifications

Let us first examine factions, as they are one of our system's central themes. For simplicity, our system defines two factions: one that values nature and one that values development. Each faction has its own name, ID, member list, faction size, renown, state, and faction goals. Factions are mainly used to track and house characters. These aforementioned properties affect some of the characters' volition, as we will discuss in

later sections. Faction effects are still in development; their primary purpose at the time of writing is to house characters, help establish the character's initial simulation, and provide (possible) character goals.

6.2 Characters

Characters are the central part of our systems. Many elements come together to create our characters, including a character's personality, goals, faction, and relationships. The following subsections will cover our character components in greater detail.

6.2.1 On general character properties.

Each character includes a basic set of properties used to define a character. These properties include:

- **Character name and ID**. Each character has a generated name and a unique identifier determined at run time. These properties are used to reference a character on our system.
- **Character health**. Character health is a crucial component of faction characters due to the combat-centric nature of faction interactions. Characters can lose health points, influencing their actions and renown values.
- **Character race.** Each character has an associated race; the system currently contains Orcs and Humans for testing purposes.

- **Renown**. Each character has a global renown value. The renown value can increase or decrease depending on the character's actions or game events. Renown is constructed similarly to a player's fame/infamy bar in fantasy games.
- **Faction**. Each character can reference their faction class and faction properties, such as a faction's strength, warcry, goals, and name. As mentioned earlier, faction classes are currently under development and primarily act as housing and tracking units for our characters. They still play a role in influencing a few character properties at runtime, such as characters' relationships (as we will soon discuss).

6.2.2 On personality.

Character personalities are structured from the concept of archetype. We believe archetypes help make characters more recognizable from a bird's eye view. Each character is assigned an archetype at run time. The assigned archetype influences a character's base personality traits. Each character's personality comes with a set of mutable and immutable traits. Mutable traits are added to highlight character change and growth. We note that some traits are immutable, reflecting a character's core personality (previous sections of this chapter highlight this phenomenon). We note these archetypes make up our characters' base personality traits. Archetypes also play

a role in our event system and in determining character actions and assigned goals (more on that soon). Currently, our system includes these defined archetypes[37]:

- **The Hero.** The hero character fights for justice and follows the rules of the land. Thus, a hero character has high *lawful* and *violent* traits but low *lying* and *chaotic* traits.
- **The Lover.** The lover is a dreamy character with a heart of gold and often has a *naive* outlook on life. Thus, the lover has relatively high *optimistic*, *empathetic*, and *naive* traits. On the other hand, the lover has low *pessimistic*, *apathetic*, and *worldly* personality traits.
- **The Outlaw.** Unlike the hero, the outlaw is chaotic and uses their slick tongue to conjure falsehoods. The outlaw character has high *clever*, *chaotic*, and *lying* traits but low *lawful* traits.
- **The Explorer.** As the name suggests, this character likes to explore the world around them. They are often *calm*, *worldly* but abhor *violence* and can be somewhat *naive*.
- **The Social.** As the title suggests, this character loves to *chat* (has chatty trait). They can be a little awkward and may overthink things thanks to their highly *pessimistic* traits.

[37] Please note that personality traits are italicized in-text for illustrative purposes. The archetypes and traits are influenced by [1].

- **The Innocent**. The prey of many characters, the innocent, is often *naive*, *empathetic*, and *optimistic*. They hardly ever *lie*, nor do they learn from their experiences. They have low *worldly* traits.
- **The Sage.** The sage is a wise character, often known for their intellect and guidance. These characters are *clever*, *calm*, and *worldly*. They are seldom *violent* or *naive*.
- **The Loner.** The loner character is often *empathetic* and *optimistic*.

The above highlighted some character traits belonging to each archetype. Each character contains a larger set of mutable and immutable traits. While the character's archetype determines their basic traits (and removes contradictions), other traits are assigned randomly upon inception. The personality class currently houses 17 traits, 11 of which are mutable. The mutable traits could change gradually due to a character's goals or participation in an event. The trait values range from one and are maxed out at one hundred.

6.2.2 On values.

Each character will have a number of values and beliefs initialized upon their creation. Similar to character personality traits, these values range from one to a hundred. At the time of writing, our system contains ten values. These values are generally themed around the factions and family units. What a character initially values is determined by a multitude of factors like their initial faction, their archetypes, traits, and relationships. For instance, a character that has a strong relationship with family members will care about *family*, whereas a lover archetype will care about *love* as a

value. We note that what a character cares about can change as a result of their interactions with other characters or achieving character goals.

6.2.3 On goals.

Goals are written in a way that is easily authorable. Each goal is composed of the following:

- Goal name and ID: an identifier to track and check the goal's progress or status.
- Character preconditions. These conditions must be true for a character to complete the goal. Conditions can vary depending on the authored goal. We authored some goals that rely on the character's successful or attempted actions.
- World preconditions. These are conditions based on world events that must be true to achieve a goal. We note that this condition is optional.
- Effects. Effects happen as a result of achieving a goal. These effects are personal and belong to the characters involved in achieving the goal. At the time of this writing, effects can change a character's renown, values, and personality traits.
- World effects. Similar to effects, these effects are global and alter the world in some way. For instance, a character succeeding in killing two enemies affects their traits, such as confidence and renown, but also alters their faction's strength and standing within the world. Similar to world preconditions, these effects are optional.

- Priority. A numerical value can be assigned to each goal. We currently range priorities from one to three, with three being the highest priority.
- Goal type. Finally, goal types can be personal or faction-based at the time of writing. Personal goals can be thought of as goals a character may want to achieve, while faction goals are assigned for the betterment of a faction. Faction goals are also longer and take more time to fulfill, as hinted at in the earlier scenario.

We currently have the groundwork set up for our goals. These goals are currently implemented but not attached as part of the simulation. Further testing is required before we fully integrate this component.

Example authored goal:

```
Goals goal_X = new(v1: 5, "Recruit multiple entities",
    c => c.knowledge.ReturnTimesSucseededAtAnAction("Recruit") > 2, //precondition
    w => true, //default world condition
    c =>
    {
        c.ApplyGoalEffects(1, "confidence");                //effects
        c.ApplyGoalEffects(2);                             //renown
    }, 3,                                                 //priority
    w => { w.IncareaseInitFactionSTR(2); },              //faction effects
    Goals.GoalType.faction);                            //goal type
```

6.2.4 Relationships.

Each character has a reference to the characters they met in the world. We keep track of three relationship stats, the main types of which are relationship type, relationship kind, and relationship value. We can think of relationship types as thematic descriptions, like romantic, familial, or friendships. In contrast, relationship kind signifies the specifics of a relationship, such as identifying another NPC as an enemy,

partner, son, or friend. We believe that having two metrics creates interesting dynamics for our characters. For instance, we can have scorned lovers if they see each other as enemies but share a romantic relationship type. Finally, a relationship's value is the numerical number attached to the character's feelings. Like other metrics in our project, relationship values range from one to a hundred.

By default, all characters have a reference to other characters as strangers (relationship type) and an unknown (relationship kind). If a character is instantiated within the same faction, the relationship kind will likely be that of a faction member. Characters within the same faction will likely start off being friends or acquaintances, although there is a 30 percent chance of creating enemies within the same faction. Other non-faction members are most likely set up as strangers. We note relationships also amount to numerical values that range from one to one hundred. Different events and encounters can affect a character's relationship. Relationship effects are nonreciprocal, meaning that if one character liked someone from an interaction, the other may not feel the same way. Character relationships are temporary; we are currently restructuring them in another format to allow for more variety in our relationships.

6.2.5 Personal knowledge base.

Each character has a reference to their internal knowledge base, which lists a reference to their actions or information about other characters, such as what they think a character values or likes. (The concept of lying is underdeveloped at the moment; more information under social actions.)

A character's personal knowledge base is used in conjunction with goals, events, and character actions, to increase the odds of a successful action or goal. For instance, a character may reference the number of times they attempted a social action in their utility function or in achieving a goal that asks them to be braver (e.g., attempting to attack an enemy a few times).

6.3 Events and Actions

Currently, the game runs via events and scenarios. Before we address events, one of our primary systems, we would like to briefly discuss scenarios. Scenarios act as physical components that accompany events. Scenarios (and a scenario manager) mark the physical locations and coordinates where the event(s) takes place. It essentially tells characters to move to specific locations and partake in an event. We can imagine the scenario as a stage, whereas each event is a scene, and a collection of events in a scenario represents an act for a character.

We are currently in the testing phase of our project and thus opted for a text-based UI whereby we can select particular events and view the results of each interaction for different characters. We hope to eventually flesh out our world and run the game by invoking multiple events simultaneously. We will use scenarios as a main gameplay component, by which players can view different scenarios around the game-world. However, we note that gameplay is still under development at the time of writing and will serve as future work. Events, as mentioned earlier, act as scenes in a play. Like goals, events are written with authoring in mind. Authoring events involves:

- **Event name and ID.** These are used to reference a particular event and track its status in the game.
- **Event type.** We currently structure three types of events.
    - **Farming events:** As mentioned earlier, resources are needed in this game and act as a source of conflict. Thus, farming these resources is structured as events that require some characters to farm. This event and its gameplay (mining, gardening, hunting) are under development and require the character's location and gameplay, serving as part of our future work.
    - **Social exchange events.** Social exchange events invoke characters that meet certain conditions to participate in social actions. The character's actions are determined by their utility functions.
    - **Large faction events**. As the name suggests, these events involve all faction members. Instances include participating in raid scenarios or defending their base. This event type is currently under development, as it mainly serves our gameplay portion.
- **Event Preconditions**. Each event has a precondition that involves one or more characters. These conditions are required for an event to take place. For instance, we may author an event in which an injured character runs across an enemy faction. In this case, the conditions would be that one character has to have low health, with characters A and B belonging to different factions.

Selecting characters for events is currently done through a UI for testing purposes.

- **Event Effects.** Effects, like goal effects, can alter the situation with additional modifications dependent on the author's scenario. For instance, if a faction invasion event comes to pass, the faction loses strength due to stolen resources. We note that events can affect characters as well; these effects are optional and are added depending on the authoring scenario, as each social action and goal affects characters in different ways.

Events currently work with a UI through which we can select eligible characters to enact in the event. We note that events also highlight the initiator and receiver; we will elaborate on the UI in a later section. Upon selecting a social-based event (such as characters coming across each other), the initiator and receiver characters enact a few rounds of social exchange. What a character chooses depends on their utility functions. Each character has a different utility for different actions due to their archetypes, values, or relationships. We currently created these actions:

- **Common social actions:** We currently have *insult*, *small talk*, *greet*, and *flirt* as possible actions that invoke their namesakes. They are typical social actions found in simulation-based games or RPGS. Performing these actions can affect a character's relationship or personality traits.
- **Recruitment-based actions.** *Inquiring*, *recruiting*, and *convincing* are faction-related social actions. -*Inquire about*- is currently under development. Through inquiring, an NPC may learn more about a character's values and add to their

own knowledge. In the future, we hope to use a character's knowledge base to create a common ground or foundation through which we can highlight change.

- **Hostile actions**. *Attacking* is a possible action, depending on the event or characters involved. Those with a highly *violent* trait or outlaw archetypes may attack first and ask questions later. Attacking leads to a loss of health.
- **Fleeing action**. Fleeing is a possible action or reaction a character can take due to their interaction or personality. Fleeing ends an encounter early.

The initiator selects one of the actions mentioned above based on a utility function, and the receiver reacts by selecting an applicable reaction based on a utility from possible reactions. For instance, if character A insults character B, character B is limited to a discrete set of reactions whereby an insult has potential reactions including *insult, beg, look away, flee,* or *attack*. We note that each action has a defined set of possible reactions tailored to work with the selected action. A utility function is calculated for both actions and reactions based on a multitude of factors, including how many times an action was attempted, the initiator and receiver relationship values, general values, and their personality traits or archetypes. We are updating, implementing, and testing these utility functions; they are currently a work in progress, especially as we finetune and balance them. The following section will provide an example scenario and explain our current simulation and future goals.

6.4 Current Demo and UI Configurations

While we made some progress toward creating a game environment and animations, our primary focus is concentrating on the system's core loop, which involves events and character social actions. To test events and actions, we created a small UI to select an event from our list of authored events. The UI then filters characters based on the selected event. We can also reference character information (values, relationships, traits) and run that event, reading what transpires for three rounds. Figure 44 depicts our current UI configurations, while Figure 45 showcases the dialogue effects of running an event that forces two opposing faction members to meet.

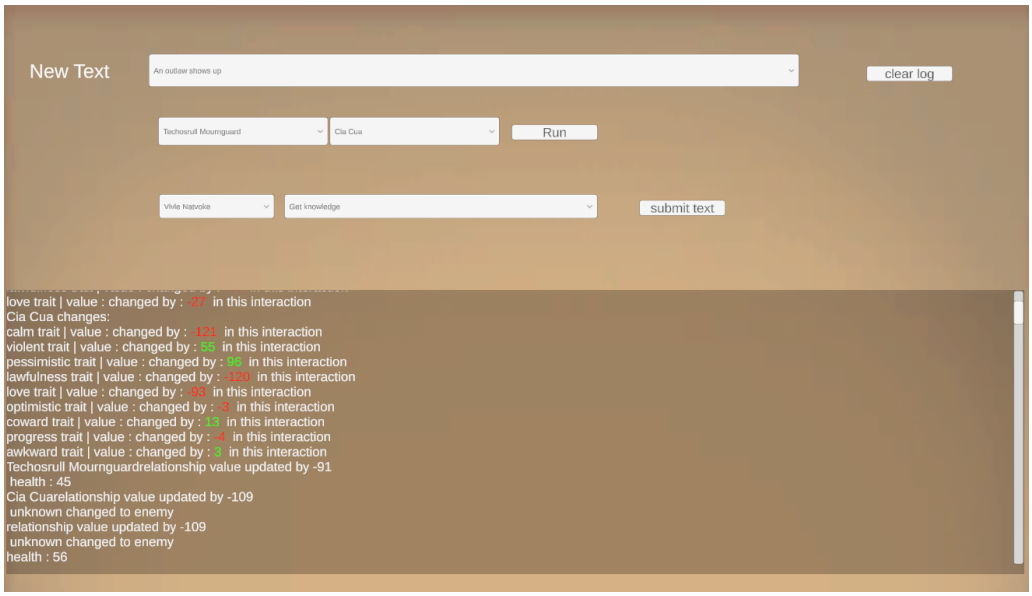

Figure 44 depicts the UI system used to test events.

The depicted scenario shows how the interaction started with an insult from the initiator. Upon getting insulted, the receiver insulted the initiator back. We can see the situation escalating after the initiator attacked the receiver, to which the receiver begged

(as a reaction). Following this, the initiator and receiver insulted each other and went on their separate ways. After a three-round exchange (each social exchange has three rounds of actions and reactions unless an NPC flees or exits an exchange), we can see a change in traits and values where applicable. The above scenario showed that the receiver's calmness and lawfulness traits decreased, whereas the pessimism trait increased.

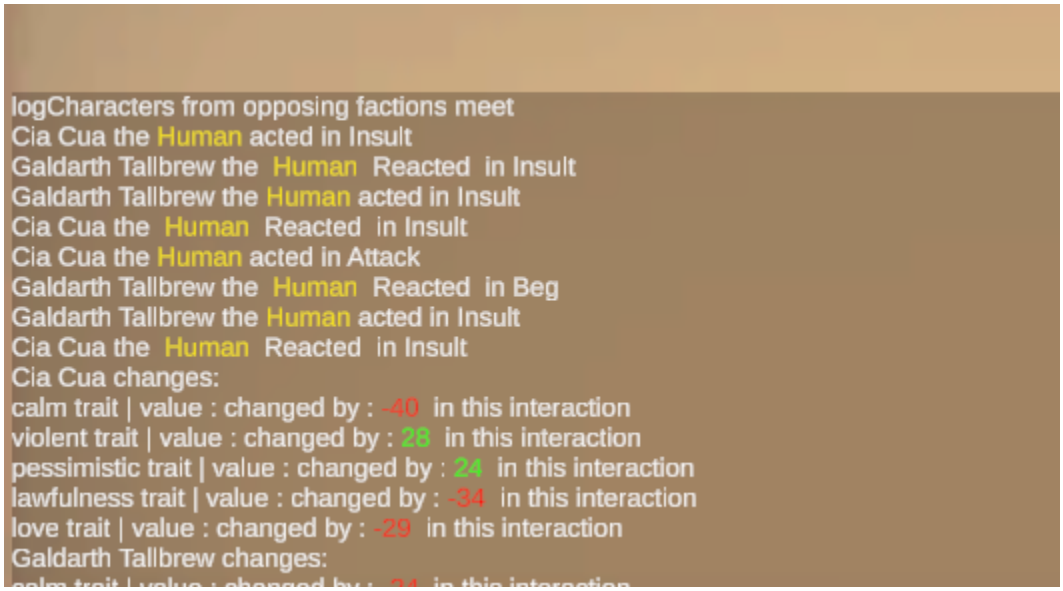

Figure 45 depicts the log results from running the event titled *two characters from opposing factions meet.*

6.5 Conclusion, Game Design and Future Work

After balancing the core loop, updating some structural elements (e.g., relationships), and incorporating others such as goals, we will focus on reiterating over our game design concepts.

We envision a Sims-like game in which characters converse using animations and UI pop-ups. Players will view the world from an outward perspective in a sandbox-like game, zooming in on specific characters to learn more about them. Figure 46 depicts future directions.

We also hope to expand our set of actions to incorporate additional subactions, such as actions serving as in-betweens (taken from animation concepts, in-betweens add a possible subaction as a transition point between the character's main actions and reactions), expanded actions (actions that are considered a higher form of the base action, for instance, flirt eventually unlocks serenade) and additional social actions found in similar games (e.g., sabotage, gossip, and hug).

We also plan on incorporating time-based mechanics since our system aims to highlight character change and growth as a concept. Similar to other simulation-based games(e.g., Sims and Two-point Hospital), players can control the flow of the game, including speeding up the game, pausing it, and slowing it down. We also plan on expanding our scenario components so that players can cycle through active scenarios or read the summary of transpired events.

While we know that considerable development and time are needed to bring this project to fruition, we believe that a demo like this can serve and aid us in understanding how to develop a system to support character change, an important and often missing aspect of character believability. We believe that a faction-based game offers an excellent venue for conveying small changes to background characters that can make them seem more believable on a larger scale.

Figure 46: the figure depicts the two factions' zones and character interactions via UI pop-ups, currently under development.

Conclusion

In this dissertation, we explored three avenues of research concerning video game characters, including the character's believability, sense of beliefs, and morality.

One of our initial goals in our work is to explore what believability entails for NPCs within video games. Chapters One and Two explored notions of believability.

In **Chapter One**, we discussed how believability is often ambiguous; researchers frequently confuse a believable character with a realistic one. Our research uncovered too many variations between the two character accounts for them to be synonymous, coming to two differing terms: realistic characters and believable characters.

We then dived deeper into believability. We argued that character believability is based on the player's perception and results as a combination of criteria that a character holds; these include a character's personality, motivation, emotion, roles, sociability, and change, among others. We argued that by analyzing these criteria, one could determine and understand how a character is perceived to be believable by players. Furthermore, we argued that evaluating believability is often based on the player's perception and, therefore, requires qualitative measures and questioning to understand how believability is conveyed to our players.

We then investigated a series of games and academic systems that highlight believability notions, coming to the conclusion that character change is often an underdeveloped and understudied area of believability.

In examining autonomous and believable systems, we also noticed that one common element that emerged is belief modeling. We argue that a character's values and beliefs can dramatically affect believability criteria, whereby we can see clear relationships between a character's values and motivation, on the one hand, and perceived personality, on the other.

In examining social and believability-based systems, we found that most implemented values are presented as part of an NPC's social aspects (with shallow belief modeling) or as the system primarily focuses without contextualizing it within the scope of believable character design. We believe that a gap exists in employing values and beliefs to a character's actions, especially as it pertains to a character's sense of morality and believability.

Thus, **Chapter Three** defines morality within our dissertation. We also highlight standard implementation methods for morality within popular games, culminating in two main designs: reputation scales and state machines. Through our analysis, we argue that standard reputation scales tend to shift moral systems into a binary state of good or evil, whereby morality often includes many shades of gray. While heavily scripted characters are perceived to have diverse morality, background characters usually lack believability. We argue that utilizing values and beliefs delivers

a promising method of diverse character morality, creating a sense of uniqueness in characters, especially background characters.

In **Chapter Four**, we dive deeper and review beliefs and values as a topic. We also propose using known theories to implement diverse morality via an NPC's values and beliefs. Thus, we introduced Lakoff's *Moral Politics,* a political book that showcases morality as a collection of metaphorical values used as a basis for our projects.

We built two primary iterations of Argument Box, a game built to learn more about value-based design. Argument Box went through multiple studies, including pilot studies, to ensure that the design functions as intended. Argument Box included two primary studies, the first of which asked if players can perceive values within our system. Through the first study, we learned that players successfully identified values in numerous ways, establishing the baseline for our second study.

Through our second study, we identified that a relationship exists between an NPC's values and that of believability and morality. We discovered multiple avenues in which values play a role in NPC believability and morality. Specifically, there is a strong association between values and a character's personality. We also learned that a character's values impacted their perceived motivation, change, and the overall illusion of life. We also learned that values affected how players perceived NPC morality; some players described morality according to broader definitions, while others noted specific definitions as moral behavior. Some players were also able to perceive morality as being multifaceted, adding a sense of uniqueness to characters. Though additional

research is needed to verify the results, we believe nonetheless that our work shows promising results in terms of linking morality, values, and notions of believability.

Lastly, in exploring morality, we came to the notion that most moral systems contain faction designs by nature, presenting morality as an often two-sided stance, one with a faction and one against said faction. Thus, our concluding chapter reviews morality through a faction lens. In doing so, we created a (game) faction taxonomy highlighting how morality and systems operate.

In developing our taxonomy, we concluded that game factions differ from other factions perceived in various media. We analyzed game-based and other media factions, coming to the realization that a key discrepancy exists: character growth (change).

As a reminder, character change was determined as an understudied area in character believability. Through analyzing faction designs, we concluded numerous ways in which change functions, including whom it affects, how it's processed and triggered, what motivates change, and how it relates to character beliefs and values.

Current and Future work

As presented in **Chapter Nine**, we are currently in the process of creating a faction-based game, one that highlights change and values as an aspect of believability. In future work, we hope to develop this system further and learn more about change and its effects on character believability, values, beliefs, and morality.

Bibliography

[1]

2k and Irrational Games, *BioShock Infinite*. (2013).

[2]

R. P. Abelson, “Differences Between Belief and Knowledge Systems,” *Cognitive Science*, vol. 3, no. 4, pp. 355–366, 1979, doi: 10.1207/s15516709cog0304_4.

[3]

R. P. Abelson and J. D. Carroll, “Computer Simulation of Individual Belief Systems,” *American Behavioral Scientist*, vol. 8, no. 9, pp. 24–30, May 1965, doi: 10.1177/000276426500800908.

[4]

R. AlJammaz, M. Mateas, and N. Wardrip-Fruin, “Modeling Morality-Based Argumentation for Believable Game Characters: A Design Postmortem,” *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 19, no. 1, Art. no. 1, Oct. 2023, doi: 10.1609/aiide.v19i1.27514.

[5]

R. AlJammaz, Y. She, and M. Mateas, “Argument Box,” in *AIIDE Workshops*, 2021. Accessed: Dec. 06, 2024. [Online]. Available: https://scholar.google.com/scholar?cluster=12051650494555301234&hl=en&oi=scholarr

[6]

R. Aljammaz, N. Wardrip-Fruin, and M. Mateas, “Towards an Understanding of Character Believability,” in *Proceedings of the 18th International Conference on the Foundations of Digital Games*, in FDG ’23. New York, NY, USA: Association for Computing Machinery, Apr. 2023, pp. 1–9. doi: 10.1145/3582437.3582466.

[7]

R. AlJammaz, N. Wardrip-Fruin, and M. Mateas, “Navigating Faction Systems: Insights and Recommendations for More Believable NPCs in Video Games,” in *Proceedings of the 19th International Conference on the Foundations of Digital Games*, in FDG ’24. New York, NY, USA: Association for Computing Machinery, Jul. 2024, pp. 1–11. doi: 10.1145/3649921.3650012.

[8] Amplitude Studios, *Humankind*. (2021).

[9] C. Arzate Cruz and J. A. Ramirez Uresti, “HRLB^2: A Reinforcement Learning Based Framework for Believable Bots,” *Applied Sciences*, vol. 8, no. 12, Art. no. 12, Dec. 2018, doi: 10.3390/app8122453.

[10] R. Aylett and S. Louchart, “Towards a narrative theory of virtual reality,” *Virtual Reality*, vol. 7, no. 1, pp. 2–9, Dec. 2003, doi: 10.1007/s10055-003-0114-9.

[11] R. S. Aylett, S. Louchart, J. Dias, A. Paiva, and M. Vala, “FearNot! – An Experiment in Emergent Narrative,” in *Intelligent Virtual Agents*, T. Panayiotopoulos, J. Gratch, R. Aylett, D. Ballin, P. Olivier, and T. Rist, Eds., Berlin, Heidelberg: Springer, 2005, pp. 305–316. doi: 10.1007/11550617_26.

[12] S. Azad and C. Martens, “Addressing the Elephant in the Room: Opinionated Virtual Characters,” in *Experimental AI in Games Workshop at the 14th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE’18)*, Edmonton, Canada, 2018. Accessed: Dec. 06, 2024. [Online]. Available: https://www.researchgate.net/publication/331024627_Addressing_the_Elephant_in_the_Room_Opinionated_Virtual_Characters

[13] S. Azad and C. Martens, “Lyra: Simulating Believable Opinionated Virtual Characters,” *Proceedings of the AAAI Conference on Artificial*

Intelligence and Interactive Digital Entertainment, vol. 15, no. 1, Art. no. 1, Oct. 2019, doi: 10.1609/aiide.v15i1.5232.

[14] J. Bates, "The role of emotion in believable agents," *Commun. ACM*, vol. 37, no. 7, pp. 122–125, Jul. 1994, doi: 10.1145/176789.176803.

[15] Bay 12 Games, *Dwarf Fortress*. (2022).

[16] Bethesda Game Studios, *The Elder Scrolls IV: Oblivion*. (2006).

[17] Bethesda Game Studios, *The Elder Scrolls V: Skyrim*. (2011).

[18] Bethesda Game Studios, *Fallout 4*. (2015).

[19] E. Bevacqua, R. Richard, and P. De Loor, "Believability and Co-presence in Human-Virtual Character Interaction," *IEEE Computer Graphics and Applications*, vol. 37, no. 4, pp. 17–29, 2017, doi: 10.1109/MCG.2017.3271470.

[20] BioWare, *Neverwinter Nights*. (2002).

[21] BioWare, *Dragon Age: Origins*. (2009).

[22] S. Bjork and J. Holopainen, *Patterns In Game Design*, 1st edition. Boston, Mass: Charles River Media, 2004.

[23] Blizzard Entertainment, *World of Warcraft*. (2004).

[24] B. Bucklew, J. Grinblat, Caelyn Sandel, and N. DeCapua, *Caves of Qud*. (2015).

[25] Bungie and Halo Studios, *Halo*. (2021 2001).

[26]

J. Cameron, *Avatar*, (2009).

[27] J. Cassas-Roma, M. Nelson, J. Arnedo-Moreno, S. Gaudl, and R. Saunders, “Towards Morally Autonomous NPCs: The Case of The Elder Scrolls IV: Oblivion,” presented at the Artificial Intelligence and Simulated Behaviour, Falmouth, Apr. 2019. Accessed: Dec. 06, 2024. [Online]. Available: https://repository.falmouth.ac.uk/3330/

[28] J. Cassell *et al.*, “Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents,” in *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, in SIGGRAPH ’94. New York, NY, USA: Association for Computing Machinery, Jul. 1994, pp. 413–420. doi: 10.1145/192161.192272.

[29] G. Castellano *et al.*, “Towards Empathic Virtual and Robotic Tutors,” in *Artificial Intelligence in Education*, H. C. Lane, K. Yacef, J. Mostow, and P. Pavlik, Eds., Berlin, Heidelberg: Springer, 2013, pp. 733–736. doi: 10.1007/978-3-642-39112-5_100.

[30] M. Cavazza, F. Charles, and S. J. Mead, “Character-based interactive storytelling,” *IEEE Intelligent Systems*, vol. 17, no. 4, pp. 17–24, Jul. 2002, doi: 10.1109/MIS.2002.1024747.

[31] M. Cavazza, J.-L. Lugrin, D. Pizzi, and F. Charles, “Madame bovary on the holodeck: immersive interactive storytelling,” in *Proceedings of the 15th ACM international conference on Multimedia*, in MM ’07. New York, NY, USA: Association for Computing Machinery, Sep. 2007, pp. 651–660. doi: 10.1145/1291233.1291387.

[32] J.-A. Cervantes, L.-F. Rodríguez, S. López, and F. Ramos, “A biologically inspired computational model of Moral Decision Making for autonomous agents,” in *2013 IEEE 12th International Conference on*

Cognitive Informatics and Cognitive Computing, Jul. 2013, pp. 111–117. doi: 10.1109/ICCI-CC.2013.6622232.

[33] K. Charmaz, *Constructing Grounded Theory*, Second edition. London ; Thousand Oaks, Calif: SAGE Publications Ltd, 2014.

[34] J. Cleese and G. Chapman, "Argument Clinic," *Monty Python's Flying Circus*, BBC, 1972.

[35] N. Cointe, G. Bonnet, and O. Boissier, "Ethical Judgment of Agents' Behaviors in Multi-Agent Systems," in *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, in AAMAS '16. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, May 2016, pp. 1106–1114.

[36] ConcernedApe, *Stardew Valley*. (2016).

[37] F. Darabont, "The Walking Dead, Season 6," AMC, 2016 2015.

[38] F. Darabont, "The Walking Dead, Season 9," AMC, 2019 2018.

[39] B. De Carolis, C. Pelachaud, I. Poggi, and M. Steedman, "APML, a Markup Language for Believable Behavior Generation," in *Life-Like Characters: Tools, Affective Functions, and Applications*, H. Prendinger and M. Ishizuka, Eds., Berlin, Heidelberg: Springer, 2004, pp. 65–85. doi: 10.1007/978-3-662-08373-4_4.

[40] D. M. Degens, G. J. Hofstede, J. McBreen, S. Mascarenhas, A. Paiva, and A. J. M. Beulens, "When agents meet: empathy, moral circle, ritual, and culture," presented at the Workshop on Emotional and Empathic Agents help on AAMAS 2012, Valencia, Spain, 2012. Accessed: Dec. 09, 2024. [Online]. Available: https://research.wur.nl/en/publications/when-agents-meet-empathy-moral-circle-ritual-and-culture

[41]
V. Demeure, R. Niewiadomski, and C. Pelachaud, “How Is Believability of a Virtual Agent Related to Warmth, Competence, Personification, and Embodiment?,” *Presence*, vol. 20, no. 5, pp. 431–448, Oct. 2011, doi: 10.1162/PRES_a_00065.

[42]
J. Dias and A. Paiva, “Feeling and reasoning: a computational model for emotional characters,” in *Proceedings of the 12th Portuguese conference on Progress in Artificial Intelligence*, in EPIA’05. Berlin, Heidelberg: Springer-Verlag, Dec. 2005, pp. 127–140. doi: 10.1007/11595014_13.

[43]
P. H. Ditto, D. A. Pizarro, and D. Tannenbaum, “Motivated Moral Reasoning,” *Psychology of Learning and Motivation*, vol. 50, pp. 307–338, 2009, doi: 10.1016/S0079-7421(08)00410-6.

[44]
Dontnod Entertainment, *Life Is Strange*. (2015).

[45]
Dontnod Entertainment, *Vampyr*. (2018).

[46]
K. Dorer, “Extended Behavior Networks for the magmaFreiburg Team,” *Linköping Electronic Articles in Computer and Information Science*, vol. 4, no. 17, pp. 79–83, Dec. 1999.

[47]
L. Egri, *The Art Of Dramatic Writing: Its Basis in the Creative Interpretation of Human Motives*, Revised edition. New York, NY (u.a.): Touchstone, 1972.

[48]
M. S. El-Nasr, L. Bishko, V. Zammitto, M. Nixon, A. V. Vasiliakos, and H. Wei, “Believable Characters,” in *Handbook of Multimedia for Digital Entertainment and Arts*, B. Furht, Ed., Boston, MA: Springer US, 2009, pp. 497–528. doi: 10.1007/978-0-387-89024-1_22.

[49]

S. ElSayed and D. J. King, “Affect and believability in game characters: GAME-ON’2017, 18th annual Conference on Simulation and AI in Computer Games,” *Game-On’17*, pp. 90–97, Aug. 2017.

[50] Ensemble Studios, *Age of Empires*. (1997).

[51] R. Evans and E. Short, “Versu—A Simulationist Storytelling System,” *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 6, no. 2, pp. 113–130, Jun. 2014, doi: 10.1109/TCIAIG.2013.2287297.

[52] Firaxis Games and Westlake Interactive, *Civilization III*. (2001).

[53] L. Floridi and J. W. Sanders, “On the Morality of Artificial Agents,” *Minds and Machines*, vol. 14, no. 3, pp. 349–379, Aug. 2004, doi: 10.1023/B:MIND.0000035461.63578.9d.

[54] S. Frazier, M. S. A. Nahian, M. Riedl, and B. Harrison, “Learning Norms from Stories: A Prior for Value Aligned Agents,” Dec. 07, 2019, *arXiv*: arXiv:1912.03553. doi: 10.48550/arXiv.1912.03553.

[55] Funcom, *Conan Exiles*. (2018).

[56] P. Gamez, D. B. Shank, C. Arnold, and M. North, “Artificial virtue: the machine question and perceptions of moral character in artificial moral agents,” *AI & Soc*, vol. 35, no. 4, pp. 795–809, Dec. 2020, doi: 10.1007/s00146-020-00977-1.

[57] P. Gebhard, “ALMA: a layered model of affect,” in *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, in AAMAS ’05. New York, NY, USA: Association for Computing Machinery, Jul. 2005, pp. 29–36. doi: 10.1145/1082473.1082478.

[58]

L. R. Goldberg, “An alternative ‘description of personality’: the big-five factor structure,” *J Pers Soc Psychol*, vol. 59, no. 6, pp. 1216–1229, Dec. 1990, doi: 10.1037//0022-3514.59.6.1216.

[59] P. Gomes, A. Paiva, C. Martinho, and A. Jhala, “Metrics for Character Believability in Interactive Narrative,” in *Interactive Storytelling*, H. Koenitz, T. I. Sezen, G. Ferri, M. Haahr, D. Sezen, and G. Çatak, Eds., Cham: Springer International Publishing, 2013, pp. 223–228. doi: 10.1007/978-3-319-02756-2_27.

[60] M. Grinberg and E. Todorov, “Cognitive Agent Based Simulation Platform for Modeling Large-Scale Multi-Level Social Interactions with Experimental Games,” *International Journal*, vol. 3, no. 2, 2016.

[61] M. Guimaraes, P. Santos, and A. Jhala, “CiF-CK: An architecture for social NPCS in commercial games,” in *2017 IEEE Conference on Computational Intelligence and Games (CIG)*, Aug. 2017, pp. 126–133. doi: 10.1109/CIG.2017.8080425.

[62] J. Gunn, *Guardians of the Galaxy Vol. 2*, (2017).

[63] D. F. Harrell, D. Kao, C.-U. Lim, J. Lipshin, and A. Sutherland, *Chimeria: Gatekeeper*. (2014).

[64] D. F. Harrell, P. Ortiz, P. Downs, M. Wagoner, E. Carré, and A. Wang, “Chimeria:Grayscale: an interactive narrative for provoking critical reflection on gender discrimination,” *Coimbra University Press*, 2018, Accessed: Dec. 06, 2024. [Online]. Available: https://dspace.mit.edu/handle/1721.1/135084.2

[65] F. Heider and M. Simmel, “An Experimental Study of Apparent Behavior,” *The American Journal of Psychology*, vol. 57, no. 2, pp. 243–259, 1944, doi: 10.2307/1416950.

[66] I. Horswill, “MKULTRA (Demo),” *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 11, no. 1, Art. no. 1, 2015, doi: 10.1609/aiide.v11i1.12776.

[67] B. Howard and R. Moore, *Zootopia*, (2016).

[68] P. Itagaki, *BEASTARS*. Akita Shoten, 2019.

[69] S. Johnson-Bey, M. J. Nelson, and M. Mateas, “Neighborly: A Sandbox for Simulation-based Emergent Narrative,” in *2022 IEEE Conference on Games (CoG)*, Aug. 2022, pp. 425–432. doi: 10.1109/CoG51982.2022.9893631.

[70] O. Johnston and F. Thomas, *The Illusion of Life: Disney Animation*, Subsequent edition. New York, NY: Disney Editions, 1995.

[71] M. Karpenko, *WorldBox*. (2012).

[72] M. Karpenko, *WorldBox*. (2019).

[73] M. Kreminski, M. Dickinson, and M. Mateas, “Winnow: A Domain-Specific Language for Incremental Story Sifting,” *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 17, no. 1, Art. no. 1, Oct. 2021, doi: 10.1609/aiide.v17i1.18903.

[74] G. Lakoff, *The ALL NEW Don’t Think of an Elephant!: Know Your Values and Frame the Debate*. Chelsea Green Publishing, 2014.

[75] G. Lakoff, “Chapter 5, Strict Father Morality,” in *Moral Politics: How Liberals and Conservatives Think, Third Edition*, 3rd Enlarged ed. edition., Chicago: University of Chicago Press, 2016, pp. 65–107.

[76]

G. Lakoff, “Chapter 6, Nurturant Parent Morality,” in *Moral Politics: How Liberals and Conservatives Think, Third Edition*, 3rd Enlarged ed. edition., Chicago: University of Chicago Press, 2016, pp. 108–140.

[77] G. Lakoff, *Moral Politics: How Liberals and Conservatives Think, Third Edition*, 3rd Enlarged ed. edition. Chicago: University of Chicago Press, 2016.

[78] G. Lakoff, H. Dean, and D. Hazen, “Framing 101: How to Take Back Public Discourse,” in *Don’t Think of an Elephant!: Know Your Values and Frame the Debate--The Essential Guide for Progressives*, First Edition., White River Junction, Vt: Chelsea Green Publishing, 2004, pp. 1–21.

[79] P. Lankoski and S. Björk, “Gameplay Design Patterns for Believable Non-Player Characters,” in *Proceedings of DiGRA 2007 Conference: Situated Play*, Jan. 2007. Accessed: Dec. 06, 2024. [Online]. Available: https://dl.digra.org/index.php/dl/article/view/262

[80] P. Lankoski and S. Björk, Eds., *Game Research Methods: An Overview*. Pittsburgh: lulu.com, 2015.

[81] P. Lankoski, A. Johansson, B. Karlsson, S. Björk, and P. Dell’Acqua, “AI Design for Believable Characters via Gameplay Design Patterns,” in *Business, Technological, and Social Dimensions of Computer Games: Multidisciplinary Developments*, IGI Global Scientific Publishing, 2011, pp. 15–31. doi: 10.4018/978-1-60960-567-4.ch002.

[82] P. Leong and M. Chunyan, “Fuzzy cognitive agents in shared virtual worlds,” in *2005 International Conference on Cyberworlds (CW’05)*, Nov. 2005, p. 5 pp. – 372. doi: 10.1109/CW.2005.49.

[83] J. C. Lester and B. A. Stone, “Increasing believability in animated pedagogical agents,” in *Proceedings of the first international conference on*

Autonomous agents - AGENTS '97, Marina del Rey, California, United States: ACM Press, 1997, pp. 16–21. doi: 10.1145/267658.269943.

[84] M. Y. Lim and R. Aylett, "Feel the Difference: A Guide with Attitude!," in *Proceedings of the 7th international conference on Intelligent Virtual Agents*, in IVA '07. Berlin, Heidelberg: Springer-Verlag, Sep. 2007, pp. 317–330. doi: 10.1007/978-3-540-74997-4_29.

[85] Lionhead Studios, *Fable II*. (2007).

[86] J. M. Llargues Asensio, J. Peralta, R. Arrabales, M. G. Bedia, P. Cortez, and A. L. Peña, "Artificial Intelligence approaches for the generation and assessment of believable human-like behaviour in virtual characters," *Expert Systems with Applications*, vol. 41, no. 16, pp. 7281–7290, Nov. 2014, doi: 10.1016/j.eswa.2014.05.004.

[87] A. B. Loyall, "Believable agents: building interactive personalities," phd, Carnegie Mellon University, USA, 1997.

[88] A. B. Loyall and J. Bates, "Hap A Reactive, Adaptive Architecture for Agents," *School of Computer Science, Carnegie Mellon University*, 1991, Accessed: Dec. 06, 2024. [Online]. Available: https://www.semanticscholar.org/paper/Hap-A-Reactive%2C-Adaptive-Architecture-for-Agents-Bryan-Bates/f060869ea79e2d8279bb9ce5e20b29090c861022

[89] I. M. Mahmoud, L. Li, D. Wloka, and M. Z. Ali, "Believable NPCs in serious games: HTN planning approach based on visual perception," in *2014 IEEE Conference on Computational Intelligence and Games*, Aug. 2014, pp. 1–8. doi: 10.1109/CIG.2014.6932891.

[90] Mark R Johnson, *Ultima Ratio Regum*. (2012).

[91]

Marvelous Interactive, *Harvest Moon: A Wonderful Life*. (2003).

[92]

Massive Monster, *Cult of the Lamb*. (2022).

[93]

M. Mateas, "An Oz-Centric Review of Interactive Drama and Believable Agents," in *Artificial Intelligence Today: Recent Trends and Developments*, M. J. Wooldridge and M. Veloso, Eds., Berlin, Heidelberg: Springer, 1999, pp. 297–328. doi: 10.1007/3-540-48317-9_12.

[94]

M. Mateas and A. Stern, "Façade: An Experiment in Building a Fully-Realized Interactive Drama," presented at the Game Developers Conference, 2003.

[95]

Maxis Redwood Shores, *The Sims 3*. (2009).

[96]

J. McCoy, M. Treanor, B. Samuel, B. Tearse, M. Mateas, and N. Wardrip-Fruin, "Comme il Faut 2: a fully realized model for socially-oriented gameplay," in *Proceedings of the Intelligent Narrative Technologies III Workshop*, in INT3 '10. New York, NY, USA: Association for Computing Machinery, Jun. 2010, pp. 1–8. doi: 10.1145/1822309.1822319.

[97]

J. McCoy, M. Treanor, B. Samuel, A. A. Reed, N. Wardrip-Fruin, and M. Mateas, "Prom week," in *Proceedings of the International Conference on the Foundations of Digital Games*, in FDG '12. New York, NY, USA: Association for Computing Machinery, May 2012, pp. 235–237. doi: 10.1145/2282338.2282384.

[98]

J. McCoy, M. Treanor, B. Samuel, N. Wardrip-Fruin, and M. Mateas, "Comme il Faut: A System for Authoring Playable Social Models," *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 7, no. 1, Art. no. 1, Oct. 2011, doi: 10.1609/aiide.v7i1.12454.

[99]

M. McFarland, “Magic Leap’s new AI assistant looks alarmingly human | CNN Business,” CNN. Accessed: Dec. 06, 2024. [Online]. Available: https://www.cnn.com/2018/10/12/tech/magic-leap-ai-assistant/index.html

[100] R. McKee, *Story: Substance, Structure, Style and the Principles of Screenwriting*, 1st edition. New York, NY: ReganBooks, 1997.

[101] R. McKee, *Character: The Art of Role and Cast Design for Page, Stage, and Screen*. Grand Central Publishing, 2021.

[102] Merrit Kopas, *Lim*. (2012).

[103] Miquela, “Miquela (@lilmiquela) • Instagram photos and videos.” Accessed: Dec. 06, 2024. [Online]. Available: https://www.instagram.com/lilmiquela/

[104] J. Mooney and J. M. Allbeck, “Rethinking NPC intelligence: a new reputation system,” in *Proceedings of the 7th International Conference on Motion in Games*, in MIG ’14. New York, NY, USA: Association for Computing Machinery, Oct. 2018, pp. 55–60. doi: 10.1145/2668084.2668091.

[105] L. Morais, J. Dias, and P. A. Santos, “From caveman to gentleman: a CiF-based social interaction model applied to conan exiles,” in *Proceedings of the 14th International Conference on the Foundations of Digital Games*, in FDG ’19. New York, NY, USA: Association for Computing Machinery, Aug. 2019, pp. 1–11. doi: 10.1145/3337722.3337746.

[106] J. H. Murray, *Hamlet on the Holodeck, updated edition: The Future of Narrative in Cyberspace*, Updated ed. edition. Cambridge, Massachusetts: The MIT Press, 2017.

[107]

Naughty Dog, *The Last of Us*. (2013).

[108]

M. Nelson, “Prototyping Kant-Inspired Reflexive Game Mechanics,” May 29, 2012, *Social Science Research Network, Rochester, NY*: 2115479. Accessed: Dec. 06, 2024. [Online]. Available: https://papers.ssrn.com/abstract=2115479

[109]

M. J. Nelson, “Moral Calculus in Video Games.” Accessed: Dec. 06, 2024. [Online]. Available: https://www.kmjn.org/notes/moral_calculus_in_videogames.html

[110]

NetherRealm Studios, *Injustice 2*. (2017).

[111]

Nintendo EAD, Flagship, Grezzo, and Nintendo EPD, *The Legend of Zelda*. (2024 1986).

[112]

Nintendo EAD and Nintendo EPD, *Super Mario*. (2023 1985).

[113]

Nintendo EPD, *The Legend of Zelda: Tears of the Kingdom*. (2023).

[114]

Number None, *Braid*. (2008).

[115]

M. Oliveira and P. A. Santos, “A model for socially intelligent merchants,” in *Proceedings of the 14th International Conference on the Foundations of Digital Games*, in FDG ’19. New York, NY, USA: Association for Computing Machinery, Aug. 2019, pp. 1–8. doi: 10.1145/3337722.3337729.

[116]

OpenAI, *ChatGPT 3.5*. (2023).

[117]

A. Ortony, “On Making Believable Emotional Agents Believable,” in *Emotions in Humans and Artifacts*, R. Trappl, P. Petta, and S. Payr, Eds., The MIT Press, 2003, pp. 189–211. Accessed: Dec. 06, 2024. [Online]. Available:

https://direct.mit.edu/books/edited-volume/4405/chapter/188605/On-Making-Believable-Emotional-Agents-Believable

[118] A. Paiva *et al.*, "Caring for agents and agents that care: building empathic relations with synthetic agents," in *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems, 2004. AAMAS 2004.*, Jul. 2004, pp. 194–201. Accessed: Dec. 09, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/1373479

[119] A. Paiva *et al.*, "Learning by Feeling: Evoking Empathy with Synthetic Characters," *Applied Artificial Intelligence*, vol. 19, no. 3–4, pp. 235–266, Mar. 2005, doi: 10.1080/08839510590910165.

[120] Paradox Development Studio, *Crusader Kings III*. (2020).

[121] I. Poggi, C. Pelachaud, F. de Rosis, V. Carofiglio, and B. De Carolis, "Greta. A Believable Embodied Conversational Agent," in *Multimodal Intelligent Information Presentation*, O. Stock and M. Zancanaro, Eds., Dordrecht: Springer Netherlands, 2005, pp. 3–25. doi: 10.1007/1-4020-3051-7_1.

[122] Positech Games, *Redshirt*. (2013).

[123] D. V. Pynadath and S. C. Marsella, "PsychSim: Modeling Theory of Mind with Decision-Theoretic Agents".

[124] Quantic Dream, *Beyond: Two Souls*. (2013).

[125] M. O. Riedl and R. M. Young, "Narrative Planning: Balancing Plot and Character," *jair*, vol. 39, pp. 217–268, Sep. 2010, doi: 10.1613/jair.2989.

[126] Rockstar Games, *Red Dead Redemption 2*. (2018).

[127]

Rockstar North, *Grand Theft Auto IV*. (2008).

[128]

A. Russo and J. Russo, *Avengers: Infinity War*, (2018).

[129]

A. Russo and J. Russo, *Avengers: Endgame*, (2019).

[130]

J. Ryan, "Curating Simulated Storyworlds," UC Santa Cruz, 2018. Accessed: Dec. 06, 2024. [Online]. Available: https://escholarship.org/uc/item/1340j5h2

[131]

J. Ryan and M. Mateas, "Simulating Character Knowledge Phenomena in Talk of the Town," in *Game AI Pro 360: Guide to Character Behavior*, CRC Press, 2019.

[132]

J. Ryan, A. Summerville, M. Mateas, and N. Wardrip-Fruin, "Toward Characters Who Observe, Tell, Misremember, and Lie," *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 11, no. 3, Art. no. 3, 2015, doi: 10.1609/aiide.v11i3.12825.

[133]

B. Samuel, J. Ryan, A. J. Summerville, M. Mateas, and N. Wardrip-Fruin, "Bad News: An Experiment in Computationally Assisted Performance," in *Interactive Storytelling*, F. Nack and A. S. Gordon, Eds., Cham: Springer International Publishing, 2016, pp. 108–120. doi: 10.1007/978-3-319-48279-8_10.

[134]

J. Sanghvi, G. Castellano, I. Leite, A. Pereira, P. W. McOwan, and A. Paiva, "Automatic analysis of affective postures and body motion to detect engagement with a game companion," *Proceedings of the 6th international conference on Human-robot interaction*, pp. 305–312, Mar. 2011, doi: 10.1145/1957656.1957781.

[135]

M. Sellers, “Otello: A Next-Generation Reputation System For Humans and NPCs,” *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 4, no. 1, Art. no. 1, 2008, doi: 10.1609/aiide.v4i1.18688.

[136] J. Shaheed and J. Cunningham, “Agents making moral decisions,” 2008. Accessed: Dec. 09, 2024. [Online]. Available: https://www.semanticscholar.org/paper/Agents-making-moral-decisions-Shaheed-Cunningham/137089885c7427d42be808ad2c2cc92795d74da5

[137] L. J. Skitka, C. W. Bauman, and E. G. Sargis, “Moral Conviction: Another Contributor to Attitude Strength or Something More?,” *Journal of Personality and Social Psychology*, vol. 88, no. 6, pp. 895–917, 2005, doi: 10.1037/0022-3514.88.6.895.

[138] G. Smith and J. Carette, “Design Foundations for Emotional Game Characters,” *Eludamos: Journal for Computer Game Culture*, vol. 10, no. 1, Art. no. 1, 2019, doi: 10.7557/23.6175.

[139] K. Sterelny, “Dennett, D. C., ‘Brainstorms: Philosophical Essays on Mind and Psychology,’” *Australasian Journal of Philosophy*, vol. 59, no. n/a, p. 442, 1981.

[140] Sucker Punch Productions, *Infamous*. (2009).

[141] A. Summerville and B. Samuel, “Kismet: A Small Social Simulation Language,” in *Joint Proceedings of the ICCC 2020 Workshops*, 2020. Accessed: Dec. 06, 2024. [Online]. Available: https://www.semanticscholar.org/paper/Kismet%3A-A-Small-Social-Simulation-Language-Summerville-Samuel/4646f8631f1783f20fcc0af119b15a3cd58bec11

[142]

R. Sun, “The Motivational and Metacognitive Control in CLARION,” in *Integrated Models of Cognitive Systems*, W. D. Gray, Ed., Oxford University Press, USA, 2007, pp. 63–75.

[143] R. Sun, “The CLARION cognitive architecture: Toward a comprehensive theory of the mind,” in *The Oxford handbook of cognitive science*, New York, NY, US: Oxford University Press, 2017, pp. 117–133.

[144] Supermassive Games, *Until Dawn*. (2015).

[145] Telltale Games, *The Walking Dead: A Telltale Games Series*. (2019 2012).

[146] Telltale Games, *The Wolf Among Us*. (2013).

[147] Telltale Games, *Batman: The Telltale Series*. (2016).

[148] Toby Fox, *Undertale*. (2015).

[149] J. Togelius, “A Procedural Critique of Deontological Reasoning,” in *Proceedings of DiGRA 2011 Conference: Think Design Play*, Jan. 2011. Accessed: Dec. 06, 2024. [Online]. Available: https://dl.digra.org/index.php/dl/article/view/537

[150] J. Togelius, G. N. Yannakakis, S. Karakovskiy, and N. Shaker, “Assessing Believability,” in *Believable Bots: Can Computers Play Like People?*, P. Hingston, Ed., Berlin, Heidelberg: Springer, 2012, pp. 215–230. doi: 10.1007/978-3-642-32323-2_9.

[151] S. S. Tomkins and S. Messick, “Computer Simulation of ‘Hot Cognition,’” in *Computer Simulation of Personality*, First Edition., John Wiley and Sons, 1963, pp. 277–298.

[152]

M. Toy, G. Wichman, K. Arnold, and J. Lane, *Rogue*. (1980).

[153] A. M. Turing, "Computing Machinery and Intelligence," in *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*, R. Epstein, G. Roberts, and G. Beber, Eds., Dordrecht: Springer Netherlands, 2009, pp. 23–65. doi: 10.1007/978-1-4020-6710-5_3.

[154] Ubisoft Montréal, *Assassin's Creed II*. (2009).

[155] Ubisoft Montréal, *Assassin's Creed The Ezio Collection*. (2016).

[156] Ubisoft Montréal, *Far Cry Primal*. (2016).

[157] Valve, *Portal*. (2007).

[158] Virtual Beings Summit, "Home Page," Virtual Beings Summit. Accessed: Dec. 06, 2024. [Online]. Available: https://www.virtual-beings-summit.com

[159] S.-W. Wang, C.-Y. Huang, and C.-T. Sun, "Modeling self-perception agents in an opinion dynamics propagation society," *SIMULATION*, vol. 90, no. 3, pp. 238–248, Mar. 2014, doi: 10.1177/0037549713515029.

[160] L. Wanner *et al.*, "KRISTINA: A Knowledge-Based Virtual Conversation Agent," in *Advances in Practical Applications of Cyber-Physical Multi-Agent Systems: The PAAMS Collection*, Y. Demazeau, P. Davidsson, J. Bajo, and Z. Vale, Eds., Cham: Springer International Publishing, 2017, pp. 284–295. doi: 10.1007/978-3-319-59930-4_23.

[161] N. Wardrip-Fruin, *Expressive Processing: Digital Fictions, Computer Games, and Software Studies*. Cambridge, Mass. London: The MIT Press, 2012.

[162] N. Wardrip-Fruin, M. Mateas, S. Dow, and S. Sali, “Agency Reconsidered,” in *Proceedings of DiGRA 2009 Conference: Breaking New Ground: Innovation in Games, Play, Practice and Theory*, Jan. 2009. Accessed: Dec. 06, 2024. [Online]. Available: https://dl.digra.org/index.php/dl/article/view/369

[163] S. G. Ware and C. Siler, “The Sabre Narrative Planner: Multi-Agent Coordination with Intentions and Beliefs,” *AAMAS Conference proceedings*, May 2021, Accessed: Dec. 06, 2024. [Online]. Available: https://par.nsf.gov/biblio/10300907-sabre-narrative-planner-multi-agent-coordination-intentions-beliefs

[164] S. Ware and R. M. Young, “Glaive: A State-Space Narrative Planner Supporting Intentionality and Conflict,” *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 10, no. 1, Art. no. 1, 2014, doi: 10.1609/aiide.v10i1.12712.

[165] H. Warpefelt, “The Non-Player Character: Exploring the believability of NPC presentation and behavior,” Department of Computer and Systems Sciences, Stockholm University, 2016. Accessed: Dec. 06, 2024. [Online]. Available: https://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-128079

[166] H. Warpefelt and B. Stråat, “Breaking immersion by creating social unbelievabilty,” in *Proceedings of AISB 2013 Convention. Social Coordination: Principles, Artefacts and Theories (SOCIAL. PATH)*, 2013, pp. 92–100.

[167] B. Weber, M. Mateas, and A. Jhala, “Applying Goal-Driven Autonomy to StarCraft,” *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 6, no. 1, Art. no. 1, Oct. 2010, doi: 10.1609/aiide.v6i1.12401.

[168]

Yann Leroux, *Experimental study of apparent behavior. Fritz Heider & Marianne Simmel. 1944*, (Dec. 26, 2010). Accessed: Dec. 06, 2024. [Online Video]. Available: https://www.youtube.com/watch?v=n9TWwG4SFWQ

[169] G. N. Yannakakis, P. Spronck, D. Loiacono, and E. André, "Player Modeling," in *Artificial and Computational Intelligence in Games*, vol. 6, S. M. Lucas, M. Mateas, M. Preuss, P. Spronck, and J. Togelius, Eds., in Dagstuhl Follow-Ups, vol. 6. , Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2013, pp. 45–59. doi: 10.4230/DFU.Vol6.12191.45.

[170] G. N. Yannakakis and J. Togelius, "Modeling Players," in *Artificial Intelligence and Games*, Springer, 2018, pp. 203–255.

[171] R. Zubek, *Elements of Game Design*, First Edition. Cambridge, Massachusetts: The MIT Press, 2020.

Appendix A

Chapter Eight Survey questions

Please insert your generated game code here:

Through your gameplay you may have encountered three different characters. Please answer the following for each applicable character. (If the character did not show up in your session please write NA for Not applicable in the short follow up questions.)

The questions are on a scale of 1 to 5, with 5 being yes/loved/helpful and 1 being no/hated/unhelpful.

Do you feel that the characters **Amrock, Leo and Carrot** are aware of the world around them?

Please note, there are only three characters in this game. They may change color depending on your gameplay. We attached a couple of photos with the questions below.

Do you feel that the character **Amrock** is aware of the world around them?

unaware 1 2 3 4 5 aware

Why or why not? (Please write "NA" if you did not meet this character.) *

Do you feel that the character **Leo** is aware of the world around them?

unaware 1 2 3 4 5 aware

Why or why not? (Please write "NA" if you did not meet this character.)

Do you feel that the character **Carrot** is aware of the world around them?

unaware 1 2 3 4 5 aware

Why or why not? (Please write "NA" if you did not meet this character.)

Is it easy to understand what **Amrock** is thinking about?

no, not at all 1 2 3 4 5 yes, extremely so

Why or why not? (Please write "NA" if you did not meet this character.)

Is it easy to understand what **Leo** is thinking about?

no, not at all 1 2 3 4 5 yes, extremely so

Why or why not? (Please write "NA" if you did not meet this character.) *

Is it easy to understand what **Carrot** is thinking about?

no, not at all 1 2 3 4 5 yes, extremely so

Why or why not? (Please write "NA" if you did not meet this character.)

Do you think **Amrock** was persuadable?

not at all persuadable 1 2 3 4 5 easily persuaded

Do you think **Leo** was persuadable?

 not at all persuadable 12345 easily persuaded

Do you think **Carrot** was persuadable?

 not at all persuadable 1 2 3 4 5 easily persuaded

Do you believe that **Amrock** had a personality ?

 Yes No NA

If so, can you please describe it? (Please write "NA" if you did not meet this character.)

Do you believe that **Leo** had a personality ?

Yes No NA

If so, can you please describe it? (Please write "NA" if you did not meet this character.)

Do you believe that **Carrot** had a personality ?

Yes No NA

If so, can you please describe it? (Please write "NA" if you did not meet this character.)

Do you find the characters' behavior predictable ? yes no some of them are predictable

Can you elaborate on why you think that is / is not true (for any of the characters)?

Did you find these characters to be consistent?

Yes- No- some of them are consistent

Can you elaborate on why you think that is / is not true (for any of the characters)?

Can you picture social links between characters? (Do you think these characters know each other or have relationships?)

yes, they seem social - no, not at all -some seem social, some do not

Can you elaborate on why you think that is / is not true (for any of the characters)?

Did you find any of these characters to be memorable?

Amrock - Leo -Carrot - Amrock did not show up -Leo did not show up - Carrot did not show up

Would you like to elaborate?

Did you find any of these characters to be interesting?

Amrock - Leo - Carrot - Amrock did not show up - Leo did not show up - Carrot did not show up

Would you like to elaborate?

Did you find any of these characters to be lovable?

Amrock Leo Carrot Amrock did not show up Leo did not show up Carrot did not show up

Would you like to elaborate?

Did you find any of these characters to be hateful?

Amrock - Leo - Carrot - Amrock did not show up - Leo did not show up - Carrot did not show up.

Overall, what do you think of the characters you encountered?

Do you think the characters hold any moral values? Please elaborate.

What do you think of the characters' morality?

What values do you think the character **Amrock** cares about?

Values / level	is repulsed by this value	hates this value	is indifferent about this value	holds to this value	deeply holds this value	I think this value did not show up	NA (this character did not show up)
preserving the natural order				x			
being normal is important			x				
scared of spreading influence and ideologies							

a sense of character and constitution is important							
strength and will power is important							
abides and believes in authority							
if we were to help others, we need to first help ourself							
giving opportunities for growth is important							
empathy towards others is important							
it is important to nurture others							
distribution and fairness are important							
developing oneself comes from helping others develop							
one's happiness is important							
justice through retribution and restitution is of paramount							
preserving the natural order							
being normal is							

important							
scared of spreading influence and ideologies							
a sense of character and constitution is important							
strength and will power is important							
abides and believes in authority							
if we were to help others, we need to first help ourself							
giving opportunities for growth is important							
empathy towards others is important							
it is important to nurture others							
distribution and fairness are important							
developing oneself comes from helping others develop							
one's happiness is important							
justice through retribution and							

restitution is of paramount							

This question requires at least one response per row

What values do you think the character **Leo** cares about?

Values / level	is repulsed by this value	hates this value	is indifferent about this value	holds to this value	deeply holds this value	I think this value did not show up	NA (this character did not show up)
preserving the natural order				x			
being normal is important			x				
scared of spreading influence and ideologies							
a sense of character and constitution is important							
strength and will							

power is important							
abides and believes in authority							
if we were to help others, we need to first help ourself							
giving opportunities for growth is important							
empathy towards others is important							
it is important to nurture others							
distribution and fairness are important							
developing oneself comes from helping others develop							
one's happiness is important							
justice through retribution and restitution is of paramount							
preserving the natural order							
being normal is important							
scared of spreading influence and ideologies							

a sense of character and constitution is important							
strength and will power is important							
abides and believes in authority							
if we were to help others, we need to first help ourself							
giving opportunities for growth is important							
empathy towards others is important							
it is important to nurture others							
distribution and fairness are important							
developing oneself comes from helping others develop							
one's happiness is important							
justice through retribution and restitution is of paramount							

This question requires at least one response per row

What values do you think the character **carrot** cares about?

Values / level	is repulsed by this value	hates this value	is indifferent about this value	holds to this value	deeply holds this value	I think this value did not show up	NA (this character did not show up)
preserving the natural order				x			
being normal is important			x				
scared of spreading influence and ideologies							
a sense of character and constitution is important							
strength and will power is important							
abides and							

believes in authority							
if we were to help others, we need to first help ourself							
giving opportunities for growth is important							
empathy towards others is important							
it is important to nurture others							
distribution and fairness are important							
developing oneself comes from helping others develop							
one's happiness is important							
justice through retribution and restitution is of paramount							
preserving the natural order							
being normal is important							
scared of spreading influence and ideologies							
a sense of character and							

constitution is important							
strength and will power is important							
abides and believes in authority							
if we were to help others, we need to first help ourself							
giving opportunities for growth is important							
empathy towards others is important							
it is important to nurture others							
distribution and fairness are important							
developing oneself comes from helping others develop							
one's happiness is important							
justice through retribution and restitution is of paramount							

This question requires at least one response per row

Can you compare the morality of these characters with characters in other games you've played that have a morality system? Popular examples include the Bioshock series, the Elder Scrolls games (as well as most RPGs), Telltale games, and World of Warcraft, among many. please write NA if you have not played any games with any moral systems in place

Do you think the morality displayed in this game make the characters more or less engaging?

not at all engaging 1 2 3 4 5 more engaging

Do you think their morality makes the character more / less believable? (Believability in this case refers to a lifelike quality, seeming alive the way fictional characters can.)

not at all believable 1 2 3 4 5 yes, very much so

What values resonated from your gameplay with these characters and why or why not?

Who is your favorite and least favorite character and why?

Were there any conversations or elements in the game that struck you as odd or surprising?

Why or why not?

How many in-game days did you play?

- one day (game froze, did not restart)
- two days (game froze)

- three days (finished the game and got an ending screen)
- did not play the game
- one day (ran out of time) (had about 1-3 topic conversations)
- two days (ran out of time) (had about 4-6 topic conversations)

Other:

Any other comments or feedback?

End of Survey questions