Modeling Morality-Based Argumentation for Believable Game Characters: A Design Postmortem

Rehaf AlJammaz[1,2], **Michael Mateas** [1,2], **Noah Wardrip-Fruin** [1,2]

[1] University of California, Santa Cruz
[2] Expressive Intelligence Studio (EIS)
raljamma@ucsc.edu, nwardrip@ucsc.edu, mmateas@ucsc.edu

Abstract

An ability to morally reason is crucial to the believability of many fictional characters, from Jane Austen's heroines to the denizens of *The Good Place.* These works often foreground the complexity of moral questions and the circumstances under which different forms of behavior might be justified. Morality is also foregrounded in many games, from *Black and White* to *Mass Effect 3.* Yet, most in-game characters judge other characters (or the player) based on a single reputation scale or binary values of right and wrong. There has been little exploration in games of the relationship between character values and beliefs and moral reasoning. In keeping with this year's conference theme, "Oh the Humanity," this design postmortem paper describes the design and development of Argument Box, a model of moral argumentation and reasoning based on Lakoff's metaphor theory of moral politics. We describe our design approach, iterations, and authoring concerns — covering what went right and wrong in our attempts to model morality-based argumentation for believable game characters.

Introduction

Believable characters are essential in creating story worlds (Mateas 2001), engaging players, and immersing them (Warpefelt 2016). Yet the research literature often confuses *believable* characters with *realistic* characters (Aljammaz, Wardrip-Fruin, and Mateas 2023). Believable characters can engage audiences and encourage suspension of disbelief. Realistic characters are meant to act like human beings, in line with research findings from disciplines such as psychology or linguistics. Developing believable characters is one of the crucial aspects of game design.

Creating believable characters can include many aspects, including establishing a character's personality, motivation, social relationships, and growth, overall maintaining an illusion of life (Mateas 2001; Loyall 1997). Here we focus particularly on beliefs and values. A character's beliefs and values can have a cascading effect on believability aspects; through a character's beliefs, we can inform a character's motivation and goals. Beliefs can also individualize characters, creating unique personas. Furthermore, a character can gradually change their beliefs and ideologies, providing a basis for character growth and change.

We consider how beliefs and values can shape a character, particularly when used as a moral reasoning device. Morality is a core aspect of many games. Nevertheless, most games employ surface-level reasoning, often judging the player or other characters based on a single reputation scale or binary values of right and wrong. There needs to be more exploration of the relationship between character values and beliefs and moral reasoning. To that end, we present our project, Argument Box (AlJammaz, She, and Mateas 2021).

In Argument Box, the player's interactions are entirely through dialogue. In the course of dialogue, players interpret the utterances of characters. These utterances both make specific (surface-level) arguments about the moral actions of other characters and also reveal (through their selection and framing) deeply held moral beliefs that motivate the arguments. The player makes counter-arguments, also framed in terms of deeper beliefs, and sees how characters respond. This both provides further insights into the characters' beliefs and displays progress (or backsliding) in the attempt to change their minds. We believe this value-based-reasoning system adds depth to the traditional black-and-white morality systems seen in games such as *Mass Effect, Undertale,* and *Bioshock*, to name a few (BioWare 2007; Fox 2015; 2K Games 2007).

Unsurprisingly, designing value-based morality systems is a difficult task. Argument Box went through multiple design iterations. In each of our design iterations, we discovered issues that arose in developing these value-based reasoning systems, including communication issues, authoring problems, and design considerations. This paper will discuss lessons learned in a postmortem-like style, covering an overview of our developed systems as we progressed from our initial concept (AlJammaz, She, and Mateas 2021) to our current demo.

Related Work

This section will explore the relationship between social simulation systems and beliefs. We then examine morality as found in the wild and academia. We finally look at Lakoff's work that informs the underlying values of our system.

Belief Modeling in Social Simulation Games

Predictably, many social simulation games incorporate belief modeling in their designs. CiF-CK (Guimaraes, Santos, and Jhala 2017), for instance — an extension of the Comme il Faut (McCoy et al. 2011) social architecture — added characters who believe in false information alongside their social networks. Connan Exiles (Cif-EX) (Morais, Dias, and Santos 2019) extended character beliefs to incorporate other characters beyond the interacting character. In Talk of The Town (Ryan et al. 2015) — a historical, social simulation — characters can reason about their held beliefs; these beliefs are mutable and affected by various elements such as a character's memory and social network. Horswill's system MK ULTRA (Horswill 2015, 2018) allows players to inject false beliefs in NPCs, manipulating their knowledge bases to accomplish player goals and solve puzzles. Versu (Evans and Short 2013) — an episodic storytelling simulation — utilizes a model of belief in which characters share a public view of the world and specific instances of individualized beliefs; these beliefs inform goals that can cause a significant change in the narrative. Lastly, Azad et al.'s system, Lyra (Azad and Martens 2019), is a social simulation in which characters interact in politically charged groups, enables shared topic beliefs between characters, as informed by a character's private "attitude" or belief.

Morality in Games

Morality and games have a long history together. In earlier decades, game morality took the form of a protagonist fighting a villain. We can see many examples, from Sonic fighting Eggman to Mario rescuing Peach from Bowser (SonicTeam 1991; NintendoEAD 1990). Since then, game morality has evolved.

Nowadays, a variety of moral systems exist in games. Some games, like Telltale's *Batman* and *The Walking Dead,* enforce moral choices through high-stakes moments (TelltaleGames 2016, 2012), while others enforce morality through a player's lawful actions in the game's world, sporting a fame and infamy scale. Many RPG games follow the latter category; examples include wanted signs in *GTA,* character reactions in *Fable,* and guards in *Oblivion* (RockstarGames 1997; BigBlueBoxStudios 2008; BethesdaGameStudios 2006).

The approaches in some games, like *Fallout 4'*s (BethesdaGameStudios 2015) reputation system, base morality on the player's perspective and chosen alignment. Others, like *World of Warcraft* (WOW), take the player's alignment (horde or alliance) as opposing forces (BlizzardEntertainment 2004). Guards in WOW, for instance, reactively attack the player, despite the lore giving these characters values and beliefs. Unfortunately, these moral systems tend to judge morality as a scale of black to white, with shades of grey. We believe there is potential in, instead, creating value-based moral systems, giving each character a sense of uniqueness via their beliefs.

The few academic systems in this area tend to take more interesting approaches than commercial games, examining mortality through specific thought experiments, cognative modeling or moral theories. For example, Togelius (Togelius 2011) created a procedural prototype that abides by the categorical imperative principle. Using the same moral theory, Nelson (Nelson 2012) implemented a prototype that focuses on creating and breaking rules. Other projects, such as Harrell's chimera experiments (Harrell et al. 2018) showcase societal issues via interactive narratives grounded in cognitive and linguistic models. Examples include Chimeria: Gatekeeper (Harrell n.d), a narrative scenario that examines character identities, and Greyscale(Harrell et al. 2018), a chimera application that examines gender discrimination through character interactions.

We note morality also exists in hypertext and link-based games. Hypertext games typically employ content-based morality rather than systemic and procedural morality systems, which is our focus in this paper. Other games use alternate controllers as an added dimension for player choices; work by Sullivan et al. (Sullivan et al. 2018) showcases thread-cutting as a mechanic contributing to player agency.

Moral Politics

Our presented systems are heavily influenced by George Lakoff's work *Moral Politics* (Lakoff 2010). Lakoff is known for his work as a philosopher and cognitive linguist. In his work, Lakoff identifies societal and moral metaphors and relates them to the human experience. In some of his relatively recent works (*Moral Politics* and *Don't Think of an Elephant* (Lakoff 2010, 2014)), Lakoff identifies two broad systems of moral metaphors that society falls under, the strict father (SF) and the nurturant parent (NP) systems. As their names imply, they metaphorically represent society as a family system. A strict father, the head of the household, is harsh and stern; the metaphor system values elements like character, strength, listening to authorial figures, and independence. On the other hand, the nurturant parent system focuses on empathy, nurturing individuals, and fairness.

Both of these family systems utilize a set of moral metaphors and values they believe in to evaluate the morality of actions. The strict father, for instance, believes in moral boundaries (conforming to norms, seeing deviating from them as immoral), moral health (the "diseased mind argument," obligated to stop the spread of bad influences), and strength (seen as self-discipline or courage). The nurturant parent, on the other hand, values self-development (the development of nurturant abilities, not harmful abilities like torture), social nurturance (nurturing social ties), and empathy (seeing form another person's perspective).

Our work uses a subset of each family system's metaphors to account for our characters' moral judgments. We reference these metaphors in writing as *deep values.*

Version One: Aspirations and Beginnings

Gameplay Overview

Like the comedic Monty Python skit titled *"Argument Clinic"* (MontyPython 2009) in which a man purchases arguments from a clinician, in Argument Box NPCs procure arguments from the player in a local shop. The core gameplay loop consists of arguments going back and forth between the player and an NPC. The game starts when an NPC brings up a character they have mixed feelings about, either in a positive or negative light. The player's primary goal is to dissuade the NPC from judging the character negatively by engaging the NPC's core moral values.

The following sections will first describe our initial design and prototype. We will cover the systems overview, design issues, and problematic elements arising from the initial prototype. We then expand on lessons learned and design iterations covering what went right and wrong. Lastly, we will present the current version. We will cover design adjustments and developmental changes, after which we end our paper with a series of informal playtesting results.

Reviewing a City of Shapes

Our initial system was fully described in (AlJammaz, She, and Mateas 2021) However, a summary of this design is provided here in order to ground our explanation of the process of design evolution.

Our first version featured a city of shapes where cube-like NPCs approached the player with gossip they wished to unpack. The player could agree with or oppose their ideals conversationally.

This first system separated NPCs in the world into two types, NPCs that the player converses with, called conversational NPCs (CNPCs), and NPCs that the CNPCs talk about, called background NPCs (BNPCs). We imported the latter type from *Talk of the Town,* a historical town simulator (Ryan et al. 2015). The gameplay generally began when a CNPC selected a BNPC to discuss. The selection process was filtered based on compelling behavior patterns found on the BNPC characters. Once a BNPC was selected, the CNPC mapped the initial pattern found on the BNPC to an unbiased rumor. For example, the pattern *InLoveWithAFriendsSpouse* gets translated as "Have you heard that the [BNPCNAME] is in love with their friend's spouse?"; This was then mapped to the CNPC's surface-level biased opinion of the pattern (e.g., *InLoveWithAFriendsSpouse* could be mapped to the surface value (SV) *LoveIsForFools*, viewing the act as immoral). The CNPC could take either a positive or a negative stance.

The player then had the option to disagree or agree with the CNPC's opinion. If the player disagreed, the CNPC could push back or let go of the topic, conceding to the player. The CNPC pushed back or doubled down if they held that surface value strongly. If they did, the game loop transitioned into the inner model, called *deep values.* As explained in the related work section, these deeper values were based on Lakoff's categorization of strict father (SF) and nurturant parent (NP) metaphors.

In the deep value loop, the selected behavior pattern was mapped to one of the deep values of the CNPC's primary moral model (SFM or NPM). For instance, the pattern *InLoveWithAFriendsSpouse* was mapped to the deeper values *low_moralBoundries* as part of the CNPC's SFM. If, on the other hand, the NPC held the NPM, then this behavior pattern would have been mapped to a deep value such as *low_empathy*. This inner loop then consisted of the player selecting a conversational approach that adjusted the player's presented argument as part of the strict or nurturant deep values. Once the player selected a model, they encountered multiple options matching their selected model. At this point, the player could also change the selected topic to a different behavior pattern held by the BNPC. Figure 1 shows the system's conversational loop. Our earlier work (AlJammaz, She, and Mateas 2021) elaborates on this version's system and moral modeling specifications.

Hard to Author and Read

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

Flow, Loop Design and Control. 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 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
    of friends must be good shapes.
```

As noticed from the conversation above, the transition between the morality of butchering animals to the BNPC's 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

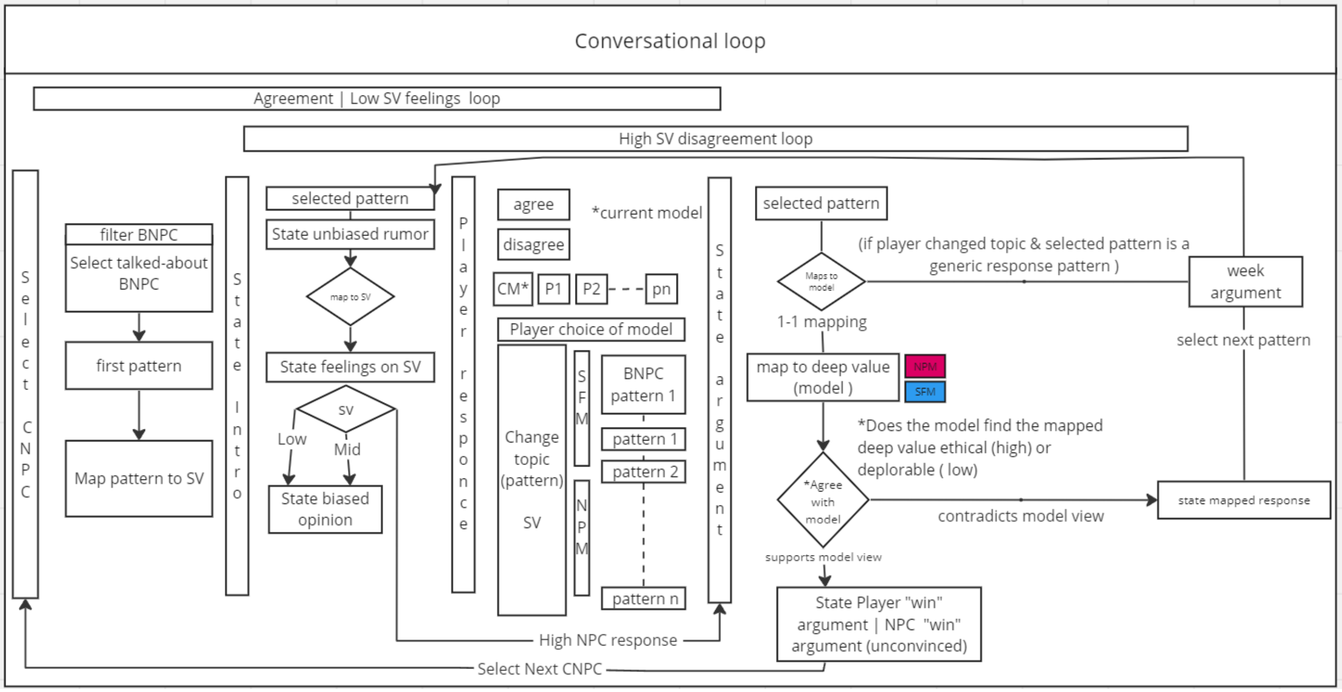


Figure 1: Argument Box: version one's conversational loop

to do given that the system randomly assigned a topic's surface values (low, mid, high) in its initialization phase

Context and Authoring Burdens. Unsurprisingly, authoring conversational flows at different levels was a taxing effort for our writer. As covered earlier, each BNPC behavior 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 them, including transitions such as unbiased rumors and introductions. We had 29 surface values (including pro and con variations), authored at three surface value levels as well as introductions, agreements, focus areas, quick responses, and disagreement texts; this leads to about 16-19 strings per surface value (551 strings). Additionally, our deep values (moral model) contained responses for 54 behavior patterns, each mapped to a single deep value in a pro or con stance. 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). Given that these responses could be selected dynamically in many different orders depending on the state of the moral argument, transitions between responses 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 due to transitional issues, miscommunication, or logical bugs.

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 became contradictory and confusing.

Player Feedback and Deep Value Impact. Our initial 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.

In later sections we will illustrate how our current system handles these design issues. Next, we list elements that were not problematic but posed design constraints and trade-offs.

- **Reusing a cast of characters.** At the start of the project, we chose to import our cast of characters from *ToTT* (Ryan et al. 2015). 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 restricted 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.

- **Separating our cast of characters.** The first version initially 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 a CNPC could hold. This strategy, however, came at the expense of the player losing familiarity and ability to interact with BNPCs.

Version Two: Simplifying the Model

The Half-Way Mark

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* (Byron Howard 2016; Itagaki 2016); 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 (elaborated in the following section).

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).

A Mammal Society: 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. Similarly 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. Figure 2 shows sample gameplay.

Modification and Changes

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

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.

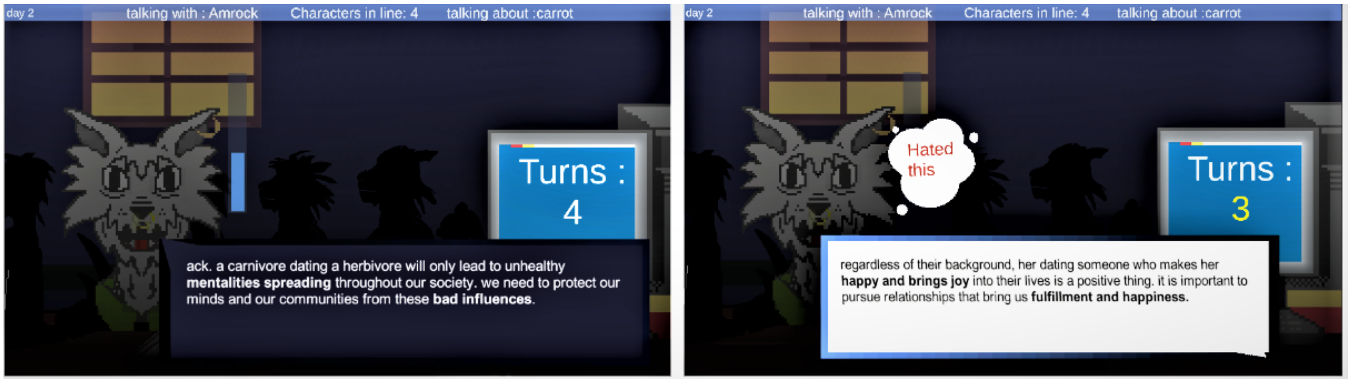
For example, suppose the talked-about NPC has the pattern *DatingOtherSpecies,* and the conversational NPC uses the surface value *RomancingAnotherSpecies* with a con moral stance. In that case, the player is directly given a shuffled list of pro-stance deep values applicable to the pattern *DatingOtherSpecies.* These deep values are a combination of the metaphorical SF and NP deep values affiliated with each surface value; they contain up to 12 deep values for both pro and con versions, depending on compatibility. Examples include *MoralHealth* and *MoralBoundries* for the SF model and *Empathy* and *Happiness* for the NP model.[1] Figure 3 shows how deep values are made more explicit to the player, by bolding parts of the text that reference the deep value.

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 how it relates surface values to deep values. Each surface value is associated with up to 12 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.

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

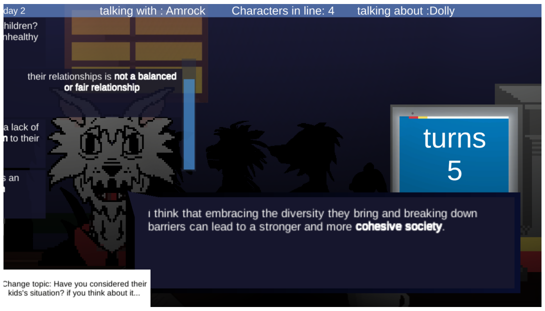
[1] Each of the 12 selected deep values are contextualised for the selected pattern, if applicable. For instance, *Moral Boundaries* views anything that deviates from the norm as immoral. In this case *DatingOtherSpecies* translates as *"A carnivore dating a herbivore is* ***unnatural.****"*



ack. a carnivore dating a herbivore will only lead to unhealthy **mentalities spreading** throughout our society. we need to protect our minds and our communities from these **bad influences**.

regardless of their background, her dating someone who makes her **happy and brings joy** into their lives is a positive thing. it is important to pursue relationships that bring us **fulfillment and happiness**.

Figure 2: Sample conversation and reaction. On the left we see the value (in bold) the CNPC is arguing from, on the right the value the player is arguing from

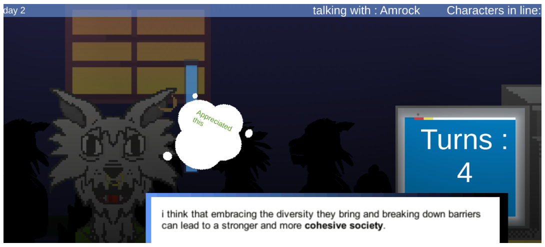


their relationships is **not a balanced or fair relationship**

ı think that embracing the diversity they bring and breaking down barriers can lead to a stronger and more **cohesive society**.

Change topic: Have you considered their kids's situation? if you think about it...

Figure 3: Bold text in white highlights the statement's deep value in short-hand and expanded forms



i think that embracing the diversity they bring and breaking down barriers can lead to a stronger and more **cohesive society**.

Figure 4: An NPC's persuadablity bar and thought bubble graphic

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 understand which values are most important to the NPC. To generate the player's conversational choices, the system shuffles 4-5 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.

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 4.

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 communication issues and misunderstandings of 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 (OpenAI 2023) 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 summarise the overall flow of our game loop and conversation structures.

Character Structures

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 (Goldberg 1990). 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 the *coward* pattern, the system may add the pattern *ScaredOfCarni* 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.

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.*

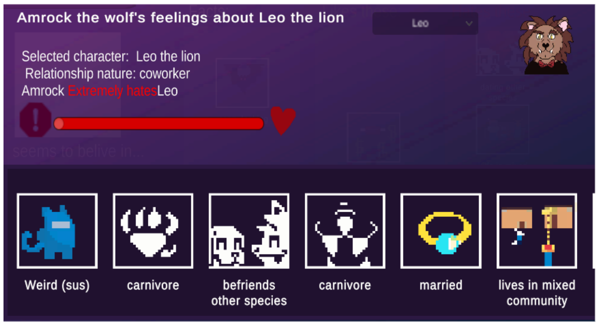
```
I [CNPCOpinion] [BnpcName] the [
    talkedAboitAnimalSpecies]. I  [
    SvFeelings] that [pronoun] is [
    romanticRelationship] other species!
```

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 the keywords indicating the NPC’s deep values are bolded. The persuasibility score is then adjusted 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(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. 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.



Amrock the wolf's feelings about Leo the lion

Selected character: Leo the lion
Relationship nature: coworker
Amrock Extremely hatesLeo

seems to belive in...

Figure 5: Sample UI deisgn. The screen depicts the NPC Amrock's feelings about Leo, a character Amrock hates and Leo's list of patterns

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 5 highlights these UI designs.

Informal Play-Tests

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 current 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 naive to the experience. During our play-tests, players were instructed to use the "think aloud method" (Lankoski, Björk et al. 2015) elaborating aloud on their game-play experience. Here we highlight feedback attained from our players for the current version. 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, strict father, so I'm selecting something close to 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, "Oh 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 decreased their relationship with Dolly, the talked-about NPC).

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 conversations made sense and natural. Although, There were a few instances where the NPC sent "mixed signals,"; this 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* 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.

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 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 1-3 associated patterns (and 1-12 pro and con deep value arguments, respectively). We note that the variability of these patterns (and cared-about surface

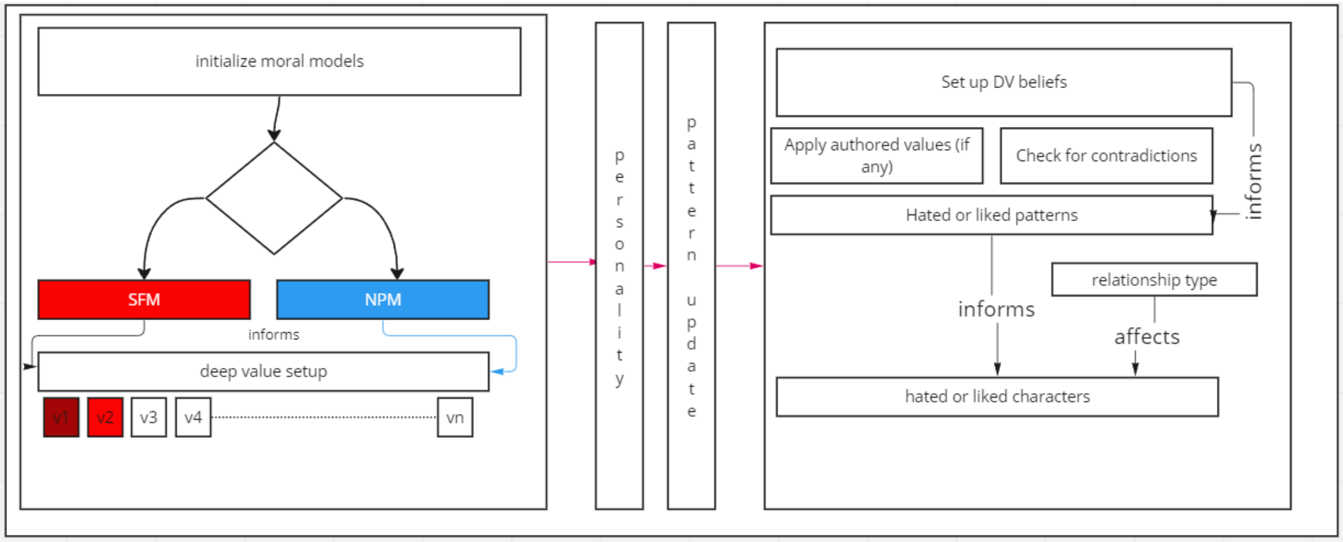


Figure 6: Argument Box character setup diagram

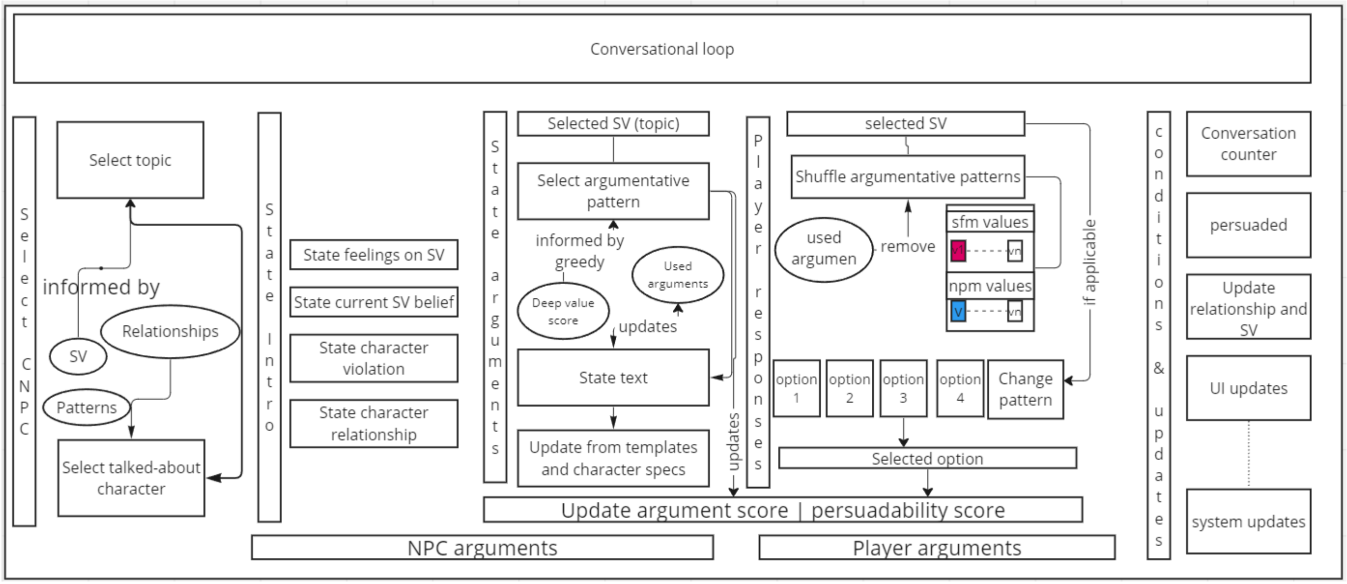


Figure 7: Argument Box: version two the conversational system diagram

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, such as missing stylized and bolded text in some strings or cut-off texts in other boxes.

Conclusion

We consider this paper a crucial part of our strategy for creating a more believable character via moral reasoning. We believe moral reasoning can help with character change and development, a feature of believability identified in our earlier work (Aljammaz, Wardrip-Fruin, and Mateas 2023). While our first system focused on identifying and building an initial candidate system (AlJammaz, She, and Mateas 2021), this paper focuses on an overview of the two systems, considering the game from a design perspective. Throughout this postmortem-style paper, we covered design strategies, issues faced, and adjustments made. In our future work, we plan to include a formal evaluation(from a believability perspective) and expand the technical details of our current system.

We hope others can benefit from our design mistakes and lessons learned, particularly in simplifying value-based systems, presenting feedback, and conversational flow. We hope that this project can be seen as an example of an alternative morality-based system for future developers.

References

2K Games, I. G. 2007. BioShock. [Playstation 3]. 2K.

AlJammaz, R.; She, Y.; and Mateas, M. 2021. Argument Box. In *AIIDE Workshops, Experimental AI in Games (EXAG)*.

Aljammaz, R.; Wardrip-Fruin, N.; and Mateas, M. 2023. Towards an Understanding of Character Believability. In *Proceedings of the 18th International Conference on the Foundations of Digital Games*, 1–9.

Azad, S.; and Martens, C. 2019. Lyra: Simulating believable opinionated virtual characters. In *Proceedings of the AAAI conference on artificial intelligence and interactive digital entertainment*, volume 15, 108–115.

BethesdaGameStudios. 2006. The Elder Scrolls IV: Oblivion. [PlayStation 3].Bethesda Softworks, 2K.

BethesdaGameStudios. 2015. Fallout 4. [PlayStation 4]. Bethesda Softworks.

BigBlueBoxStudios, L. 2008. Fable II. [Xbox 360]. Microsoft Game Studios, Feral Interactive.

BioWare. 2007. Mass Effects. [xbox 360]. Electronic Arts.

BlizzardEntertainment. 2004. World of Warcraft. [Microsoft Windows]. Blizzard Entertainment.

Byron Howard, R. M. 2016. Zootopia. Walt Disney Studios Motion Pictures.

Evans, R.; and Short, E. 2013. Versu—a simulationist storytelling system. *IEEE Transactions on Computational Intelligence and AI in Games*, 6(2): 113–130.

Fox, T. 2015. Undertale. [Microsoft Windows]. Toby Fox.

Goldberg, L. R. 1990. An alternative" description of personality": the big-five factor structure. *Journal of personality and social psychology*, 59.6: 1216.

Guimaraes, M.; Santos, P.; and Jhala, A. 2017. CiF-CK: An architecture for social NPCs in commercial games. In *2017 IEEE Conference on Computational Intelligence and Games (CIG)*, 126–133. IEEE.

Harrell, D. n.d. Chimeria: Gatekeeper. [Web]. Computation, and Expression Laboratory (ICE Lab).

Harrell, D. F.; Ortiz, P.; Downs, P.; Wagoner, M.; Carré, E.; and Wang, A. 2018. Chimeria: Grayscale: an interactive narrative for provoking critical reflection on gender discrimination. *MATLIT*.

Horswill, I. 2015. Mkultra. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 11, 223–225.

Horswill, I. 2018. Postmortem: MKULTRA, an experimental AI-based game. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 14.1, 45–51.

Itagaki, P. 2016. *BEASTARS*, volume 1 of *BEASTARS*. Akita Shoten.

Lakoff, G. 2010. *Moral politics: How liberals and conservatives think*. University of Chicago Press.

Lakoff, G. 2014. *The all new don't think of an elephant!: Know your values and frame the debate*. Chelsea Green Publishing.

Lankoski, P.; Björk, S.; et al. 2015. *Game research methods: An overview*. Lulu. com.

Loyall, A. B. 1997. Believable Agents: Building Interactive Personalities. Technical report, Carnegie-mellon Univ Pittsburgh Pa Dept of Computer Science.

Mateas, M. 2001. An Oz-centric review of interactive drama and believable agents. In *Artificial intelligence today: Recent trends and developments*, 297–328. Springer.

McCoy, J.; Treanor, M.; Samuel, B.; Wardrip-Fruin, N.; and Mateas, M. 2011. Comme il faut: A system for authoring playable social models. In *Proceedings of the AAAI conference on artificial intelligence and interactive digital entertainment*, volume 7, 158–163.

MontyPython. 2009. Argument Clinic. YouTube. URL: https://www.youtube.com/watch?v=DkQhK8O9Jik.

Morais, L.; Dias, J.; and Santos, P. A. 2019. 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*, 1–11.

Nelson, M. 2012. Prototyping kant-inspired reflexive game mechanics. In *Proceedings of the 2012 Workshop on Research Prototyping in Games*.

NintendoEAD. 1990. Super Mario World. [Super Nintendo Entertainment System]. Nintendo.

OpenAI. 2023. *ChatGPT [GPT-3.5 architecture]*. OpenAI.

RockstarGames. 1997. Grand Theft Auto. [CD-ROM]. Rockstar Games.

Ryan, J.; Summerville, A.; Mateas, M.; and Wardrip-Fruin, N. 2015. Toward characters who observe, tell, misremember, and lie. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 11.3, 56–62.

SonicTeam. 1991. Sonic the Hedgehog. [Sega Genesis]. Sega.

Sullivan, A.; McCoy, J. A.; Hendricks, S.; and Williams, B. 2018. Loominary: crafting tangible artifacts from player narrative. In *Proceedings of the Twelfth International Conference on Tangible, Embedded, and Embodied Interaction*, 443–450.

TelltaleGames. 2012. The Walking Dead. [PlayStation 4].Telltale Games.

TelltaleGames. 2016. Batman: The Telltale Series. [PlayStation 4]. Telltale Games.

Togelius, J. 2011. A Procedural Critique of Deontological Reasoning. In *DiGRA Conference*.

Warpefelt, H. 2016. Chapter 6 Typology of Non-player characters. In *in The Non-Player Character: Exploring the believability of NPC presentation and behavior.*, 99–104. PhD diss.