# Mental State Modelling: A Philosophy of Mind Approach to Emergent Believable Behaviour in Non-Player Characters

SOLOMON INYANG, The Australian National University, Australia
PENNY SWEETSER, The Australian National University, Australia
ANNE OZDOWSKA, The Australian National University, Australia

Video game worlds allow for meaningful social interaction with non-player characters (NPCs), but NPCs are often not perceived as believable by players, which negatively impacts player immersion. Existing believable agent architectures attempt to achieve believability by giving NPCs human-like qualities such as personality and emotion. However, most believable agent methodologies disregard the importance of modelling internal mental processes. We use philosophical theories of mind to create a believable agent architecture called Mental State Modelling (MSM). The MSM architecture provides NPCs with internal mental states that can influence their interactions with players, other NPCs and the environment to produce emergent and more believable behaviours. Eighteen participants took part in an experimental study that compared the perceived believability of NPCs made using the MSM architecture with an alternative Finite State Machine (FSM) technique. We found that the MSM architecture was able to portray most of the intended believability qualities in NPCs and achieve a higher overall believability than the FSM technique. We conclude that utilising a philosophy of mind-based believable agent architecture could allow developers to successfully create more believable NPCs for games.

CCS Concepts: • **Computing methodologies** → *Theory of mind*; *Intelligent agents*; • **Applied computing** → **Computer games**.

Additional Key Words and Phrases: believable agents, human-AI interaction, philosophy of mind

**ACM Reference Format:**
Solomon Inyang, Penny Sweetser, and Anne Ozdowska. 2024. Mental State Modelling: A Philosophy of Mind Approach to Emergent Believable Behaviour in Non-Player Characters. 1, 1 (October 2024), 18 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 Introduction

Video games present unique interaction opportunities by creating immersive experiences that engross players in fictional worlds. Games facilitate immersion by allowing players to directly interact with lifelike NPCs and build meaningful relationships with them. These lifelike non-player characters can take various forms, including being built into the player's UI or acting as embodied agents that are situated in the same environment as the player. Despite this unique potential, NPCs developed using common game AI approaches like Finite-State-Machines (FSM) [22] or Behaviour Trees [28] are often limited to pre-scripted behaviours. NPCs that exhibit static behaviours clash with other highly immersive aspects of many game worlds. For example, players and reviewers have criticised AAA RPG games like the recently released Starfield [35], for their lack of believable NPC behaviour which often reduces the overall immersiveness of the game environment.

New AI platforms like ConvAI and Inworld have emerged to improve the conversational capabilities of game characters through features like environmental awareness, reasoning and natural language understanding and generation [1, 2]. These features support the development of NPCs that are more context aware and responsive, but the effectiveness of these tools for producing complex believable behaviour has yet to be demonstrated in a commercial game. In an effort

Authors' Contact Information: Solomon Inyang, The Australian National University, Canberra, Australia, u6701825@anu.edu.au; Penny Sweetser, The Australian National University, Canberra, Australia, larst@affiliation.org; Anne Ozdowska, The Australian National University, Canberra, Australia, .

2024. Manuscript submitted to ACM

to investigate the potential of using generative AI to simulate realistic human behaviour, Park et al. [32] describe an agent architecture that stores, synthesises and applies relevant memories to generate believable behaviour using a large language model (LLM). The proposed architecture allows NPCs to simulate human behaviour by remembering past actions and objectives, reflecting on past experiences and creating dynamic plans for evolving circumstances, which leads to more believable actions. However, using generative AI for NPCs is computationally expensive, difficult to implement and not currently practical for multiple NPCs [32].

The research presented in this paper introduces a new approach to developing NPC behaviour called Mental State Modelling (MSM). The MSM architecture aims to enhance the believability of NPCs by assigning internal mental states to them. Compared to existing methods which focus on specific believability qualities like memory or language, the MSM architecture takes a holistic systems theory approach to achieving NPC believability. It allows believable and dynamic behaviour to emerge from the systemic interactions of internal mental states that are each associated with different believability qualities such as emotion, personality and motivation.

The human mind is a complex system from which human-like qualities, like personalities, emotions, and social ideologies emerge based on the interactions of internal mental processes [6]. However, most existing believable agent architectures adopt a behaviourist approach to AI, in which behaviour is represented purely through observable actions without much consideration for internal mental phenomena [41]. As such, we ask the research question: **How can philosophy of mind theories be integrated into a believable agent architecture to develop NPCs that are perceived as more believable by players compared to NPCs that use a current game AI approach?**

To answer our research question, we first investigated how philosophy of mind concepts could be integrated into a believable agent architecture that produces emergent and believable NPC mental processes and behaviour. We developed the Mental-State-Modelling (MSM) architecture to create believable NPCs that exhibit emergent believable behaviour. The MSM architecture combines considerations for emergent behaviour [40] with the philosophical theory of functionalism [33] to develop NPCs whose believability qualities are expressed as emergent phenomena that arise as a result of different mental state interactions.

In video games, the most common strategy for developing NPC behaviour is to transition between actions after specific conditions are met. FSMs and behaviour trees provide good examples of this kind of approach. Another approach is to use utility AI. Utility AI relies on the premise that an action will be selected when the utility provided by that action exceeds the utility provided by others. In contrast to these methods, the MSM architecture provides a framework for NPCs to dynamically select the appropriate behavioural response for a situation based on their emotions, beliefs and other internal mental properties they are experiencing.

A second goal of this research was to address the disconnect between believable agent research and AI approaches that are used in the games industry. We developed an experimental study and data analysis technique that asked players to compare the believability of two groups of NPCs in a game, where one group used the MSM architecture to portray the believability qualities of personality, emotion, self-motivation, change, social interaction and situatedness, and the other group used an FSM implementation that did not account for these qualities. We found that the MSM architecture was able to portray most of the intended believability qualities in game NPCs and achieved a higher overall believability than the FSM implementation.

The contributions of our research are threefold. First, we present a **unified believable agent architecture** based on well-established philosophy of mind theory for creating NPCs that can exhibit emergent believable behaviour. Second, a **reproducible experimental methodology** for evaluating the believability of two contrasting NPC AI systems in a game is developed. Third, **insights on applying the MSM architecture** to existing games to allow developers to craft believable NPC behaviours for different scenarios are presented.

## 2 Related Work

Our review of related work begins with the concept of a believable agent, and the qualities of believable agents that have been proposed in the literature. We also explore existing believable agent architectures that have been developed to achieve the qualities of emotion, personality, and social interaction. Finally, we discuss the basic principles of the philosophical theory of functionalism and how it could be applied to the development of believable agents.

### 2.1 Believable Agents

Bates [29] first defined a believable agent as a virtual character that is both autonomous (exhibits reactive, goal-driven behaviour in a dynamic environment) and believable (possesses rich human-like characteristics such as personality and emotion). Bates proposed an initial set of qualities that were necessary for virtual characters to be considered believable agents, which included personality, emotion, self-motivation, change, social relationships, and consistency of expression. Loyall [27] later expanded the set of believability qualities by including the illusion of life quality, which is comprised of various sub-qualities including situatedness, reactivity, and responsiveness. Over time, the illusion of life quality has been further expanded to include environment, social, and interaction awareness [8, 42].

Our research focuses on the believability qualities of personality, emotion, self-motivation, change, and social relationships as outlined by Bates [29]. For illusion of life, we only considered the sub-quality of situatedness, defined by Loyall [27] as the ability of a believable agent to be situated in the game environment in which they exist. This sub-quality was chosen because it encompasses multiple adjacent illusion of life sub-qualities, including reactiveness, responsiveness, and awareness [17]. Consistency of expression was omitted as it is reliant on extensive research in areas such as facial expression, gesture, and natural language, which was outside the scope of this research. The believability qualities studied in this research are outlined in Table 1.

Table 1. Believability Qualities.

| **Believability Quality** | **Description** |
|---|---|
| *Personality* | Unique characteristics that produce individuality in an agent's behaviour. |
| *Emotion* | Expression of an agent's emotional states, which are consistent with their personality. |
| *Self-Motivation* | Goal-driven behaviour based on the agent's internal drives and desires. |
| *Change* | Growth and change of the agent in ways that are consistent with their personality. |
| *Social Interaction* | The agent's ability to engage in social interactions with other characters, which influence, and are influenced by, their relationships with these characters. |
| *Situatedness* | The agent is aware of their surroundings and is able to sense and react to things that happen around them. |

### 2.2 Emotion

Most believable agent architectures implement the Ortony, Clore, and Collins (OCC) model [31]. The OCC model specifies 22 types of emotional reactions, including joy, love, anger, and pity. These emotional reactions involve an appraisal process, in which an individual evaluates their current environment, as well as intrinsic factors such as their current arousal level, to determine the most suitable emotional reaction for a given situation. Additionally, some studies have attempted to expand on the OCC model by integrating subjective emotional experience [3], emotional history [7], and emotional expression through facial expressions [5] and gestures [24]. A notable limitation of existing approaches is that emotions are typically not influenced by other qualities like self-motivation and change [29]. Furthermore, believable agents that use the OCC model are often limited to the 22 types that are specified in the original model, which leaves out some emotions such as sadness and jealousy [20].

### 2.3 Personality

Alongside the development of emotional computational models for believable agents, some researchers have taken a different path to believable agents by attempting to imbue them with human-like personality traits. The Five-Factor model of personality (FFM) [37], which is the dominant personality theory, states that an individual's personality is positioned somewhere on a continuum of five different personality traits: openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism. Encoding the personality traits represented in the FFM model within a set of rules for behaviour has proven to be an effective and popular method of creating a sense of personality in NPCs. To develop agents that possess both personality and emotion, Allbeck and Badler [3] combined the OCC model of emotion and the FFM to create embodied conversational characters that can exhibit emotional reactions based on their personality traits.

A limitation of previous works is that the personalities of agents usually do not change during gameplay. Inability to change can cause dynamic game worlds to no longer make sense if a character's personality is not responsive to significant game events. While there have been some studies conducted on creating adaptive personalities for NPCs using neural networks [34], our research aims to create believable NPC behaviour by allowing personality, among other qualities, to arise dynamically through the NPC's mental state interactions.

### 2.4 Social Interaction

Parallel to the work on emotion and personality, the third believability quality that has received significant attention in the literature is social interaction. Research in this area aims to allow NPCs to engage in social interactions with other characters, including the player character, and build social relationships that will influence these interactions. Much of the work in simulating social behaviour has involved implementing psychological models of personality and emotion that facilitate various types of social interactions. For example, previous work by Doce et al. [13] applied the FFM model to develop agents that demonstrate personality-based facial expressions and coping behaviour when in a group.

Faur et al. [16] present one of the few studies that investigate internal cognitive processes in NPCs. This work uses the FFM model to simulate the dynamics of personality within a social context. The agents developed in this work possess two competing senses of self, an actual self – how the agent's personality dynamics actually appear as modelled by the FFM – and an ideal self – how the agent believes that they ought to be. There is currently limited research on NPCs that can possess a sense of perspective and point of view. The MSM architecture aims to address this research gap by facilitating internal perceptual states that allow for multiple perspectives of a situation or event.

### 2.5 Functionalism

Functionalism is a computational theory of the mind that was proposed by Putnam [33] in 1967. Putnam [33] proposed a version of this theory now known as *machine functionalism* which claims mental states are properties of the brain that can be defined purely in terms of their functional role within the mind. He considered the mind to be any system that takes in sensory inputs and produces behavioural outputs based on the interactions of internal mental states. Putnam [33] stated that the relationships between sensory inputs, behavioural outputs, and mental states are specified by a set of rules which are stored in a theoretical data structure within the mind called a “Machine Table”. The main claim of Functionalism is that the mind is essentially an abstract mathematical model of computation equivalent to a Turing Machine [43], in which different symbols are manipulated to produce outputs based on a set of rules.

To develop emergent believable NPC behaviour, the believability qualities described in the literature must emerge from a complex system that mimics the human mind. Functionalism affords a frame of reference for modelling mental states algorithmically. Therefore, to create believable agents that can exhibit emergent believable behaviour, we propose a believable agent architecture grounded in functionalism called Mental State Modelling.

## 3 Methodology

We developed a novel believable agent architecture called Mental State Modelling (MSM) to create a method for developing believable NPCs. The MSM architecture adapts Putnam’s theory of functionalism [33] into a computational model that allows NPCs to recognise environmental stimuli as sensory inputs, process this stimuli and produce specific behavioural outputs. Developers can assign an initial set of mental states to each NPC and specify a set of rules for how each mental state interacts with sensory inputs and other mental states to produce behavioural outputs. The interactions of an NPC’s mental states can also allow them to acquire entirely new mental states based on situations in the game. By allowing complex behaviour to arise from each NPC’s internal processing of the game world, the MSM architecture produces emergent behaviour in NPCs.

### 3.1 Architecture Overview

Our MSM architecture, illustrated in Figure 1, consists of three components: a sensory component, a mental state component, and a behaviour component. Each component corresponds to a specific part of the mind structure described in functionalism. The sensory component accounts for the sensory inputs in functionalism and allows NPCs to detect objects and sounds in the game world. The mental state component holds all of an NPC’s mental states as well as the “Machine Table” defined by Putnam [33]. The machine table is represented as a collection of scripts in which rules are outlined that dictate how mental states interact with sensory inputs and each other. The behaviour component produces the behavioural outputs outlined by Putnam [33]. In this component, a planning algorithm is used to create a behaviour tree that maps various actions to different mental states.

### 3.2 Sensory Component

The sensory component of the MSM architecture gives NPCs the ability to perceive environmental stimuli in the game world that may impact their mental states. In the implementation described in Section 4, each NPC is equipped with a vision system and an auditory system that allows them to detect environmental stimuli in the game world. Objects in the game environment, including NPCs themselves, possess tags that describe their visual properties which can be detected by the vision system.

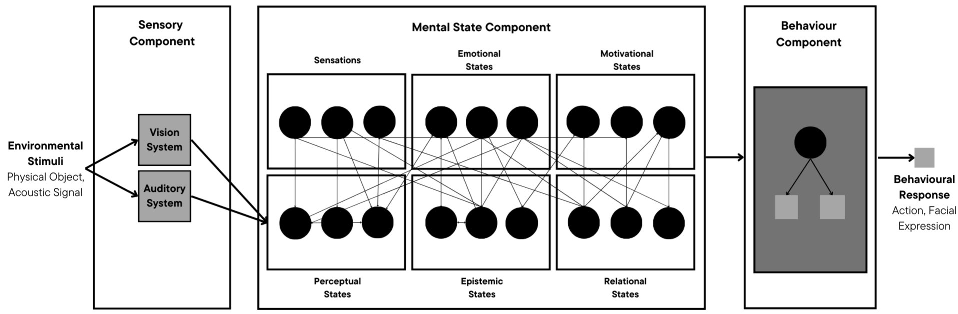
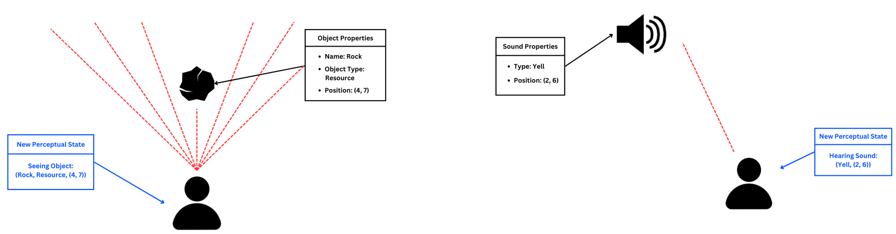



Fig. 1. A conceptual model of the MSM Architecture.

The vision system (shown in Figure 2a) replicates visual perception by allowing NPCs to automatically detect the visual properties of objects in the game world that enter their field of view. NPCs can detect objects within their field of view by sending out raycasts (represented as dashed lines in Figure 2a) from the NPC's current position in the direction they are facing. If a raycast collides with an object, the NPC will be able to identify the visual properties that are assigned to the detected object and store information on the object's properties internally as a perceptual state.




(a) Vision system (b) Auditory system

Fig. 2. An NPC acquiring a new perceptual state through the vision system (a) and auditory system (b).

The auditory system (shown in Figure 2b) is designed to mimic auditory perception by allowing NPCs to identify the type and location of acoustic signals that are emitted in the game world. In the game described in Section 4, objects in the game world can broadcast sound sources that act as acoustic signals carrying information about the location, direction and type of sound (e.g., screaming, singing, etc). An NPC can then detect these signals through their auditory system which measures broadcasted sounds that are within a certain distance from each NPC. Broadcasted sounds are also stored as perceptual states and, together with the visual information gathered from the vision system, they provide

NPCs with information on their environment that can be linked to other categories of mental states in order to select an appropriate behavioural response to the NPCs' surroundings.

### 3.3 Mental State Component

The mental state component describes a set of rules for how mental states process inputs from the sensory component and coordinate with other mental states. Despite serving as the basis for many different philosophical and psychological theories including functionalism [33], the concept of a mental state remains poorly defined in the scientific literature [11]. Therefore, for the MSM architecture we developed a categorisation of mental states that draws influences from philosophy of mind theories that we believed could provide insights on achieving the six believability qualities outlined by Loyall [27]. The theories that inspired our categorisation are listed in Table 2.

Table 2. List of philosophy of mind theories that inspired the mental states categories in the MSM architecture.

| Theory | Key Philosopher | Key Argument |
|---|---|---|
| Transcendental Idealism [4] | Immanuel Kant | Human understanding of their environment is constructed from sensory impressions perceived through senses like taste, smell, touch and hearing. |
| Causal Theory of Perception [26] | John Locke | Objects in the environment possess perceptible qualities that cause humans to experience sensations. |
| Basic Emotion Theory [14] | Paul Ekman | There are 6 basic emotions - anger, fear, enjoyment, sadness and surprise - that allow humans to adapt to changes in their environment. |
| Theory of Rational Changes of Belief [18] | Peter Gärdenfors | Humans are capable of updating their beliefs about their environment by constantly revising their beliefs when new pieces of information are acquired. |
| Humean Theory of Motivation [36] | David Hume | Humans possess mental states like desires which motivate them to perform actions in the world. |
| The I-Thou Relationship [10] | Martin Buber | Humans experience two types of relationships: "I-Thou" an "I-It" which differ in their meaningfulness. |

Our mental state categorisation is depicted in Figure 3. We establish 6 broad categories of mental states: sensations, perceptual states, emotional states, epistemic states, motivational states, and relational states. While each believability quality may be closely related to a specific mental state category, creating NPCs that possess all the qualities found in the literature and exhibit emergent believable behaviour is contingent on the specification of dynamic interaction rules between mental states from all categories. For example, achieving the quality of personality involves giving each NPC distinct emotions, motivations, knowledge, and beliefs that will cause each of them to interpret the game world differently and consequently exhibit behaviours that are different from one another.

We categorised sensations as mental states that provide NPCs with information on their own physical state (e.g., hunger, thirst). Perceptual states were categorised as mental states that allow NPCs to process information on the physical states of other objects. Sensations and perceptual states are particularly relevant to the quality of situatedness as they give NPCs awareness of themselves and their surroundings. Epistemic states were categorised as mental states that represent an NPC's knowledge and beliefs about their environment. Sensations, perceptual states and epistemic states change dynamically based on events that occur in the game world, which trigger changes to an NPC's emotional states,

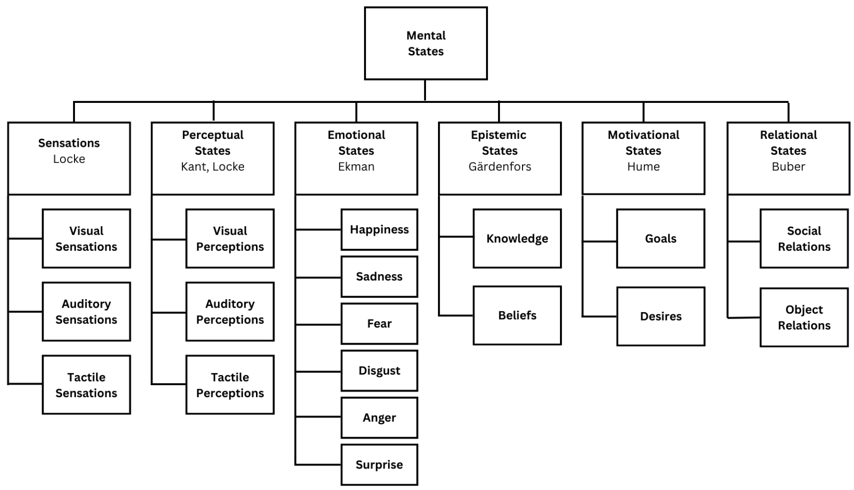



Fig. 3. A taxonomy of mental state categories including influential philosophers whose work contributed to each category.

motivational states and relational states. The dynamic nature of the mental states described in the MSM architecture allows NPCs to possess the believability quality of change.

We categorised emotional states as mental states that link the environmental information received by sensations and perceptual states to emotional responses. The emotional state sub-categories in the MSM architecture are inspired by Basic Emotion Theory [14]. In Basic Emotion Theory, Ekman [14] defines six types of basic emotional states: happiness, sadness, fear, disgust, anger, and surprise. Each basic emotional state has a specific duration and intensity, and from these emotions, more complex emotions can arise that are combinations of two or more basic emotions. For example, jealousy can be interpreted as a mixture of anger and sadness. Developers can map specific facial animations or gestures to each emotional state and can utilise animation blending to represent more complex emotions. Overall, emotional states allow NPCs to possess the emotion believability quality by allowing NPCs to exhibit a variety of expressions that are associated with different emotional states.

We categorised motivational states as mental states that serve a functional role in allowing NPCs to engage in goal-driven behaviour, which is crucial for achieving the self-motivation quality. We define two sub-categories of motivational states: goals and desires. Each motivational state contains an intensity that determines how strongly that goal influences the selection of behaviours.

Relational states were categorised as mental states that describe an NPC's relationship with other objects in the game world. We define two sub-categories of relational states based on the work of Buber [10] who identifies two distinct types of relationships: "I-Thou" relations and "I-It" relations. "I-Thou" relations are *social* and represent more meaningful interactions between multiple NPCs, whereas "I-It" relations are *objective* and describe an NPC's relationships with inanimate objects in the environment. "I-Thou" and "I-It" relationships are represented in the MSM architecture through social and objective relational states respectively.

### 3.4 Behaviour Component

The behaviour component of the MSM architecture allows NPCs to produce behavioural outputs based on perceived environmental stimuli and mental state interactions. For the MSM implementation described in Section 4, behavioural outputs are produced using a Behaviour Tree Generation (BTG) algorithm[12]. The BTG algorithm combines the AI techniques of behaviour trees and AI planning by allowing behaviours trees to be automatically generated for different mental states through a backward chaining algorithm [12].

The BTG algorithm is depicted in Figure 4. First, a utility selector node is added to the root node of an NPC's behaviour tree. This node selects and executes the child node with the highest utility value. After the utility selector node is added, each of the NPC's mental states are assigned a utility value and then added as children to this node. Finally, to generate a sequence of low-level actions for each mental state, the mental state is treated as a goal condition and a planner is used to select actions that are available to the NPC which could satisfy this goal. If necessary, new mental states and action sequences can be added to the NPCs' behaviour trees through re-planning.

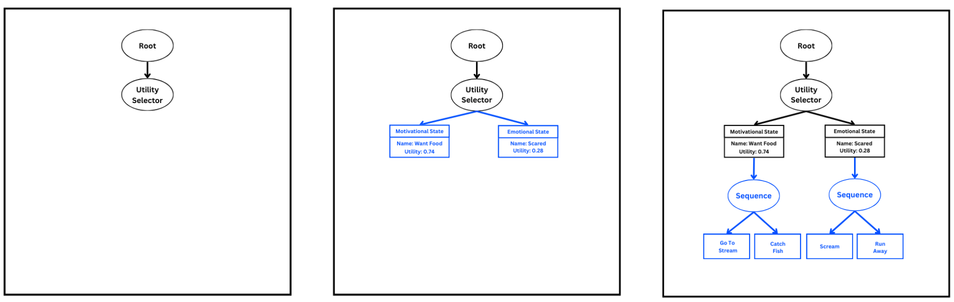



Fig. 4. A diagram that illustrates the utility selector node, goal mental states and action sequences being added to an NPC's behaviour tree.

## 4 Experimental Study

We developed an experimental study and evaluation criteria that adapts two existing approaches for evaluating believability proposed by Gorman et al. [19] and Bogdanovych et al. [8] respectively. Gorman et al. [19] tested overall perceived believability through a series of experiments where participants were shown two videos side-by-side featuring a group of virtual agents. One video depicted virtual agents that incorporated a specific believability quality in their behaviour, while the other video portrayed the same group of agents that had this quality disabled. Bogdanovych et al. [8] expanded on this evaluation method by adding an interactive component that required participants to communicate with two groups of virtual agents situated in an immersive environment. This was done to test the believable quality of awareness. Participants were then asked to rate how closely the behaviour of the agents in each video resembled human-like behaviour using the following Likert-type scale: 1:Definitely Human; 2: Probably Human; 3:Not Sure; 4: Probably Artificial; 5: Definitely Artificial.

Our modified evaluation method compared the perceived believability of a group of virtual agents that were designed using the MSM architecture to portray each believability quality to that of a group of agents that used an alternative FSM implementation (a commonly used technique in the games industry that is not designed for believability [30]). Furthermore, rather than testing each believability quality in isolation, we asked participants to evaluate all believability qualities being assessed (personality, emotion, self-motivation, change, social interaction, and situatedness) simultaneously. This was done to evaluate how well the MSM architecture was able to integrate multiple believability qualities to improve overall perceived believability.

### 4.1 Design and Setup

We conducted an experiment with eighteen participants using a within-subjects design (all participants played both versions of a game in a random order). The experiment consisted of an initial briefing, playing two versions of a game, and a survey. The total time for the experiment was thirty minutes. Participants were first briefed on the experiment and game, before playing the game through the online game-hosting platform itch.io. After playing each version of the game, participants completed an online survey in which they answered questions about that game version's AI behaviour. The experiment protocol was approved by the university's Human Ethics Committee.

The two versions of the game were developed to create NPCs that either used the MSM architecture or an alternative FSM implementation. Apart from the difference in AI systems, the only other change to the game conditions was the addition of game messages that provided the player with information on new psychological effects that NPCs were experiencing when using the MSM architecture. These additional messages were provided to communicate the internal mental state changes of the NPCs to the player. All other game conditions remained the same between both versions. To minimise order effects in our within-subjects design, participants were randomly assigned to begin either version of the game using an online pseudo-random-number generator.

After each version of the game, participants were sent a link to a survey which contained a series of questions on the believability of the NPCs in that version. Survey responses were collected anonymously and data was aggregated. Each survey asked participants to rate their agreement with six Likert statements, with responses that ranged from "Never" to "Always". These statements were derived from the believability qualities outlined in the literature, with each statement corresponding to a specific quality. The Likert statements are described in Table 3.

Table 3. Believability questionnaire statements.

| **Relevant Quality** | **Likert statement** |
|---|---|
| *Emotion* | Survivors expressed different emotions based on what was happening in their environment. |
| *Self-Motivation* | Survivors had their own goals and motivations. |
| *Change* | The behaviour of the survivors changed throughout the game. |
| *Social Interaction* | Survivors interacted with one another. |
| *Situatedness* | Each survivor reacted to what was happening around them. |
| *Personality* | The reactions of each survivor were consistent with their personality. |

### 4.2 Game Implementation

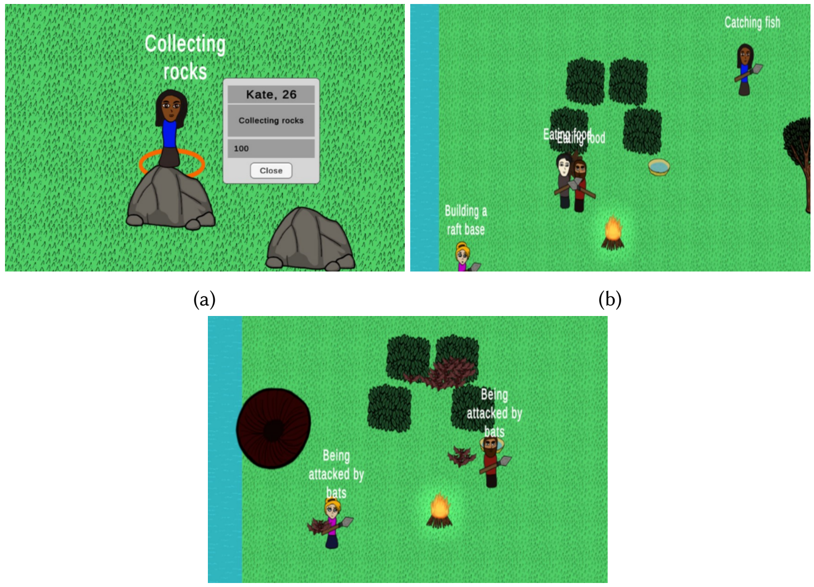
To test the believability of NPCs made with the MSM architecture, we developed two versions of an experimental game called "Occult Island" using the Unity game engine [21]. Occult Island is a simulation game where the player assigns survival tasks to a group of NPCs that have been ship-wrecked on an island, and then sequentially selects supernatural events to trigger on the island each day. After seven in-game days, a final boss automatically spawns on the island. If the survivors have completed all the necessary survival tasks before this happens, they will be able to escape from the island and the player wins. Otherwise, all the survivors are eaten by the boss and the player loses. There has been substantial research that claims extreme and unusual environments can have significant psychological impacts on individuals in the real world as they attempt to cope with their unpredictable circumstances [38]. Therefore, the goal of our game was to assess whether simulating the effects of extreme and unusual environmental changes on the NPCs' mental states in the MSM version of the game could improve perceived believability by players.

*4.2.1 FSM Version.* We implemented two alternative systems for NPC behaviour for comparison: the MSM architecture and a Finite-State-Machine (FSM) system. In the FSM version, none of the survivors possessed mental states. Instead, the behaviour of each survivor was dictated by three FSMs which contained behaviour states that the survivors' transitioned between when specific conditions were met. The three FSMs were: task, needs, and event. Survivors transitioned between different states within an FSM depending on the circumstances of the game world. The task FSM (Figure 5a) contained states for completing one of four survival tasks: finding food, collecting water, building a shelter, and building a raft. Each of the four survivors was assigned a specific survival task by the player at the beginning of the game.

To reflect the survival nature of the game, the needs FSM (Figure 5b) contained states for satisfying the survivors' two physiological needs: eating food and drinking water. When a survivor's hunger or thirst fell below a certain threshold, that survivor would transition to the appropriate state in the needs FSM. Once their need was satisfied, the survivor would then return to their designated survival task.

Finally, the event FSM (Figure 5c) contained various states for reacting to the different supernatural events that could be triggered on the island by the player, such as the appearance of a swarm of vampire bats. The purpose of the event FSM was to provide NPCs with actions that they could perform to resolve the game's events and progress to the next day. When the player triggered an event on the island, all survivors would transition to a particular state in the event FSM. Once the event was completed, the survivors would return to the state they were in prior to the commencement of the event.

*4.2.2 MSM Version.* In the MSM version of the game, the sensory, mental state and behaviour components of the MSM architecture were implemented as Unity packages. Each survivor began the game with a unique set of sensations, perceptual, emotional, motivational, epistemic, and relational states. The sensory component of the MSM architecture equipped each survivor with a sensory system that could detect nearby objects and sounds in their environment through raycasting and then activate perceptual states and epistemic states that stored environmental information. For example, a survivor standing within a certain distance of another survivor yelling had an active "hearing yell" perceptual state. Figure 6a illustrates the process of multiple survivors perceiving a nearby survivor yelling and reacting to that sound. Survivors also had two sensations: hunger and thirst which corresponded to two desire motivational states of wanting food and wanting water respectively. When either of these desires reached a specific intensity value, the survivors carried out actions to satisfy that desire.




(a) (b)

(c)

Fig. 5. Survivors collecting rocks (a), eating food (b) and reacting to a swarm of bats (c) in the FSM implementation.

Events that were triggered on the island increased the intensities of the survivors' emotional states and led to different emotional responses. Emotional responses were depicted through a variety of facial expressions and behaviours that were each mapped to one type of emotional state (e.g., sadness, fear, anger). For example, a survivor that first encountered a supernatural creature on the island experienced a "scared" emotional state, adopted a scared facial expression and would momentarily freeze in their current position. Compared to the FSM version where survivors would immediately return to their previous state after an event was resolved, survivors in the MSM version experienced lingering psychological effects as a result of the events. For example, certain events like the zombie invasion left the survivors with feelings of shock and anxiety long after they had concluded.

To give the survivors unique personalities, we allowed each of them to have distinct emotional responses to the same event. The duration of emotional responses were also tailored to reflect different personalities. Figure 6b shows the various unique emotional states that survivors experienced as a result of the zombie invasion event. The relational states for each survivor were used to signify their relationships with the other survivors and changed based on game events. Figure 6c depicts a change in relationship between two survivors. Ultimately, we theorised that the use of mental states to produce dynamic characteristics and behaviours for each survivor would improve the player's perceived believability of the NPCs as survivors that were stranded on a deserted island.

### 4.3 Participants

There were twenty participants who started the experiment, but only eighteen completed it. Two of the twenty participants did not complete the survey and their data was excluded. All participants were students between the ages of eighteen and twenty-five.

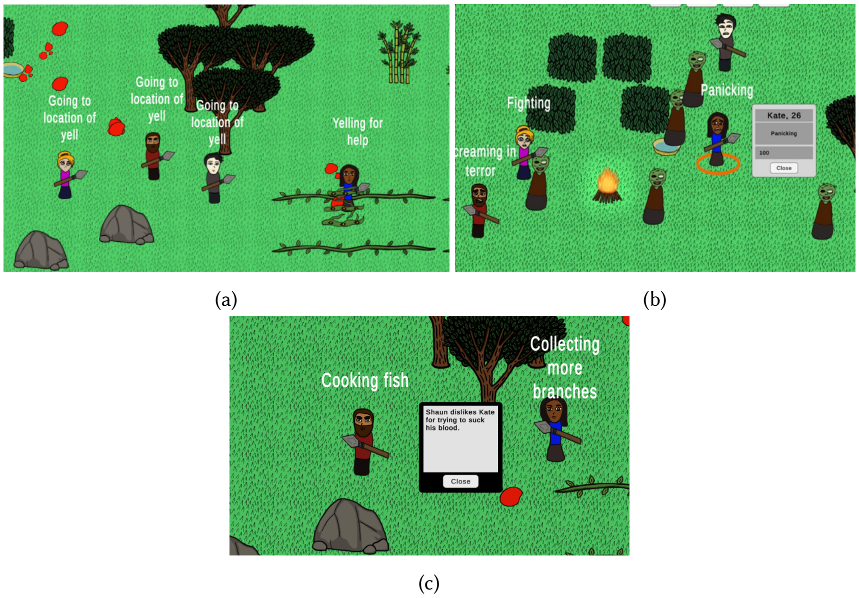



Fig. 6. Survivors reacting to a nearby yelling sound (a), displaying unique emotional reactions to an event (b) and experiencing changes in their relationships (c) in the MSM implementation.

### 4.4 Hypothesis Testing

We developed the following hypotheses to investigate our research question: **How can philosophy of mind theories be integrated into a believable agent architecture to develop NPCs that are perceived as more believable by players compared to NPCs that use a current game AI approach?**

- **H1:** Players will perceive the believability qualities of personality, emotion, self-motivation, change, social interaction, and situatedness more often in characters that use the MSM architecture compared to characters that use an FSM implementation.
- **H2:** Characters that use the MSM architecture will have a higher overall perceived believability than characters that use an FSM implementation.
- **H3:** Individual believability qualities in the MSM architecture will be positively correlated.

### 4.5 Believability Measures

To measure believability between the two NPC systems, we adapted the believability index equations used by Gorman et al. [19]. Each participant rated an NPC's ability to portray a specific believability quality on a scale of 1 (Never) to 5 (Always), and the degree to which an NPC using a given AI system was able to portray the believability quality in question was the normalised difference between the participant's rating and the value corresponding to 'Always':

$$h_p(NPC_i) = \frac{|r_p(NPC_i) - A|}{A - B}$$

Where $h_p(NPC_i)$ was the frequency to which participant $p$ judged the NPC to be portraying the believability quality, $r_p(NPC_i)$ was the participant's rating of NPC $i$, and $A$ and $B$ represented the values on the rating scale corresponding to 'Always' and 'Never' respectively. In other words, $h_p(NPC_i)$ will be 0 if the participant stated that the NPC never portrayed the believability quality, 1 if they always portrayed that quality, and a value in between for ratings in the middle. The believability index for any participant is:

$$b_n = 1 - \sum_{0<p<=n} h_p(NPC_i)$$

The believability index $B$ of the NPCs for believability quality $Q$, based on the rating given by $m$ participants and using AI behaviour system $S$ was:

$$B_S(Q) = \frac{\sum h_p(NPC_i)}{m}$$

Hence, testing H1 was equivalent to testing whether the following equation held true for each believability quality assessed in the experiment:

$$\Delta = |B_{MSM}(Q) > B_{FSM}(Q)|$$

Where $\Delta$ represented the contribution of that believability quality towards perceived believability [8]. Testing H2 was also reduced to computing the overall perceived believability score $OB$ by taking the mean of all believability index values in each AI system:

$$OB_S = \frac{\sum B_S(Q)}{m}$$

which was then compared to see if the overall score of the MSM architecture was greater than the overall score of the FSM system:

$$OB_{MSM} > OB_{FSM}$$

## 5 Results

We used one-tailed paired t-tests to test our three hypotheses. For H1, we tested the difference between the FSM believability index $B_{FSM}$ and the MSM believability index $B_{MSM}$ for each believability quality (see Table 4). The only quality in which the difference between $B_{FSM}$ and $B_{MSM}$ was not statistically significant was self-motivation. For H2, the overall perceived believability of the NPCs for each AI system was calculated by taking the mean of all believability quality index values for that system. The overall perceived believability for the FSM system was 0.52 ($SD$ = 0.15), while the overall perceived believability of the MSM architecture was 0.80 ($SD$ = 0.05). The difference between these values was also statistically significant ($t(5) = 4.06$; $p = .005$). For H3, one-tailed paired t-tests showed that all believability qualities were positively correlated (Table 5).

## 6 Discussion

Our research question for this study was: **How can philosophy of mind theories be integrated into a believable agent architecture to develop NPCs that are perceived as more believable by players compared to NPCs that use a current game AI approach?**. To address this question, we drew inspiration from the philosophy of mind literature to develop a believable agent architecture called Mental State Modelling (MSM) which allows developers to design computational models of the mind for NPCs. An experimental evaluation of the MSM architecture was carried out to test three hypotheses:

Table 4. Believability index values and t-test results for Emotion (E), Social Interaction (S), Change (C), Personality (P), Situatedness (S) and Self Motivation (SM).

| Believability Quality | $B_{FSM}$ | $B_{MSM}$ | P-value | t-value |
|---|---|---|---|---|
| E | 0.38 | 0.85 | <.001 | 5.24 |
| SI | 0.36 | 0.78 | < .001 | 4.86 |
| C | 0.42 | 0.76 | < .001 | 4.57 |
| P | 0.49 | 0.75 | .001 | 3.56 |
| S | 0.69 | 0.89 | .003 | 3.11 |
| SM | 0.76 | 0.76 | .5 | 0 |

Table 5. Pearson correlation results. Correlations with an asterisk next to them were statistically significant.

| | P | E | SM | C | SI | S |
|---|---|---|---|---|---|---|
| P | | 0.21* | 0.36* | 0.13 | 0.06 | 0.24* |
| E | 0.21* | | 0.39* | 0.27* | 0.38* | 0.42* |
| SM | 0.36* | 0.39* | | 0.44* | 0.05 | 0.05 |
| C | 0.13 | 0.27* | 0.44* | | 0.25* | 0.05 |
| SI | 0.06 | 0.38* | 0.05 | 0.25* | | 0.23* |
| S | 0.24* | 0.42* | 0.05 | 0.05 | 0.23* | |

- **H1:** Players will perceive the believability qualities of personality, emotion, self-motivation, change, social interaction, and situatedness more often in characters that use the MSM architecture compared to characters that use an FSM implementation.
- **H2:** Characters that use the MSM architecture will have a higher overall perceived believability than characters that use an FSM implementation.
- **H3:** Individual believability qualities in the MSM architecture will be positively correlated.

With regard to H1, our findings indicate that the MSM architecture was better at portraying the believability qualities of personality, emotion, change, social interaction, and situatedness compared to the FSM implementation, but was not better at portraying the self-motivation quality. Self-motivation has been identified as an important quality of believable agents [9, 27, 29] and Bragt [9] theorised that self-motivation could be the basis for other qualities such as emotion and personality to emerge. However, when Bogdanovych et al. [8] compared two groups of virtual agents that either did, or did not demonstrate self-motivation, they found that self-motivation did not contribute to an improved sense of character believability. Our findings are consistent with those presented by Bogdanovych et al. [8]. We found that self-motivation might not contribute to a player's perceived believability of an NPC. Our research provides new evidence that suggests that self-motivation might not be the central quality of believability of NPC behaviour in a game. In relation to H2, despite the absence of self-motivation as a contributing factor of believability, our results indicate that NPCs that exhibit behaviour using the MSM architecture have a higher overall perceived believability compared to NPCs that use an FSM implementation for behaviour. Finally, for H3, the results of the one-tailed paired t-tests presented in Section 5 indicate that all believability qualities were positively correlated.

Researchers have highlighted cumbersome development work and overly complex agent architectures as key issues that have prevented the believable agent architectures presented by the research community from being adopted in the games industry [15, 44]. Additionally, it has been noted that most agent architectures lack proper evaluation procedures, and the benefits of adopting these architectures over current industry-standard AI approaches have not been clearly demonstrated [23]. We have presented an approach for implementing believability qualities in NPCs by specifying mental state interactions that can lead to emergent behaviour. Our categorisation of mental states describes a broad group of mental state categories that can interact to produce each of the believability qualities outlined in the literature. Furthermore, the MSM architecture is flexible enough to account for more complex psychological effects that are not accounted for in traditional believable agent architectures such as hallucination through altered perceptual states. Our study provides empirical data that suggests NPCs created using the MSM believable agent architecture are considered to be more believable overall than NPCs that use the popular FSM game AI technique.

### 6.1 Implications

Based on our findings, we propose that the MSM architecture can be applied to the development of believable NPCs for any game that aims to create immersive experiences for players. Adopting the MSM architecture in the design and implementation of NPC behaviour can allow developers to populate their game worlds with NPCs that exhibit emergent believable behaviours. For example, NPCs could express dynamic emotional reactions to surprising choices that the player makes, a feature that is notably missing from RPGs like Starfield [35].

Compared to common game AI solutions like FSMs and behaviour trees, game designers that follow the MSM architecture do not need to pre-script behaviours for their NPCs. Instead, designers can adopt a more systemic approach to designing NPCs by specifying rules for how different mental states interact and produce behavioural responses in different situations, and then observing what behaviours emerge based on these rules. Furthermore, the MSM architecture can serve as a more holistic AI framework that is not tied to any specific game series or technology and includes mental properties like sensations, knowledge and beliefs which are rarely incorporated in other systems.

### 6.2 Limitations and Future Work

Despite the insights provided by our study, there are some limitations that warrant future work. In terms of evaluation, there is still no clear consensus in the literature on what aspects of NPC behaviour contribute to believability [25]. Future work could involve assessing the MSM architecture using alternative metrics and methodologies for evaluating believability. In addition to this, the MSM architecture was only evaluated by eighteen participants. To improve the generalisability of the claims made in this study, and prove that this architecture is suitable for widespread adoption in the games industry, further evaluation is required on a larger sample size.

Each of the system components of the MSM architecture could be developed further to facilitate more complex emergent behaviour. The sensory component could be expanded through the addition of new sensory data types such as smell. The current implementation of the mental state component is also heavily reliant on having pre-designed interaction rules between mental states. Future work could involve exploring how to use logical inference to dynamically generate new mental state transition rules as a game progresses.

Another future research direction could involve testing alternative methods for mapping mental states to behaviours, specifically in relation to the emerging field of generative AI. For example, Large Language Models have been used in combination with other cognitive architectures similar to the MSM architecture to generate natural, human-like dialogue responses for virtual characters based on cognitive processes [39].

## 7 Conclusions

In this paper, we presented a new believable agent architecture, called Mental State Modelling, that facilitates emergent believable behaviour. The MSM architecture was developed through an interdisciplinary approach that applies functionalism and other philosophical theories of mind to the modelling of a complex computational mind system. The architecture was then used to create NPCs that portrayed 5 of the 6 believability qualities that were tested and had a higher overall perceived believability compared to NPCs developed using a common FSM approach. The focus on emergent behaviour means that the architecture is not overly convoluted or difficult to use, which is one criticism of existing believable agent architectures. Our research serves as important preliminary work on incorporating functionalism, complex systems modelling, and emergent behaviour into the development of believable NPCs and indicates a promising direction for future work in this area.

## References


[1] [n. d.]. Convai - Conversational AI for Virtual Worlds — convai.com. https://convai.com/. [Accessed 27-09-2024].

[2] [n. d.]. Inworld: AI-powered gameplay — inworld.ai. https://inworld.ai/. [Accessed 27-09-2024].

[3] Jan Allbeck and Norman Badler. 2002. Toward representing agent behaviors modified by personality and emotion. *Embodied conversational agents at AAMAS* 2, 6 (2002).

[4] Henry E Allison. 2006. Kant's transcendental idealism. *A companion to Kant* (2006), 111–124.

[5] Ruth Aylett, Christopher Ritter, Mei Yii Lim, Frank Broz, Peter E McKenna, Ingo Keller, and Gnanathusharan Rajendran. 2019. An architecture for emotional facial expressions as social signals. *IEEE Transactions on Affective Computing* 12, 2 (2019), 293–305.

[6] Pablo Barbosa, Danielle Silva, Geber Ramalho, and Patricia Tedesco. 2010. Simulating the emergence of social relationship networks in groups of believable agents: The X-BARIM model. In *Brazilian Symposium on Artificial Intelligence*. Springer, 133–142.

[7] Christoph Bartneck. 2002. Integrating the occ model of emotions in embodied characters. (2002).

[8] Anton Bogdanovych, Tomas Trescak, and Simeon Simoff. 2016. What makes virtual agents believable? *Connection Science* 28, 1 (2016), 83–108.

[9] Jasper Bragt. 2010. *Towards believable characters in the virtual storyteller.* Master's thesis. University of Twente.

[10] Martin Buber. 1970. *I and Thou.* Scribner, New York,.

[11] Stephen Burwood. 2004. Philosophy of mind. In *Fundamentals of Philosophy*. Routledge, 234–261.

[12] Michele Colledanchise, Diogo Almeida, and Petter Ögren. 2019. Towards blended reactive planning and acting using behavior trees. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 8839–8845.

[13] Tiago Doce, Joao Dias, Rui Prada, and Ana Paiva. 2010. Creating individual agents through personality traits. In *International Conference on Intelligent Virtual Agents*. Springer, 257–264.

[14] Paul Ekman. 1999. Basic emotions. *Handbook of cognition and emotion* 98, 45-60 (1999), 16.

[15] Salma ElSayed and David J King. 2017. Affect and believability in game characters: A review of the use of affective computing in games. In *GAME-ON'2017, 18th annual Conference on Simulation and AI in Computer Games*. EUROSIS, 90–97.

[16] Caroline Faur, Céline Clavel, Sylvie Pesty, and Jean-Claude Martin. 2013. PERSEED: a self-based model of personality for virtual agents inspired by socio-cognitive theories. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 467–472.

[17] Razvan V Florian. 2003. Autonomous artificial intelligent agents. *Center for Cognitive and Neural Studies (Coneural), Cluj-Napoca, Romania* (2003).

[18] Peter Gärdenfors. 1988. *Knowledge in flux: Modeling the dynamics of epistemic states.* The MIT press.

[19] Bernard Gorman, Christian Thurau, Christian Bauckhage, and Mark Humphrys. 2006. Believability testing and bayesian imitation in interactive computer games. In *From Animals to Animats 9: 9th International Conference on Simulation of Adaptive Behavior, SAB 2006, Rome, Italy, September 25-29, 2006. Proceedings 9*. Springer, 655–666.

[20] Jonathan Gratch and Stacy Marsella. 2005. Evaluating a computational model of emotion. *Autonomous Agents and Multi-Agent Systems* 11 (2005), 23–43.

[21] John K Haas. 2014. A history of the unity game engine. (2014).

[22] Devang Jagdale. 2021. Finite state machine in game development. *algorithms* 10, 1 (2021).

[23] Hanneke Kersjes and Pieter Spronck. 2016. Modeling believable game characters. In *2016 IEEE Conference on Computational Intelligence and Games (CIG)*. IEEE, 1–8.

[24] Ji-Hwan Kim, Nguyen Duc Thang, and Tae-Seong Kim. 2009. 3-D hand motion tracking and gesture recognition using a data glove. In *2009 IEEE international symposium on industrial electronics*. IEEE, 1013–1018.

[25] Michael Sangyeob Lee and Carrie Heeter. 2012. What do you mean by believable characters?: The effect of character rating and hostility on the perception of character believability. *Journal of Gaming & Virtual Worlds* 4, 1 (2012), 81–97.

[26] John Locke. 1847. *An essay concerning human understanding.* Kay & Troutman.

[27] Aaron B Loyall. 1997. *Believable Agents: Building Interactive Personalities.* Technical Report. CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF COMPUTER SCIENCE.

[28] Ryan Marcotte and Howard J Hamilton. 2017. Behavior trees for modelling artificial intelligence in games: A tutorial. *The Computer Games Journal* 6 (2017), 171–184.

[29] Michael Mateas. 1997. An Oz-centric review of interactive drama and believable agents. In *Artificial intelligence today*. Springer, 297–328.

[30] Jere Miles and Rahman Tashakkori. 2010. Improving believability of simulated characters. *Journal of Computing Sciences in Colleges* 25, 3 (2010), 32–39.

[31] Andrew Ortony, Gerald L Clore, and Allan Collins. 1988. The Cognitive structure of emotions cambridge. *UK: Cambridge University Press9* (1988).

[32] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*. 1–22.

[33] Hilary Putnam. 1967. The nature of mental states. *Art, mind, and religion* (1967), 37–48.

[34] Stephen J Read, Brian M Monroe, Aaron L Brownstein, Yu Yang, Gurveen Chopra, and Lynn C Miller. 2010. A neural network model of the structure and dynamics of human personality. *Psychological review* 117, 1 (2010), 61.

[35] Nick Rodriguez. 2023. Starfield's npcs don't measure up to Baldur's Gate 3's. https://gamerant.com/starfield-baldurs-gate-3-npc-animations-compared-bg3/

[36] Michael Smith. 1987. *The Humean theory of motivation.* Oxford University Press.

[37] Stephen Soldz and George E Vaillant. 1999. The Big Five personality traits and the life course: A 45-year longitudinal study. *Journal of research in personality* 33, 2 (1999), 208–232.

[38] Peter Suedfeld. 2012. Extreme and Unusual Environments: Challenges and Responses. https://api.semanticscholar.org/CorpusID:148901326

[39] Theodore Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L Griffiths. 2023. Cognitive architectures for language agents. *arXiv preprint arXiv:2309.02427* (2023).

[40] Penny Sweetser. 2008. *Emergence in Games.* Charles River Media.

[41] Penny Sweetser and Matthew Aitchison. 2020. Do Game Bots Dream of Electric Rewards? The universality of intrinsic motivation. In *Proceedings of the 15th International Conference on the Foundations of Digital Games*. 1–7.

[42] Penelope Sweetser and Janet Wiles. 2005. Combining influence maps and cellular automata for reactive game agents. In *International Conference on Intelligent Data Engineering and Automated Learning*. Springer, 524–531.

[43] Alan M Turing and J Haugeland. 1950. Computing machinery and intelligence. *The Turing Test: Verbal Behavior as the Hallmark of Intelligence* (1950), 29–56.

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