The Author's Journey—Understanding and Improving the Authoring Process of Theory-Driven Socially Intelligent Agents

MANUEL GUIMARÃES, JOANA CAMPOS, and PEDRO A. SANTOS, Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal and INESC-ID, Lisboa, Portugal
JOÃO DIAS, CISCA, Faculty of Science and Technology, University of Algarve, Faro, Portugal and INESC-ID, Lisboa, Portugal
RUI PRADA, Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal and INESC-ID, Lisboa, Portugal

State-of-the-art agent-modelling tools support the creation of powerful Socially Intelligent Agents (SIAs) capable of engaging in social interactions with participants in various roles and environments. However, their deployment demands a labourious authoring task as it is necessary to manually define behaviour rules and create content for different interaction scenarios.

While Socially Intelligent Agents (SIAs) research has centred on the user experience, we shift focus to the authors. To understand the challenges faced by authors who create these agents, we performed an innovative analysis of the authoring experience in modern agent modelling tools. One key finding is that, while SIA concepts are generally understandable, emotional-based concepts are not as easily comprehended or used by authors. We propose a hybrid solution approach that culminated in the development of Authoring-Assisted FAtiMA-Toolkit. The augmented agent modelling tool incorporates a data-driven Authoring Assistant to boost author productivity while promoting transparency and authorial control. To evaluate the impact of this framework on the authoring experience, we conducted a user study. Results showed that authors using the Authoring-Assisted FAtiMA-Toolkit were on average able to create more SIA-related content in less time.

Our findings suggest that data-augmented, theory-grounded agent modelling tools can support the development of affective social agents by reducing the authoring burden without sacrificing the framework's clarity or the authors' control over the content.

CCS Concepts: • **Computing methodologies** → **Intelligent agents**;

Additional Key Words and Phrases: intelligent agents, affective computing, cognitive architecture, emotions, social robots

This work received Portuguese national funds from FCT - Foundation for Science and Technology through project: 10.54499/UIDB/50021/2020 and UIDB/04326/2020, UIDP/04326/2020, LA/P/0101/2020 and SLICE PTDC/CCI-COM/30787/2017.

Authors' addresses: Manuel Guimarães (corresponding author), Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal and INESC-ID, Lisboa, Portugal; e-mail: manuel.m.guimaraes@tecnico.ulisboa.pt; Joana Campos, Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal and INESC-ID, Lisboa, Portugal; e-mail: joana.campos@inesc-id.pt; Pedro A. Santos, Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal and INESC-ID, Lisboa, Portugal; e-mail: pedro.santos@tecnico.ulisboa.pt; João Dias, CISCA, Faculty of Science and Technology, University of Algarve, Faro, Portugal and INESC-ID, Lisboa, Portugal; e-mail: jmdias@ualg.pt; Rui Prada, Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal and INESC-ID, Lisboa, Portugal; e-mail: rui.prada@tecnico.ulisboa.pt.

ACM 2160-6463/2025/4-ART8
https://doi.org/10.1145/3711672

ACM Reference format:
Manuel Guimarães, Joana Campos, Pedro A. Santos, João Dias, and Rui Prada. 2025. The Author's Journey—Understanding and Improving the Authoring Process of Theory-Driven Socially Intelligent Agents. *ACM Trans. Interact. Intell. Syst.* 15, 2, Article 8 (April 2025), 40 pages.
https://doi.org/10.1145/3711672

1 Introduction

Socially Intelligent Agents (SIAs) are computer-generated characters capable of engaging in social interactions with participants in a wide range of roles and environments. For instance, researchers have been successful in creating interactive scenarios where children diagnosed with Autism Spectrum Conditions engage in controlled social interactions [8], for individuals diagnosed with high anxiety levels to practice job interviews [29] and even as a first-line diagnostic system for mental health support in universities [4].

To create human–agent interaction scenarios designers typically make use of agent modelling tools; computer programs, typically user interfaces, designed to facilitate the deployment of SIAs in a wide array of applications.

Recent generations of social agent frameworks have crossed the "academia barrier" and are now being used by an increasing number of authors unfamiliar with the affective agent field [40, 65]. While challenging, creating an interactive experience can be manageable in narrow domains of application, but for a serious game or social skills training content designer, using an intelligent agent framework can quickly become a burdensome task. It is up to the author of a scenario to manually create and describe how different concepts, such as goals, beliefs and actions, interact and guarantee character consistency as events unfold.

If we look at the design of human–agent interactions from a relationship point of view, it is clear that there are two key actors: the author, the one that wants to create the experience and the end-user, the one who will play, or experience it. Naturally, there is an inherent relationship between the two and the layers in between; Authors use agent modelling tools to create a scenario and end-users play this scenario through an application such as a computer screen, virtual or augmented reality glasses or even through their phones. Figure 1 illustrates this idea.

The focus of this work, however, is on the other end of the relationship, which we will refer to as the authoring experience. We believe that, to move towards the next generation of SIAs, it is necessary to rethink the design of the tools used to create them. Improving the authoring experience will subsequently lead to a higher level of human–agent interactions.

To clarify, authors are users of authoring tools, however, when we refer to the authoring experience, we refer the entire journey an author undertakes—from the initial idea, through design and development, to playtesting. The work presented here, emphasises this comprehensive journey, which we found to be typically overlooked area in the field of SIAs.

With this work we perform several different contributions to the field. First, we provide an analysis on the authoring process of theory-driven frameworks, in particular, its "Author's Journey." In general, despite the wide variety of different social physics and affective theories, most of these frameworks rely on similar concepts. Second, we present two studies that focus on understanding the authoring experience. The first evaluates how well authors understand the concepts used in the SIA field. The second study examines the effort required to create various human–agent interaction scenarios. Both studies involve both experienced and inexperienced authors and introduce innovative metrics to evaluate their key aspects. Our findings suggest issues with the comprehension of the theories traditional agent modelling tools are grounded in. Additionally, the burden placed on authors seems too heavy for inexperienced authors to handle.

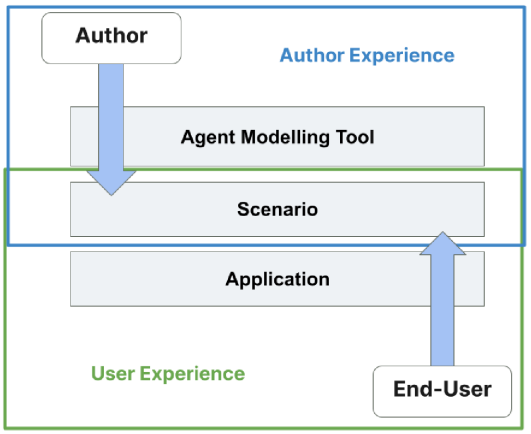

Fig. 1. Human–agent social interaction experiences have two distinct key actors: the author, the one that wants to create the experience and the end-user, the one who will play it, or experience it, typically referred to user experience in the SIA field.

Our third contribution is the development of a theory-grounded agent modelling tools that harnesses the power of data-driven approaches to facilitate its authoring experience and tackle the identified understandability issues. The framework is supported by an active Authoring-Assistant agent and by a tool that automatically generates SIA scenarios from stories. Finally, the manuscript concludes with a user study designed to assess the impact of the augmented agent modelling tool on the authoring experience. Here, we introduce a new metric to assess productivity over time when discussing authoring experience.

2 Theory-Driven Agent Modelling Tools

The design and rationale behind intelligent virtual agents represent decades of work across a wide array of disciplines, such as social sciences and human–agent interaction. Cutting-edge systems for creating realistic virtual agents can now register the gestures, body motion and gaze of an end-user and generate in real time both the verbal and non-verbal cues required to effectively communicate with the end-user, giving a startling impression of realism [52].

To create agents that were capable of emotional dynamics it was necessary to study existing work on affective behaviour and computerise it. These algorithmic interpretations of theories of emotion are typically called *computational models of emotions*. These models followed a top-down approach grounded on theoretical principles from the social sciences literature. For example, EMA is an appraisal-based computational model that follows Smith and Lazarus's cognitive-motivational-emotive psychological theory [35]. ALMA represented affective states that take into account three characteristics: emotions (short-term), moods (medium-term) and personality (long-term) [19]. ALMA implements the OCC model of emotions [51] combined with the Big Five model of personality. WASABI [7] followed the dimensional approach to emotions where it simulates the agent's emotional state as a point in the continuous 3D space using the PAD Model (Pleasure, Arousal and Dominance) [47]. Initially developed in 2005 to drive the behaviour of autonomous 3D characters in a serious game about bullying [15], **FearNot! Affective Mind Architecture (FAtiMA)** is an agent architecture with planning capabilities designed to use emotions and personality to influence the agent's behaviour based on the OCC model of emotions.

2.1 Social Agent Modelling Tools

In many ways, the first emotion-simulation frameworks can be seen as the first generation of agent modelling tools. Since the decision-making process was found to be heavily influenced by one's emotional state, researchers started using computational models of emotions as the primary motivation behind the agent's actions [6].

Several of these initial frameworks continued to be developed and grew into more complete social agent architectures. Some of these are still in use today, like FAtiMA-Toolkit, which was born out of FearNot! [5]. Through these, authors are able to define complex characters with beliefs, desires and goals, their action space, emotional reactions, dialogues, type of voice, and, in some, even specify the agent's embodiment. The new generation of agent modelling tools also attempted to address some of the complexity and accessibility issues identified by its peers in the hopes of a more widespread adoption [66].

Ensemble. [65] extends the "Comme il Faut" social physics engine [45], using network theory and predicate logic to forge emergent narratives [12]. Agents in Ensemble possess comprehensive social state knowledge, using decision-making processes rooted in sociocultural norms and social practices' structure. Unique features include the modelling of incomplete or failed actions, for example, action targets can accept or reject social practices like flirting or insulting. CiF-based emergent social physics mods have been implemented and shared in popular AAA games like Skyrim and Conan Exiles with studies indicating player preference for NPCs guided by these engines [26, 27, 48]

Virtual Human Toolkit. [29] is a well-known architecture designed to facilitate the creation of autonomous conversational characters. It was created to function as a general-purpose collection of integrated capabilities, including speech recognition, natural language processing, perception and non-verbal behaviour generation and execution [21] to make creating virtual humans easier and more accessible.

FAtiMA-Toolkit. is an open-source collection of tools and assets designed for the creation of characters (virtual or robotic) with social and emotional intelligence. As the name implies, the toolkit is the successor of FAtiMA [5] and FAtiMA-Modular [14]. Thus, it is influenced by its theory-grounded background and guided by the same principles of modularity and a character-centric approach.

Figure 2 shows a diagram of all the existing components within FAtiMA-Toolkit. Each is the result of continued iterations and experience accumulated while working with developers starting with the RAGE project[1] and afterwards with Slice Project.[2] Its official Web site[3] contains several different resources, including Starter Kits, Demonstrations and Tutorials. FAtiMA-Toolkit plays a central role in this work, an in-depth description of the toolkit was published in this journal [40]. In the following section, we will get to know more about FAtiMA and intelligent virtual agent concepts in general.

3 Creating an Interactive Social Agent Scenario Using Theory-Driven Architectures

Although the authoring process has been the focus of extensive work in various fields, such as interactive storytelling and intelligent virtual agents, in our research, we were unable to find works that provided an overview or comparison between different agent modelling tools from an author's point of view. Here we will provide a brief comparison between modern theory-driven social agent architectures while performing the various steps necessary to create a simple human–agent

[1] https://cordis.europa.eu/project/id/644187
[2] https://gaips.inesc-id.pt/slice/
[3] https://fatima-toolkit.eu/

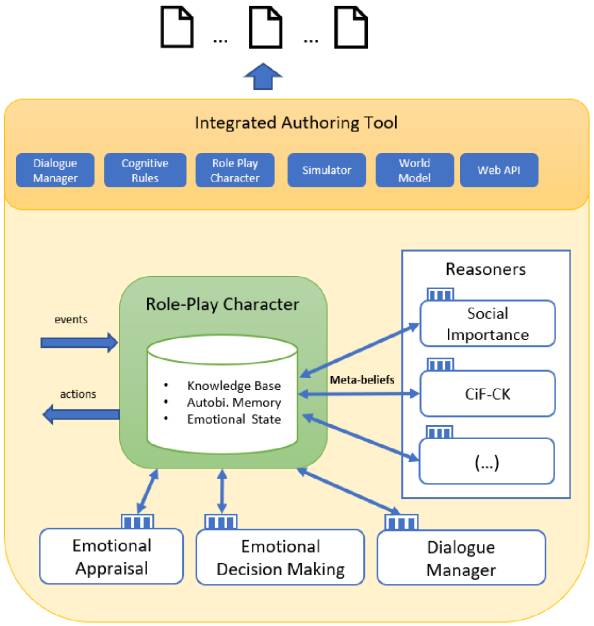

Fig. 2. FAtiMA Toolkit Components [40].

interaction scenario. The process and steps described here are based on our experience both as and working with authors of interactive experiences.

3.1 Scenario Planning

The planning stage, regardless of the field, objective or context, is characterised by the creation of some outline of the desired outcome. For example, in social learning environments, the interaction goals are typically designed to ensure participants are taught a particular topic or procedure, for instance, how to follow the standard police interview protocol [25].

A short story will be used as the desired scenario to present a fair but brief idea of the authoring process across different frameworks. Naturally, different architectures are focused on different SIA components. Thus, in each step, a description of how different state-of-the-art tools would implement those concepts will be provided. It is important to note that during this exercise, we are not focused on the exact implementation steps or details but on the overall idea.

> *John was hungry, so he went to a Restaurant. In the restaurant, John asked the waiter if they had hot dogs. The Waiter informed John that they did not serve hot dogs. John got upset and left.*

3.2 Knowledge Base

All intelligent agents have their own knowledge base. It is where information regarding the state of the world can be stored. Social intelligent agents store their beliefs about the surrounding world, including themselves and other agents. In general, the first step when implementing an agent-based experience is the development of the characters and their internal knowledge base. The concepts within it should be tied to the scenario's context. In this case, given the fact that the interaction will be played in a restaurant, it is logical for the agents to have the notions of "food" and "hunger."

In EMA [22, 23, 38], like most theory-driven frameworks, the state of the world is represented as a conjunction of propositions. Each belief corresponds to "commitments to the truth," and

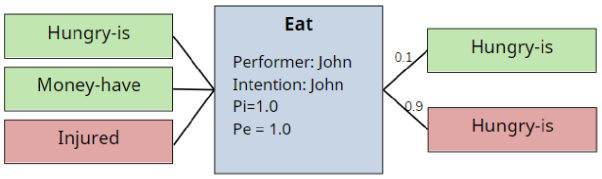

Fig. 3. Eat action representation in the EMA framework. Segments in light green represent propositions believed to be true while the light red represents propositions believed to be false.

their value is binary (although there is support for the use of probabilities): "Apple-Have—P = 1.0," "U-raised—P = 1.0," "Injured—P = 0.0" represent beliefs where the actor has an apple, its umbrella is raised, and it is uninjured. In our scenario, the representation of the initial state could be the following conjunction where the actor is Hungry, has Money and is uninjured:

Hungry-is ∧ Money-have ∧ -Injured

Comme il Faut, the predecessor of Ensemble, is a social agent architecture that places a heavy emphasis on the Social Exchanges created by authors. Thus, the representation of each character is relatively thin, consisting of a small amount of declarative information [41] that is socially relevant. The internal state of its characters contains their traits and status, which, in our case, could be:

```
trait(foodie, John), trait(stubborn, John), status(hungry, John)
```

In addition, the characters also represent their goals through their "Prospective memory," a vector of numeric volitions (desires) for characters to engage in specific Social Exchanges with specific characters. These are used as conditions that are then added up to an amount representing "how much" an agent wants to perform a particular action [43].

Overall, there is no rigid convention for how beliefs and goals are declared as long as they are consistent across the scenario and its framework. FAtiMA-Toolkit's beliefs can have any value, not just true or false:

```
Is(Hungry)=True
Has(Money)=5
Is(Injured) = Peter
```

3.3 Actions

Actions typically have both (pre)conditions and effects. Consequently, they tend to be represented as rules. If an action's conditions are verified, the agent might decide to perform that action. In addition to this, each action usually has a utility value associated with it that represents its importance or intention. The result of the execution of an action is described by its effects.

In EMA, actions are having a duration while their effects can occur asynchronously thus, at any point in time, several actions may be executed simultaneously, and several action effects may be anticipated [38]. Figure 3 displays an example of one way of representing the eat action using the EMA framework. The left side of the image represents its pre-conditions, while the effects and their probability are presented on the right. As shown, there is a chance John is hungry even after eating.

One final detail is that Pi represents the likelihood that an agent intends to execute an action while Pe represents the probability that the action can be executed [39]. Like EMA, other frameworks also model the possibility of failure or never being completely executed. For instance, the social practices of Comme il Faut, now known as Ensemble [44], allow for the target of the action to, for instance, reject an insult or accept a flirtatious invitation.

As mentioned in the previous section, the algorithm that Comme il Faut engines use to decide which action to perform is through the so-called influence rules [46]. Initiator influence rules determine a character's desire to initiate a Social Exchange with other characters. On the other hand, the responder's influence rules are used to determine whether a responder accepts or rejects the Social Exchange [41]. The following represent two initiator influence rules for a Greeting Social Exchange. If John is hungry and has the polite trait, it adds +3 to the intention of performing a Greeting however, if the relationship between John and the Waiter is lower than 0, it subtracts 3:

```
status(hungry, John) && trait(polite, John) -> +3
network(friend, John, Waiter) < 0 -> -3
```

Ensemble iterates over the architecture of Comme il Faut making Social Exchanges more complex and leading to Social practices. Social practices have a wide array of possible interaction steps; the social state is updated accordingly, and where agents evaluate influence rules and choose their response. For example, considering our scenario, "John" wants to perform a *Thank You* social practice towards the Waiter for bringing him food. John starts ranking each practice variation from the available practices list: "Sincere Thank," "Rude Thank," "Formal Thank" and so on…and then chooses the highest ranking one.

An action rule defined in FAtiMA-Toolkit [24] uses as conditions the value of beliefs present in the agent's knowledge base. The subsequent action rule captures the following phenomenon: "An Agent performs the Eat action towards [item] if the [item] is edible if the agent is hungry and if it has at least 1 [item] in their knowledge base":

```
Action: Eat
Target: [item]
Priority: 1
Conditions:
  Edible([item]) = True
  Is(Hungry) = True
  Has([item]) > 0
```

3.3.1 Consequences of Actions. Depending on the architecture, the result of each action may need to be explicitly authored. To maintain the immersiveness, the suspension of disbelief, during the interaction, authors tend to design action effects that are coherent with what happens in real life but also serve the set interaction goals.

Both EMA and Ensemble define the consequences of actions within the action itself, as shown in Figure 3, concerning EMA. Comme il Faut frameworks define the effects of actions as changes to the social state. When an action is taken, a performance occurs, and the social state is modified to reflect the agent's response via the action's effect. In the case of the scenario we are creating:

```
Social Exchange Name: Polite Thank You
Social Exchange Step: Polite Reply
Effects: network(trust, John, Waiter) -> +5

Social Exchange Name: Polite Thank You
Social Exchange Step: No Reply
Effects: network(trust, John, Waiter) -> -1
         status(angry_at, John, Waiter)
```

FAtiMA-Toolkit deals with the consequences of action through its world model. It the component that handles the effects for all different actions and, like most of its components, employs a rule-based approach [40]. For example, when an agent eats, its KB is updated so that it no longer believes it is hungry:

```
Event: Event(Action-End, [s], Eat, *)
Effects:
  Property Name : Is(Hungry)
  New Value : False
```

3.4 Emotions

The significance and vastness of the research on computers and emotion have led to the development of an entire scientific field: Affective Computing [55]. Its objective is to study how computers can recognise an end-user's emotional states, express their own emotions and respond to the end-user's emotions. Section 2 has presented some of the wide variety of works in terms of both affective theories and computational models of emotions [47, 51, 60].

GAMYGDALA [56] follows the OCC model of emotions [51] and automatically generates emotions based on the agent's goals. Designers only need to define which type of events are relevant to each goal. For example, in our case, we can consider that John has the goal of "Eat Hog-Dogs," which has a positive utility of 0.5. When John enters the restaurant, it is close to reaching its goal causing the generation of "Hope" in its affective state. When John is told that there are no hot dogs, the previously set goal is "disconfirmed" thus it generates feelings of "Disappointment." GAMYGALA functions as a black-box emotion generator, thus, it is only necessary to define the goals of the character and how events affect them:

```
EmoBrain brain = new EmoBrain();
brain.Goal("Eat Hog-Dogs", 0.7f);
brain.Belief("In a Restaurant", 1f, "John")
brain.AffectsGoal("Ordered Hog-Dogs", "Eat Hog-Dogs", 0.2f);
brain.AffectsGoal("No Hog-Dogs", "Eat Hog-Dogs", -1);
brain.Update();
```

Similar to GAMYGDALA, FAtiMA-Toolkit [40] uses an appraisal system based on the OCC model of emotions [51]. Authors can write rules that capture generic or specific actions but must also decide which appraisal variables are triggered. For example, in the following case, the act of receiving food from the agent "Waiter" triggers the "Desirability" appraisal variable which, according to the OCC model, will generate feelings of Joy:

```
Subject: Waiter
Action: Give([food])
Target: John
Appraisal Variables:
  Desirability = 5
```

ALMA's [19] affect computation also uses Appraisal Rules. In this case, there are four different types of rules: Basic Appraisal Rules, define how characters appraise events, actions and objects related to them; Act Appraisal Rules, how characters appraise their own acts and other characters' acts; Emotion Display Appraisal Rules, define how characters appraise their own and others emotion displays; and, finally, Mood Display Appraisal Rules, how characters appraise their own and others' mood displays. The framework also features a concept known as Appraisal Tags, as a symbolic

abbreviation that can be used to generalise events [20]. Finally, it should be noted that the framework uses its own XML-based affect modelling language:

```
<CharacterAffect name="John">
   <Appraisal>
      <Basic>
       <GoodEvent desirability="0.3"/>
       <BadEvent desirability="-0.3"/>
      </Basic>
      <DirectAct type="Polite Thank"performer="John">
       <GoodEvent desirability="0.5"/>
       <GoodOther agency="other" praiseworthiness="0.3"/>
      </DirectAct>
      <SelfEmotion emotion="ReproachDisplay">
       <BadEvent desirability="-0.3"/>
      </SelfEmotion>
   </Appraisal>
</CharacterAffect>
```

The above example details John's appraisal process by defining two different Base Appraisal rules containing GoodEvent and BadEvent Appraisal Tags. Whenever these occur, the agent will appraise them with the corresponding Desirability values. In addition to this, we have also defined an Act and an Emotional Display rule. The first pertains to when John performs a "Polite Thank" action, it will be appraised as a Good Event for itself, while for others, it will be seen as a Praiseworthy event. In the second rule, we are defining that when the agent detects Reproach in others, it will see it as a Bad Event. Finally, it is possible to easily associate dialogue with Appraisal Tags, as shown below:

```
John: Thank you so much [=good_event]
Waiter: Go away please [=bad_event]
```

3.5 Dialogue

At its core, social interaction is essentially communicating with someone or with something. The most common way of communicating with a SIA is through written or spoken dialogue. While in the past most of the dialogue was manually authored, and part of it still is, its importance to the human–computer interaction field has led to a wide range of different conversational systems, scripting tools and automatically generated content. [37].

Comme il Faut's instantiation of Social Exchanges consists of lines of dialogue, represented using natural language templates used by the participating characters during their exchange. The dialogue lines use a Tag system to facilitate the reutilisation and generalisation of the dialogues (through the use of %). In our Restaurant scenario, we could use the following dialogue as part of the Order Food exchange:

```
Initiator: Hello %rude%, I am ready to order
Responder: What would you liketo order  %pronoun(sir , ma'am)% ?
Initiator: Do you have HotDogs?
Responder: Nope %rude% !
Initiator : This is outrageuous, I'm leaving %angry%
```

The next generation's framework, Ensemble [65], has been supported by the work of Aljammaz et al. [1], which has integrated chatbots and a knowledge model of representation to aid the creation

Table 1. Example of Different Dialogue Actions

Current State	Next State	Meaning	Style	Utterance
Start	Greeting	-	Polite	Good Afternoon
Start	Greeting	-	Rude	Hey!
Greeting	Order	-	Polite	How can I help you?
Greeting	Order	-	Polite	How do you do?
Greeting	Leave	-	Very rude	Not you again...
Order	OrderResponse	HotDog	-	I'd like a hot-dog please
Order	OrderResponse	Pizza	-	I would like a pizza please
OrderResponse	Leave	-	-	We have no more hot-dogs
Leave	End	-	Rude	I'm outta here

and management of dialogue. The chatbots are used as a layer of abstraction between the human interactor and the characters allowing open-text input to be taken into account by the scenario. For example, the player could interact with the agent by typing "I'd like to order food." The objective of the chatbot system is to recognise and translate this line of dialogue to a particular social practice and its state. Ideally, in this case, it would be able to map it into the "Order Food" social practice.

The solution adopted in FAtiMA-Toolkit is to have an explicit dialogue tree structure without any logic. Instead, it merely informs agents of how many options they can choose from in a given dialogue state. The logic of the dialogue is handled by a specific action called: "Speak." Through it, the author can control the current and the next state of the conversation along with additional optional parameters. Just as Ensemble and CiF [45], FAtiMA-Toolkit also uses a sort of tagging system to facilitate the management of its dialogue. In this case, the following action rule defines that agents that are in a lower mood choose to Speak actions with the "Rude" tag. The dialogue state and the next state conditions allow the dialogue system to replace the "currentState" and the "nextState" variables to be replaced by values in the agent's knownledge base:

```
Action: Speak([currentState],[nextState], - ,Rude)
Target: John
Priority: 2
Conditions:
  DialogueState(John) = [currentState]
  NextState(John) = [nextState]
  Mood(SELF) < 0
```

The conditions of each action are evaluated against the beliefs of each agent. Thus, agents can use their beliefs to keep track of the state of the dialogue, both the current dialogue state and the next dialogue state. The dialogue manager will then search for a dialogue in the dialogue pool with the corresponding attributes. Table 1 represents dialogues we have created for our Restaurant scenario. In this case, the "style" flag is used to define the politeness level of the dialogue.

3.6 Agent Modelling Is an Iterative Process

The process of creating a SIA interaction scenario is fundamentally iterative, necessitating frequent interchange between distinct "authoring phases." Integral to this process is regularly adopting a "player" perspective to evaluate the scenario's adherence to designed steps, realism of dialogue and coherence of emotion generation. Swift prototyping and evaluation of scenarios by authors are

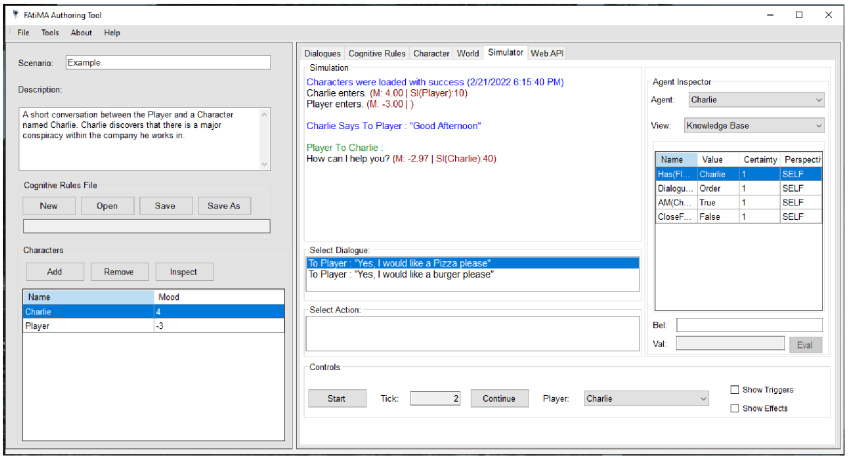

Fig. 4. FAtiMA-Toolkit Authoring Tool's Simulator component which allows authors to quickly prototype and test a SIA-based scenario [40].

crucial, with play-testing serving as an indispensable mechanism for bug detection, subsequent implementation of fixes and overall scenario refinement.

For these reasons agent modelling tools have a wide array of tools that can support debugging and testing tasks. FAtiMA-Toolkit, for instance, provides a Simulator panel where authors can play the scenario of any of the characters and test different components, as depicted in Figure 4. Through a user interface, ALMA displays the affective state of any of its characters and their reactions to different events. EMA also provides a simulation environment where agents can be initialised with a world configuration and allows them to interact with the test world [39]. To address the challenges during this authoring process, DeKerlegand et al [12] developed a game to train end-users to get familiar with the Ensemble tool's processes.

Despite the wide variety and plurality within the field, theory-driven frameworks tend to rely on similar concepts. Every character embodies a unique internal state, encapsulating their worldviews, traits and aspirations. Irrespective of decision methods, all actions necessitate conditions tied to character-specific characteristics and beliefs, with consequences influencing their surrounding world state. The OCC model of emotions, for example, creates emotions based on the interplay between events and appraisal variables, typically connected through appraisal rules. To manage dialogue, frameworks control the conversational state either via the agent's knowledge base variables or direct utterance-actions linkage. The incorporation of a "tag systems" provides extra complexity layers to performance. The analysis presented here allows us to move towards the next step when studying the authoring experience of agent modelling tools: How do authors use and manipulate the concepts present within these frameworks.

4 Understanding the Authoring Experience

During our research into the state-of-the-art of agent modelling tools, we found a lack of studies concerning the authoring experience itself. How do authors perceive, understand and use socio-emotional concepts to create interactive scenarios, when using a social agent architecture. This is a fundamental step in this work. In this section we present two different studies focused on, first, how authors perceive and understand SIA-related concepts and second, how authors use those concepts to create fully-fledged scenarios.

4.1 Study 1: Are Theory-Driven SIA Concepts Understandable?

Theory-driven agent modelling tools are rooted in Dennett [13]'s Intentional Stance, who postulated that a rational agent with beliefs, desires and other mental states exhibits intentionality; the agent's past and future behaviour is reliably predictable. Essentially, we recognise others as "intentional" beings and interpret others' minds as having "intentional states" [61]. The subject of perceived intentionality is now more relevant than ever, experts claim this is the only viable strategy for non-expert end-users to understand, predict and perhaps learn from artificial agents' behaviour in everyday social contexts [53].

This study aims to determine if the intentional-stance-based design does in fact help designers and developers in understanding and perceiving socially intelligent-based concepts.

4.1.1 Objectives. The objective of this study was to evaluate how authors, who use agent modelling tools, understand key concepts related to SIAs, such as beliefs, desires, goals, emotions and actions. By "understanding," we refer to the ability of authors to accurately comprehend and apply these concepts in practical scenarios. Additionally, to mimic demographics of agent modelling tool users, we compared the performance of experienced and inexperienced authors.

The study's results support the claim that SIA-related concepts, such as beliefs, goals, actions and dialogues, evaluated in the survey, are comprehensible and interpretable for experienced authors, and most, except for emotions, are also understandable for inexperienced authors.

4.1.2 Methods. Understandability reflects the extent to which a human can comprehend a model and the concepts underlying its inner processes [33]. In this study, we assess understandability through a series of comprehension and problem-solving tasks which both groups of participants were asked to complete.

Participants. A total of 20 participants, 15 male and 5 female participated in the study. The group was designed to consist of 10 experienced and 10 inexperienced individuals in terms of SIA authoring. Experienced participants were Computer Science students who had previous authoring experience with FAtiMA-Toolkit. Each individual contributed to the creation of at least two different authoring scenarios, at least 10 hours of authoring experience, and a 3-hour "Getting Started" workshop. Inexperienced participants were Computer Science students who never heard of or used FAtiMA-Toolkit or similar tools before. All participants received a 20€ voucher by completing the survey.

The "Experienced" group averaged 23.8 years of age ($SD = 5.7$) and, on average, spent around 23.34 minutes ($SD = 7.09$) to complete the tasks. The "Inexperienced" group averaged 22 years of age ($SD = 2.5$) and spent around 26.39 minutes ($SD = 12.45$) to complete the tasks. All participants received a 20€ voucher by completing the survey.

Metrics. The primary metric used to assess understandability levels was participant's score; the number of correct answers, the time used to complete the survey and the response latency (a measure of cognitive difficulty).

To evaluate possible learning effects of the study, the perceived competence sub-scale of the **Intrinsic Motivation Inventory (IMI)** questionnaire [42] was used. In addition to this, participants were asked to rate each question according to their perception of how much effort they put into its completion and their perceived difficulty level.

Regarding the perceived competence sub-scale, it was captured using six different items for both the conditions, before and after the tasks. Each of the items prompted participants to state their level of agreement, using 1–7 Likert scale, with perceived competence sentences such as "I am

Table 2. Tasks Level and Their Definitions

Task Level	Definition	Task Number
Specification	Creation of SIA artefacts	1, 3
Syntax	SIA artefacts structure	2
Problem-solving	Artefact manipulation, input and output	4 to 8

satisfied with my performance at this task." To measure item consistency we used the Cronbach's Alpha test.

The value for the Cronbach Alpha was as follows: for the experienced group, before the survey was 0.895, and afterwards, it was 0.9. For the inexperienced group before the survey was 0.616, which led us to ignore one of the items, leading to a 0.76. After the survey, for the same group, the Cronbach's Alpha value was 0.89. These values allowed us to average all of the items composing the scale and perform sample tests on the whole group.

Procedure. Participants were recruited through emails sent to students' mailing lists. The goal was to obtain specialised subjects for this evaluation, i.e., individuals who have already created social scenarios using FAtiMA-Toolkit and individuals with the same academic background (Computer Science) but have not interacted with FAtiMA-Toolkit before.

The study was conducted online thus, a link was sent to participants, which redirected them to an online survey, automatically assigning them a unique ID, preserving anonymity. The survey started with a description of the task, research goals and consent form. Upon acceptance of the conditions, participants received the following instructions:

(1) Step 1. Watch the following tutorial video on "What are intelligent agents?"
(2) Step 2. Complete the following questionnaire divided into different tasks (see supplemental material)

The tutorial video consisted of a short 10-minute presentation on the basic concepts behind SIAs. It included the perception-action cycle, the usual internal components of agents such as goals, beliefs and desires, and a few examples of how agent modelling architecture defines actions and emotions.

Tasks. Following the tutorial video, participants were presented with a set of 10 different tasks. Their order was fixed and designed to resemble the authoring steps in traditional agent modelling tools.

Tasks, or exercises, were divided into specification, syntax and problem-solving [54]. Specification exercises asked participants to create artefacts based on natural language descriptions, Semantic tasks queried participants about the correct syntax, finally, problem-solving exercises challenged participants to rationalise over constructs and simulate a possible output given a situation description using the SIA artefacts. The complete survey, along with all of the tasks, are in supplemental material. Table 2 presents the tasks distributed by their type.

4.1.3 Results. The dimensions studied were perceived competence, task success, perceived effort and difficulty, and task time and response latency.

IMI—Perceived Competence. An independent sample test was performed to the perceived competence of both groups, before and after the tasks. Before, a significant difference was found between inexperienced ($M = 4.03$, $SD = 0.18$) and experienced participants ($M = 4.78$, $SD = 0.94$); $t(18) = -2.146$,

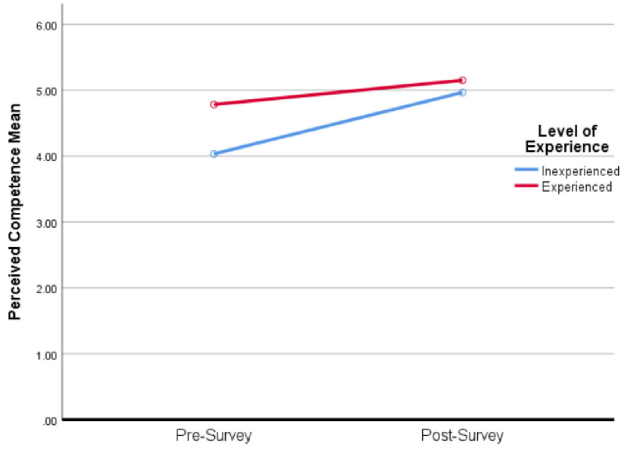

Fig. 5. Difference between conditions regarding perceived competence before and after the experience.

$p = 0.046$. In this case, the perceived competence levels were affected by their experience $F(1,18) = 4.6$, $p = 0.046$. However, the effect is not very strong, as indicated by its partial eta squared of 0.2.

After completing the tasks, while both groups increased their perceived competence levels, no significant difference was found between the inexperienced ($M = 4.97$, $SD = 0.25$) and experienced ($M = 5.15$, $SD = 0.29$) participants.

A significant difference was found in the scores for perceived competence for inexperienced authors, before the tasks ($M = 4.03$, $SD = 0.18$) and after ($M = 4.97$, $SD = 0.25$); $t(9) = -3.85$, $p = 0.004$. Regarding Experienced authors, a significant difference was also found in the scores for perceived competence, before ($M = 4.78$, $SD = 0.3$) and after ($M = 5.15$, $SD = 0.29$); $t(9) = -2.8$, $p = 0.021$ the tasks. The difference is represented in Figure 5.

Task Success. An independent-samples t-test was conducted to determine whether there was a difference between the overall scores of experienced and inexperienced participants. In total, the experiment had 10 different questions graded equally, 0 in case it was the wrong answer and 1 if it was correct. Thus, the maximum attainable score was 10.

Regarding the overall correctness of the tasks, results indicate a significant difference between inexperienced ($M = 7.7$, $SD = 1.25$) and experienced ($M = 9$, $SD = 0.94$) participants. [$t(18) = -2.6$, $p = 0.017$].

When looking at each specific type of task, in terms of "Specification" level tasks, there was no difference between the groups. Almost all experienced and inexperienced participants successfully performed Tasks 1 and 3 (8 and 9 correct answers in both groups for each task, respectively). Task 2 also showed practically no difference between the conditions since, from each group, only one person failed to complete the task correctly.

In terms of tasks related to problem-solving, differences between the groups were found. More specifically in Tasks 5, 6.1, 6.2 and 8.2. These, the most significant being Task 6 which concerned emotional appraisal and theories of emotion. A significant difference was found in Task 6.1 between inexperienced ($M = 0.3$, $SD = 0.48$) and experienced ($M = 0.9$, $SD = 0.32$) participants [$t(18) = -3.29$, $p = 0.04$]. Here, it is important to underline the fact that the 0 value corresponds to an incorrect answer while 1 corresponds to a correct response. In Task 6.2, no significant difference was found, however, the inexperienced group only provided five correct answers whereas the experienced group 8.

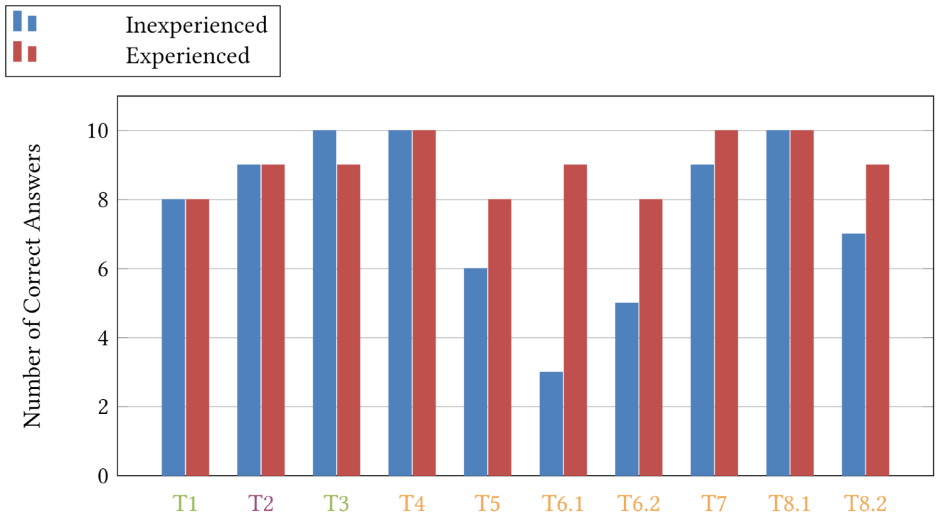

Fig. 6. Number of correct answers in each task by condition. Tasks are organised by their level: specification, syntactic and problem-solving.

Finally, no significant differences were found between the number of correct answers in Task 4, 7 and 8.1 where scores were quite high. Figure 6 captures the number of correct answers given in each task between the two groups and between different task levels.

If we were to remove emotion-related questions from the equation there is no significant difference between the overall score of inexperienced ($M = 6.9$, $SD = 1.2$) and experienced ($M = 7.3$, $SD = 0.82$) participants. In this case, highest score attainable would be 8.

Perceived Effort and Difficulty. During the survey participants were asked to rate each item according to their perception of how much effort they put into the associated task and how difficult they thought it was. Participants self-evaluated their effort using a Likert-Scale from 1–5 where 1 was "No effort at all" and 5 represented "A Lot of effort." Regarding the difficulty, participants rated the questions using a 1–5 Likert Scale where 1 was "Very Easy" and 5 was "Very Hard."

An independent-samples t-test was conducted to determine whether there is a difference between the effort and difficulty of experienced and inexperienced participants. No significant difference was found between the overall effort put into the tasks between inexperienced ($M = 2.44$, $SD = 0.48$) and experienced ($M = 2.27$, $SD = 0.38$) participants. In terms of difficulty however, a significant difference was found between inexperienced ($M = 2.53$, $SD = 0.09$) and experienced ($M = 2.16$, $SD = 0.36$) participants [$t(18) = 2.58$, $p = 0.019$].

The difference between the difficulty and effort scores was uniform throughout most of the tasks regardless of their type. The highest difficulty and effort results were found in Tasks 5, 6.1, 6.2 and 8.1, incidentally, the exact ones where the scores were lower. Participants found the most difficult question to be 6.2 ($M = 3.25$, $SD = 0.79$) and also the one that required the most effort ($M = 3.05$, $SD = 0.95$), which equates to moderate difficulty and moderate effort.

Task Time and Response Latency. Experienced participants spent around 23.34 minutes ($SD = 7.09$), while inexperienced subjects used around 26.39 minutes $SD = 12.45$) to complete all of the exercises.

An independent samples test was conducted to analyse the difference in time of completion between groups (in minutes). No significant difference was found between inexperienced participants ($M = 26.39$, $SD = 12.45$) and experienced ($M = 23.34$, $SD = 7.09$) participants. Additionally,

a Mann-Whitney U was conducted to determine whether there is a difference between the time spent watching the presentation (in minutes) and no significant difference was found between inexperienced participants ($M = 9.05$, $SD = 8.1$) and experienced ($M = 6.37$, $SD = 1.97$) participants.

The final question asked participants to estimate the time they used to complete the survey. Interestingly, both groups had similar estimates. Experienced participants averaged at 31.8 minutes ($SD = 9.95$) while inexperienced participants estimated they spent around 31.4 minutes ($SD = 7.2$).

4.1.4 Discussion. The study's results support the claim that SIA-related concepts, such as beliefs, goals, actions and dialogues, evaluated in the survey, are comprehensible and interpretable for experienced authors, and most, except for emotions, are also understandable for inexperienced authors. Experienced and inexperienced participants showed proficiency at performing tasks related to the creation of SIA artefacts (specification), of the relationship between the different constructs within the theory-driven models (syntactic).

In terms of problem-solving: using SIA concepts to complete different tasks related to the authoring of intelligent social agents, there were some significant differences between the two conditions. More specifically, in terms of the emotional models and affective states, it is clear experienced participants were more proficient at performing these tasks.

Regarding overall difficulty and effort, participants found the tasks to be between easy to medium difficulty where both groups put little to moderate effort into completing the tasks. Higher difficulty was found in the tasks related to problem-solving, coincidentally, the ones where most inexperienced authors failed. Even so, experienced authors also felt these were the hardest and the ones that required the most effort. These findings suggest that emotional concepts can be more complex than others but, after some exposure, can be just as easily understood and used.

Finally, in terms of time spent, it is important to note that while the platform used to host the experiment was able to keep track of the start and end time of the surveys, unfortunately, there were a lot of outside variables we were not able to account for. Since the test was done remotely participants had around a week to complete and often started a task and only finished it long after (in one case even, days after). Thus, within these metrics, there are a lot of disparaging results.

4.2 Study 2: How Are Theory-Driven Frameworks Used in Practice?

Agent modelling tools strive to make the transition between theory and practice to be as simple as possible. Nonetheless, the relationship between its user and the tool, in this case, the author of a SIA-scenario and an agent modelling tool, should be more carefully studied.

4.2.1 Objectives. The purpose of this second study is to analyse the output of agent modelling tools themselves. In particular, we performed a comparison between different scenarios made by different authors with the same goal; to create a realistic human–agent interaction. The results presented are not looking to assess a scenario with a qualitative mindset but, instead, provide a insights regarding which components are most used, and establish a baseline for complexity and amount of authorial effort.

We found that most of the time and effort of authors is spent on creating dialogues and decision-making-related content. Additionally and, perhaps more importantly, contrary to what was expected, authors deployed very little to no emotional-related concepts or rules, essential to the emotional capabilities of these types of systems and one of their core strengths.

4.2.2 Methods. FAtiMA-Toolkit has been used by the AI in games course of Instituto Superior Técnico of the University of Lisbon in the Computer Science Master's Degree. For the past 3 years over a hundred students have used the tool to create human–agent interaction scenarios [40].

Our analysis will focus on the work of students from the 2021/2022 semester. For their final project, students were tasked with making full use of FAtiMA-Toolkit's features to create two different social agent experiences. The structure of a FAtiMA-Toolkit authored scenario, like most traditional agent-modelling approaches, is a set of rules with conditions within different components. Scenarios have several agents with different beliefs and goals, dialogues with different dialogue states and components with rules and conditions, similar to the content in Section 3.

Procedure. To get acquainted with the tool, students had an additional 1.5-hour practical class focused on FAtiMA-Toolkit on top of the theoretical lecture regarding interactive narratives. This class, while not focused on the tool itself, approached several important topics such as the perception-action cycle, computational models of emotion and social agent architectures.

Students were tasked with creating two different scenarios. The first scenario was the implementation of a small story using the authoring tool. The objective of this task was to familiarise students with the tool, set a baseline for all the groups and serve as a checkpoint for the teachers. The following is the description of the scenario given to students:

> *"Peter was hungry so he went to a restaurant. Once there, he ordered a hot dog. The waiter told him they only served hamburgers. Peter told the waiter that was okay as well. After a while, the waiter brought Peter his food. The hamburger was burnt to a crisp. Peter complained to the waiter. The waiter told Peter they had no more hamburgers. Peter immediately left the restaurant without paying."*

In the second scenario, students were given considerable freedom regarding their "creations." The scenario had to fill a few requirements such as: using two different agents, having 15 different dialogue states and use at least three different reasoner components such as the emotional appraisal, emotional decision-making and world model. Naturally, students were encouraged to expand upon these requirements, which many of them did.

To complete their assignments, students were given 2 weeks for the deadline. To support their design and implementation process, the Toolkit's official Web site has a wide array of documentation and tutorials that are directly available to anyone who seeks them.

Sample Group. The analysis presented here uses the work of 35 different students which formed 16 groups with 2–3 authors each ($M = 2.31$, $SD = 0.7$). All of the participants were graduate students attending the first or second year of the Computer Science Master's degree at Instituto Superior Técnico, University of Lisbon.

4.2.3 Metrics. Evaluating the amount of effort necessary to create an SIA-scenario requires a metric that assesses its complexity and scale. When we refer to complexity, we refer to its cognitive complexity; the amount of work necessary to achieve the intended end-result. The disparity of complexity and scale of different SIA-agent scenarios led us to the need of defining a metric that could assess their relationship; thus we introduce the notion of "artefact."

An artefact can be: a rule, a condition, a belief or a dialogue, it captures anything that has been manually authored by the author. One action rule is one artefact, one action rule with one condition is two artefacts, two action rules are two artefacts, and so on, as shown in Table 3. Even when there are no conditions, an action rule, for instance, can be triggered and was manually authored, thus, we consider it to be the baseline.

The relationship between artefacts and the cognitive complexity is linear; the higher the number of artefacts, the higher the complexity and scale of a scenario.

Table 3. Artefact Example: A Scenario with Two Rules with 1 Condition between Them Has the Same Complexity as a Scenario *i*th One Rule with Two Conditions

Number of Rules	Number of Conditions	Number of Artefacts
1	0	1
1	1	2
1	2	3
2	1	3
4	4	8

It should be noted that the same reasoning was applied to each component of the scenario; One agent is one artefact, Two beliefs are two artefacts, three dialogues are three artefacts, and so on. The artefact metric is at the core of our analysis.

4.2.4 Results. In the analysis conducted we looked at amount of artefacts created by component type and by task.

Task 1 Results. Regarding Task 1, with 16 different participants (scenarios) to implement the previously mentioned description, on average 95.6 artefacts ($SD = 36.7$) were created. The mean amount of different dialogues was 18.2 ($SD = 14.7$) with two different agents ($M = 2.1$, $SD = 0.250$).

In addition, we asked groups to estimate the time spent learning the framework and then implementing the scenario itself. Not all of the groups responded. However, from a sample of 14 groups, each member spent, on average 2 hours ($SD = 0.68$) learning about the toolkit, and then the whole group spent around 4 hours ($SD = 1.0$) implementing the story.

Task 2 Results. Regarding Task 2, its open-ended nature led to a wide range of different scenarios, such as a blind date, an AA meeting, a blackjack game and a stand-up comedy show, among others. Overall, the number of artefacts created on average was 178.8 ($SD = 101.8$), the number of dialogues was 57.4 ($SD = 37.6$), with the average number of agents was 3 ($SD = 1.2$). The average scenarios had more than three endings ($M = 3.44$, $SD = 1.54$), and the average number of cycles was 15.38 ($SD = 12.06$). From a sample of 11 groups, each member estimated they spent around 8 hours ($SD = 5.9$) implementing their idealised scenario.

Discussion. As shown in Figure 7, students spent most of their time and effort creating dialogues and designing decision-making-related artefacts. Additionally and, perhaps more importantly, contrary to what was expected students did not use much of the emotional concepts.

Due to FAtiMA-Toolkit's hybrid approach to dialogues, as described in Section 3, the decision-making and dialogue manager components are intrinsically connected. As a result, participants that create dialogue-heavy scenarios are more prone to create more decision-making artefacts as well. However, there is no clear explanation behind the low amount of time and effort put into creating affective-driven artefacts.

The study data indicates that, to create an interactive scenario with 3 agents, around 178 artefacts, 15 interaction cycles and at least 3 different endings, an author must spend an expected 21 man-hours (8 * 2.3 working hours + 5 hours for learning) of their time.

4.3 Overall Discussion

The studies presented in this chapter provided insights into the understandability of theory-driven SIA concepts and the author's tendencies when using them within an agent modelling tool. Our

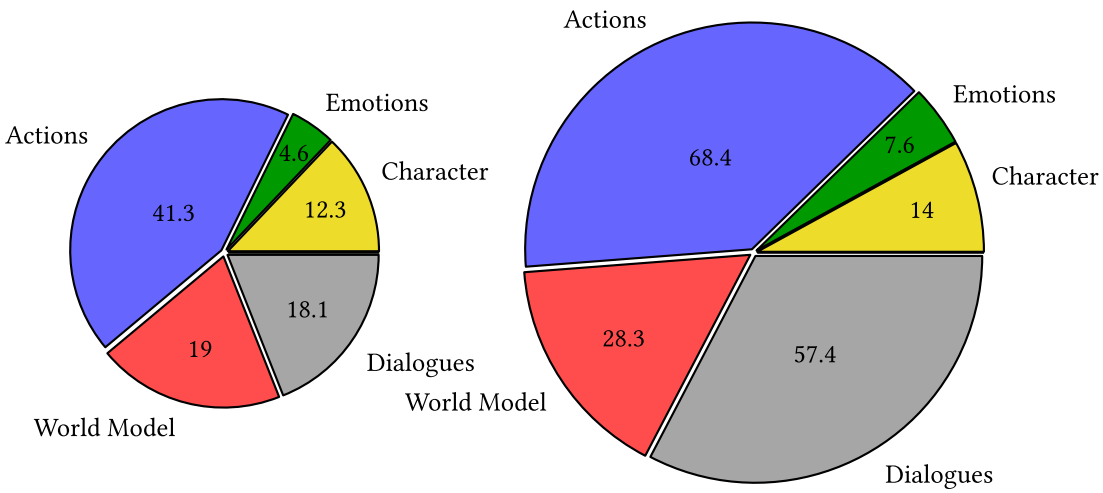

Fig. 7. *Overall Results*: Average artefact count by component. The pie chart on the left represents Task 1 results, while the one on the right represents Task 2's results.

studies show most concepts, except for emotions, appear to be relatively easy to understand for both experienced and inexperienced authors. After watching a 10-minute presentation, participants with no previous experience in modelling intelligent social agents could grasp SIA artefacts, understand the relationship between them, and successfully perform problem-solving tasks almost as well as experienced authors.

When University of Lisbon Computer Science students were tasked with creating their own human–agent interaction experiences using FAtiMA-Toolkit, most students prioritised the dialogue and decision-making components. The analysis allowed us to establish a baseline where for one person to create a complex scenario composed of at least three different agents, multiple endings, and almost 200 artefacts had to spend around 21 man-hours.

The user study and analysis revealed an apparent lack of comprehension and use of the affective components of traditional social agent frameworks. Multiple factors can explain these findings. Authors may avoid using such components due to the lack of understanding and the usual complexity of the model behind the generation of emotions. It is also possible that, since authors were tasked with creating a human–agent interactive scenario, they put most of their work into the flow of conversation, leaving other components such as emotional appraisal behind.

In the context of this work, these findings, along with the state-of-the-art research described above, give us insight into the current authoring experience problems authors face when trying to create a rich scenario. In particular, we have identified a significant issue relevant to the SIA, and Affective Computing fields. Considering the amount of research and work put into understanding how humans process and generate emotions, along with the development of a wide array of computational models of emotions, to our surprise, it seems the concepts used are not as easily understood or worked with by others, as it was previously thought.

5 Authoring-Assisted FAtiMA-Toolkit (AA-FT)

Our goal is to advance towards the next generation of affective social agents by creating better tools to model human–agent interactions and facilitate their authoring experience. Automating parts of this process could be the key to the wider adoption and proliferation of SIAs and theory-driven agent-modelling tools.

The recent advances in high-performance computing and the exponentially growing amount of structured and unstructured data obtained from various sources led to the notion of data-driven or data-oriented tools as a new paradigm of science [11, 31, 59]. Recently, one of the most popular use cases of these approaches is the development of **Large Language Models (LLMs)**. With training on extensive data and millions of parameters, LLMs have attained unparalleled generalisation potential, significantly impacting academia and industry alike [36].

One of their most renowned features is the ability of being used as-is (in zero-shot scenarios), given just a few examples (few-shot scenarios), or fine-tuned (trained a few epochs) on smaller datasets to learn a new task while maintaining the general linguistic knowledge acquired during the pre-training [9]. The amount of information consumed by these models makes it so that, inevitably, they *implicitly* encapsulate different aspects of social behaviour (e.g., identification of concepts, speech acts, emotional language) and can create the illusion that they capture user's intentions, goals and are even able to experience things and have feelings.

Nonetheless, LLM-based models suffer from a multitude of issues; *repetition* [57], they can *hallucinate false information* that can be irrelevant or incoherent with the context [57] and have trouble maintaining a persona [63] in long-term interactions [67]. Additionally, their considerate size and layers of training makes so that it is difficult understand their output or the reason behind it. Yang et al. [68] considers this type of approach to offer *low interpretability* and limited control, which might lead to undesired behaviour such as suggesting dangerous or fatal activities to users.[4]

Thus, data-driven approaches, in general, sacrifice *authorial control* and *transparency* in favour of *scalability*, where vast amounts of content can be automatically generated in a short amount of time. Moreover, LLMs are still a particularly unpredictable technology and thus, ill-advised to create social experiences in safety-critical systems (e.g., mental health and healthcare) or applications with tailored user experiences. Data-driven approaches should be heavily supervised.

One the other hand, architectures developed under the "theory-driven" umbrella provide a high degree of *authorial control* to its user[5] and the ability to create a myriad of *transparent* scenarios with various levels of *complexity*. Currently, such complexity is achieved by manually describing how SIA concepts (goals, intentions, actions, emotions, culture, etc.) interact with each other. Human–agent interaction designers must also account for end-users' decisions throughout the experience and make sure its intervenients are correctly playing their role without hindering the plot or narrative line. If handled naively, accounting for past events in the story as the narrative unfolds can lead to an exponential amount of content that needs to be created. We propose to change the state of affairs through a hybrid approach.

5.1 Design Goals

In order to create a framework that is capable of leveraging data-driven approaches to generate human–agent interaction content automatically, it was necessary to clearly define its design goals. It is of utmost importance to avoid hindering the current authoring experience of existing social agent architectures. Thus, considering the lessons learned and described in previous sections, the proposed tool strives to accomplish the following goals:

— *Context-Based*: The output of the process is to be based on information provided by authors, either using examples of behaviour or descriptions of events. As a result, the tool avoids hallucinating information or creating more data than the author asked of it.

[4] https://www.bbc.com/news/technology-59810383

[5] Scenario author/designer.

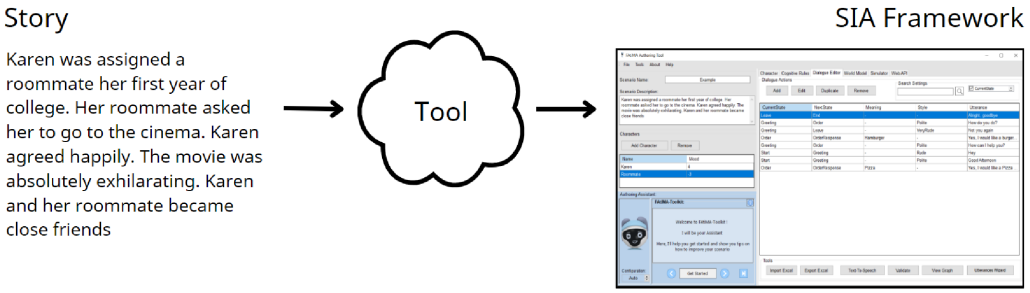

Fig. 8. Initial concept of the desired input and output using a story from the ROCStories dataset [49].

— *Automated*: Authors do not have to provide large amounts of information to output a scenario. The framework's objective is to minimise the amount of content necessary to manually create thus, asking for authors to provide vast descriptions of the scenario would defeat its purpose.
— *Auditable*: While a large amount of the process will use data-driven models, the tool is designed to be transparent and auditable where, at each step, authors can validate and change the content that is being produced.
— *Understandable*: Throughout the process, both input and output are transparent and can be interpretable for both experienced and inexperienced authors.
— *Responsive*: The iterative nature of the authoring process asks that each of its parts be robust, quick and adaptable. The automated tool generates content as rapidly as possible without sacrificing accuracy.

To accomplish the objectives stated above, our approach extracts relevant information from stories provided by authors and translates the information into an SIA-human–agent interaction scenario. Thus, the output of this process contains socially intelligent and emotional characters with context-based beliefs and goals, actions, reactions and consequences.

There are two main reasons behind the story-driven nature of the tool. First, stories play a significant role in how we as humans try to make sense of the world around us [10]. Many of our interactions with other humans are by way of storytelling. Harrison et al. suggest communication between humans and machine learning algorithms would be more natural if machine learning algorithms could use the stories told similarly to how they use demonstrations [28]. Thus, for any author, stories tend to be a natural and simple way of describing a desired interaction scenario.

Second, state-of-the-art research has shown potential regarding the use of stories as a base for models to learn social behaviour. Researchers have focused on creating automated tools for story comprehension and generating narrative plots themselves [2, 64]. In addition to this, given the central role of stories in human communication, there is an almost interminable amount of story examples available in a wide array of different modalities. Finally, researchers have developed several different easily accessible story-based datasets such as the Story Commonsense dataset [58], and ROCStories [49].

5.2 S2SIA—Story to SIA Artefacts Tool

A story-based solution allows us to access a diverse collection of data as well as use a few of the wide array of open-source NLP-focused tools. Figure 8 depicts the first concept of the tool using a story from the ROCStories dataset [49] as an example.

The name of the tool reflects its single function: *S2SIA—Story to SIA Artefacts Tool*, to automatically generate SIA artefacts based on the story provided by a human author. While tailored and integrated

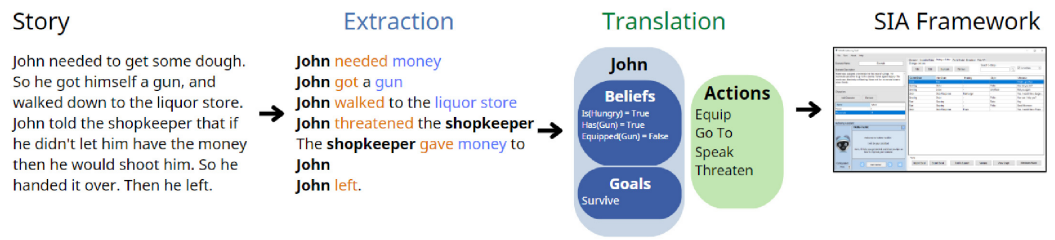

Fig. 9. Diagram of the pipeline of our proposed solution. The extraction phase takes as input the original story provided by authors and extracts information such as its actors, actions and objects. The translation phase then creates a scenario based on the extracted information.

into FAtiMA-Toolkit [40], we view the tool's output as data that could easily be used by other theory-grounded social agent modelling frameworks.

The program performs two fundamental and sequential steps. The first, named *Information Extraction*, pertains to extracting all relevant SIA concepts and events from natural language written stories. The second step, *Information Translation*, is focused on collecting all of the information gathered by the previous step and translating it into SIA concepts such as characters, beliefs, goals, action rules, appraisal rules and dialogues, among others as shown in Figure 9.

Each of these stages offers unique and different challenges. As a result, each task was designed and implemented using different approaches which are the result of the lessons learned from state-of-the-art research and our experience as human–agent interaction designers.

5.2.1 Information Extraction. The *Information Extraction* phase receives open-ended input text and generates relevant semi-structured information regarding SIA-relevant concepts. These concepts are directly extracted from the author's provided description. As we have seen in Section 3 while there is a wide range of theory-driven frameworks, their authoring and internal processes revolve around similar concepts.

The issues found in end-to-end data-driven models make them unfit to be used plainly used in our solution. To avoid sacrificing transparency, it is necessary to deconstruct the task where different complementing tools could be leveraged to extract relevant social information from author-provided data. In particular, we make full use of two Information Extraction tools: PredPatt[6] and Framenet.[7]

This process culminated in the sequence of steps presented in Figure 10. First, the natural language text goes through a process called solving co-references, where the actual name or label of the character replaces any existing pronouns in the text. This step's objective is to prepare the input to be computed in Pred Patt. Each sentence goes through Pred Patt which can extract the subject, predicate and target. The system is also prepared to deal with complex events such as "John loves going to the cinema," where the target has another predicate. In these cases, Pred Patt outputs two different subjects, predicates and targets, all associated with the same event.

The final step is to use Framenet to associate semantic frames to each event. For example, the verb "go" has several attached frames: "Being_named, Compatibility, Motion…." The semantic frame contains an illustration of an event and the relationship between its different components, entities, or participants. Thus, given a verb or a noun, such as "see.v," it provides a set of frames associated with the verb such as "Perception_experience, Graps, Touring, Categorization"[8] for example.

[6] https://github.com/hltcoe/PredPatt
[7] https://framenet.icsi.berkeley.edu/fndrupal
[8] https://framenet.icsi.berkeley.edu/fndrupal/luIndex

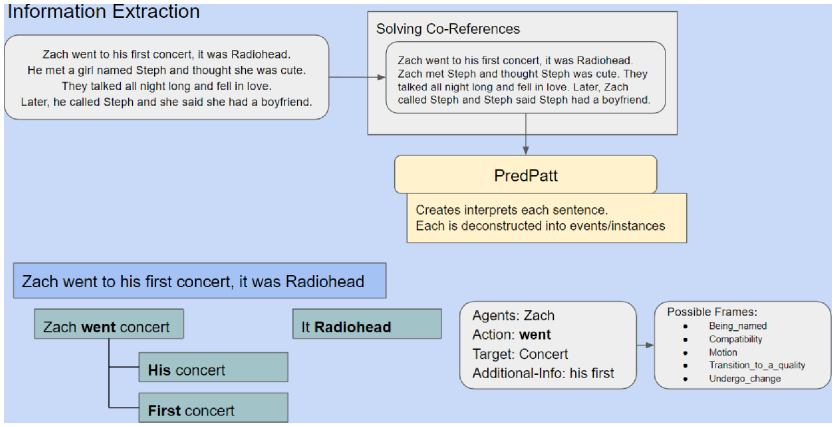

Fig. 10. Information extraction pipeline for the event: "Zach went to his first concert. It was Radiohead"

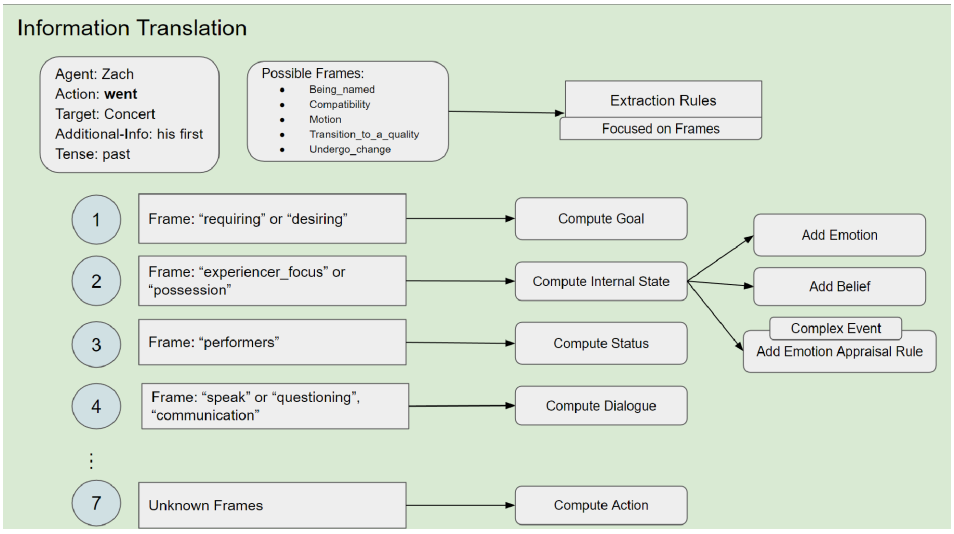

Fig. 11. Event Information Translation using a rule-based system.

5.2.2 Information Translation. The *Information Translation* phase receives as input the semi-structured information containing the extracted SIA elements from the previous stage. It translates these into fully fledged SIA artefacts that can be used to create a human–agent interactive scenario. We view this process as highly reliant on human–agent interaction designers' experience and aforementioned intuition.

A rule-based system was designed to harness these aspects and mimic the rationale behind how experienced authors create social agent interactions. The system prioritises detecting specific cases and moving towards more generic ones, including those that the system does not recognise. For example, the first rule is designed to find events that describe Goals. Thus, it looks for specific frames such as "require" and "desiring." If detected, the event is redirected to a "Compute Goal" method which adds a goal, using the initiator, action and target, to the scenario artefacts. Figure 11 describes the rule-based system.

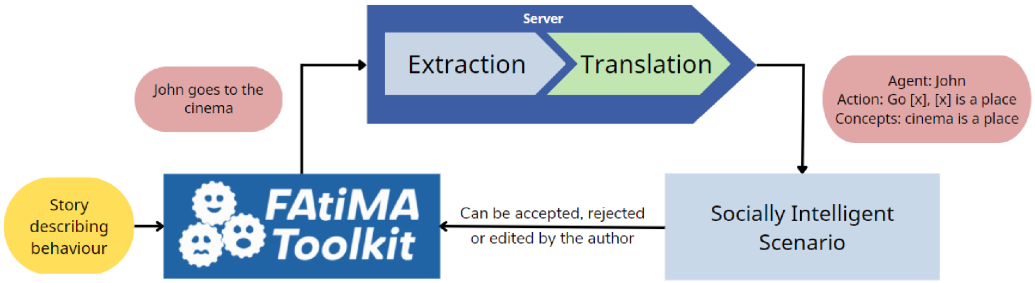

Fig. 12. The pipeline's integrated diagram.

It is important to note that this process highly depends on the frames found by the Framenet corpus and by the rules' sequence. The frame establishes the type of event, and the component-specific functions are in charge of extrapolating the rest of its elements: agents that participate in the frame, the targets of the event, locations or objects and explicit or implicit goals behind the action, among others.

This approach is further reinforced by data-oriented tools to accomplish specific tasks. For example, to detect emotion-based concepts we used Spacy, a free open-source library for NLP in Python [32]. Spacy allows us to access different models to find the relationship between emotions provided by authors and its equivalent in the OCC model through word similarity, by finding the proximity between word vectors in the embedding vector space of a language model.

An Emotion Labeller program was created that contains a table that directly maps emotions in the OCC model to their corresponding appraisal variables. The program is able to calculate, given a particular word, which is the closest emotion to it, within the word embedding model, and returns the appraisal variables associated with it.

5.2.3 Integration. To integrate the developed information extraction and translation pipeline (implemented in Python) with FAtiMA-Toolkit (C#), it was essential to consider the previously set requirements. One fundamental condition is its low demand for computational resources from its users. To achieve this, an HTTP server was developed where the S2SIA pipeline was hosted. The necessary computations are performed here and sent to the connected agent modelling tool. Our internal testing indicate that the latency of this process, 5–7 seconds, is negligible.

Figure 12 displays how the toolkit communicates with the server using a simple example such as "John goes to the cinema." Communication within FAtiMA-Toolkit and the server is handled by its Authoring Assistant. To avoid confusion regarding the different pieces that are part of this process, authors using the toolkit have access to a feature called "Story to Scenario Wizard." The name encourages the use of the feature and the curiosity of the users however, the whole process is handled by the assistant behind the scenes.

The "Story to Scenario Wizard" also provides authors with a key validation step right before the information outputted by the pipeline is transformed into a human–agent scenario. Authors are prompted with the complete output of the pipeline in a semi-structured form where they can accept, edit or refute their transformation and addition to the current scenario.

5.3 Authoring Assistant

Initially, the assistant's objective was to facilitate the authoring experience by serving as the bridge between the author, the S2SIA and the agent modelling tool itself. However, the issues raised in Section 4 and the lessons learned from state-of-the-art research, asked for an augment of its reach.

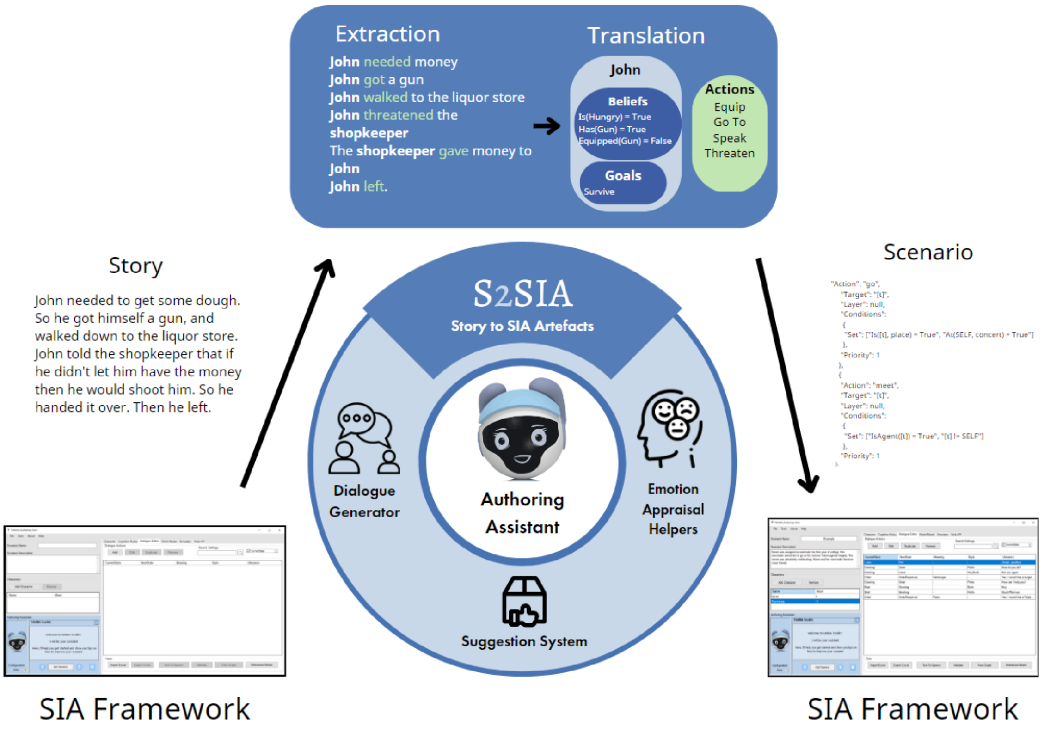

Fig. 13. The Authoring-Assistant works within the agent modelling tool to support the creation of human–agent scenarios. One of its features is the ability to not only access the S2SIA tool but to convert its outputted artefacts into a scenario within the framework.

As an active agent supporting the authoring process, the assistant can directly address understandability issues raised in Section 4. The Authoring Assistant does this by offering a wide range of emotional authoring helpers such as the ability to automatically create an appraisal rule or the ability to simulate the emotions that occur in a scenario, among others. We hope that improving how authors are exposed to computational models of emotion, such as the OCC model [51], can also lead to better comprehension behind these systems.

The data collected during the scenario's analysis, presented in Section 4, allows the definition of a baseline in terms of SIA-artefacts created in a scenario. When authors use a social agent framework tool, the assistant compares the current state of the scenario to others in its dataset. In cases where a particular component is under-prioritised, the assistant advises authors on ways to enrich the currently authored scenario. In addition to this, the Authoring Assistant also provides an in-tool tutorial of its user interface and provides debugging tools to further support the authoring task.

Finally, the flexibility of the assistant as the connector and facilitator between the author, agent modelling tool and other helpers allows us to expand its reach quickly. The assistant is also able to connect to a pre-trained language model, GPT-3 [9], and automatically generate dialogue based on the utterances already present in the scenario. Even a single dialogue line as an example is sufficient for the model to output context-based examples. Nonetheless, before using this feature, authors are encouraged to create Dialogue Actions, which will be used as examples.

Figure 13 provides an overall illustration of the proposed framework, in particular, the behind-the-scenes process of the bridge between the Authoring Assistant and S2SIA. We would like to underline

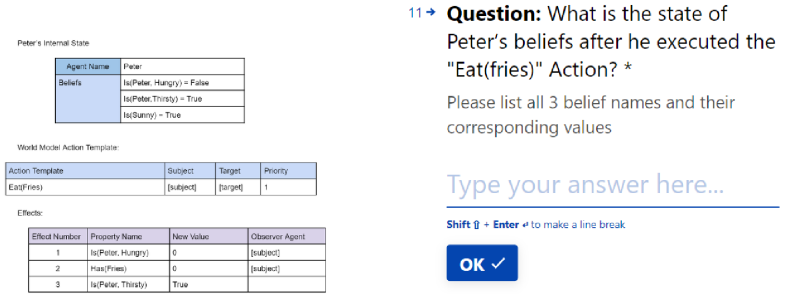

Fig. 14. Example of a concept manipulation task presented to participants.

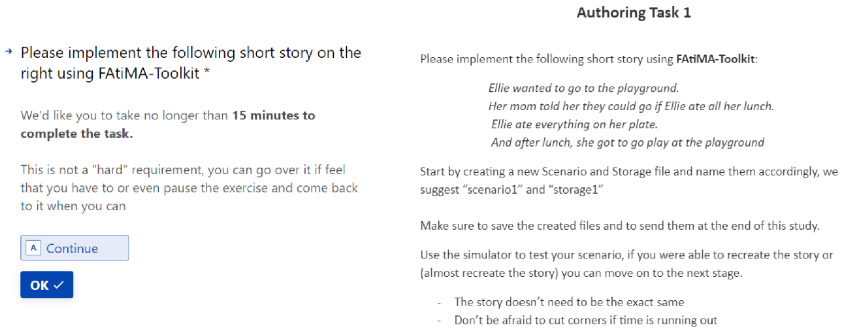

Fig. 15. A screenshot of the survey describing authoring task 1.

the fact that this system promotes authorial control and the use of understandable concepts by design. *AA-FT* remains under the theory-driven umbrella; it is merely aided by data-driven tools that the Authoring-Assistant audits at every step. We believe in the importance of having the human-in-the-loop, particularly when dealing with automated content generation. If our thesis is correct, the leveraging of data-driven approaches within a theory-grounded structure should lead to an agent modelling tool capable of automatically creating content without sacrificing its understandability and authorial control. In the following section, we will describe a user study conducted to evaluate the Authoring Assisted Authoring Toolkit's impact on the authoring experience.

6 Evaluation

The work presented in this manuscript culminated in the development of a theory-grounded agent modelling tool capable of harnessing the power of data-driven tools to facilitate its authoring experience. The primary issues behind data-centric mechanisms, led to a careful design-balance between two different approaches. It is essential to understand the impact of the introduction of this new paradigm within agent modelling frameworks and in the authoring experience itself.

6.1 Objectives

To understand the degree to which our goals were accomplished, we conducted a user study focused on assessing the impact AA-FT had on its users' authoring experience.

The study focuses on two main aspects: productivity, which refers to the effort required to create a human–agent interaction scenario using an agent modelling tool, and understandability, which examines whether the new data-driven components influence users' comprehension of the concepts and models they interact with.

Results found that, on average, authors using AA-FT significantly created more artefacts per minute than the ones using the previous version of the tool. Moreover, participants using the tool's new features were still able to understand the concepts they were manipulating and the tasks performed.

6.2 Method

We conducted a between-groups experiment with one treatment: creating an SIA-interactions using *AA-FT* vs. creating SIA-interactions using the "normal" version of *FAtiMA-Toolkit (FT)*, without the features presented in Chapter 5.

6.2.1 Participants. Participants were recruited through emails sent to students' mailing lists and through announcements shared university forums. Our goal was to obtain specialised subjects for this evaluation, i.e., individuals had previous FAtiMA-Toolkit experience and individuals who have a similar academic background (Computer Science) but had not interacted with FAtiMA-Toolkit before.

Taking into account the duration and specificity of the tasks, each participant received a 20€ voucher by completing the survey. Twenty-two different participants completed the study, with an average of 23.6 years old ($SD = 3.1$) and the majority being from a Computer Science background (95%). Ten of the participants used AA-FT while 12 used the previous version of the framework (FT).

Within each group, the distribution between levels of experience was 40%, in the AA-FT group; six participants had experience with the toolkit, and the remaining four had no previous experience with the framework. In the FT group; five users had previous experience with the same version of FAtiMA-Toolkit, while the remaining seven had no experience creating SIA scenarios.

6.2.2 Metrics. The user study and scenario analysis previously performed and described in Section 4 provided us with a solid base to build upon. Productivity was measured using an extension to the proposed "artefact" concept with the introduction of tracking of time. In terms of understandability, like before, it was determined using participant's score when completing comprehension and problem-solving tasks.

Artefacts. One crucial detail is the high variance of results in terms of time spent authoring, amount of *artefacts* created and completion of the task. Due to the nature of the survey and the soft time constraints included in each task, some participants left some of their scenarios incomplete. As a consequence, looking at the results from overall artefacts produced does not make much sense. In our view, the metrics used can and should be adapted.

A formula was used to account for the information obtained from each of the authored scenarios. In particular, the number of artefacts created and the time spent to create them. Hence, artefacts per minute, the average amount of artefacts that can be computed in a minute:

$$ArtefactsPerMinute = ArtefactNumber/MinutesSpent \quad (1)$$

The equation allows us to also measure the amount of effort necessary to create content between different components. A higher value of artefacts per minute equates to more artefacts being created in a lower amount of time.

IMI. The IMI is a multidimensional measurement device intended to assess participants' subjective experience related to a particular task [34].

An adaptation of the traditional IMI questionnaire was used in the study to better assess the authors' subjective experience of using FAtiMA-Toolkit. In particular, participants answered questions

pertaining to interest/enjoyment, perceived competence, effort, value/usefulness and perceived choice sub-scale questions.

Model Understandability. To assess how the new data-driven paradigm might have affected the framework we used a similar procedure as in Chapter 4 to gauge participants understandability of the model and their concepts. Participants completed five comprehension, concept-manipulation and problem-solving tasks revolving around SIA's concepts before and after completing the tasks, as described in Figure 14.

Each of the tasks pertained to a different component of SIAs and was graded according to their difficulty. Out of five questions, participants could achieve the maximum grade of 10. The results also allow us to find possible learning effects of using the framework. The items were the same before and after authoring tasks. The survey and these questions are presented in supplemental material.

Procedure. It was important to take into consideration the fact that the survey should be able to be completed by both experienced and inexperienced authors of agent modelling tools. Thus, creating a tutorial where participants could learn to use the tool was essential, the resulting survey had six different sections:

(1) Consent Form and General Information
(2) FAtiMA-Toolkit Tutorial
(3) Pre-Task Questions: model understandability and perceived competence
(4) *Task 1*: Create a human–agent interaction scenario based on Story 1 using FAtiMA-Toolkit
(5) *Task 2*: Create a human–agent interaction scenario based on Story 2 using FAtiMA-Toolkit
(6) Post-Task Questions: model understandability and perceived competence

Participants using AA-FT had access to an additional tutorial with information regarding how to use the newest features. In total, we estimated that to complete the tutorial, experienced participants needed around 40–50 minutes and inexperienced authors 60–75 minutes. All participants received a 20€ voucher by completing the survey.

The "FAtiMA-Toolkit Tutorial" section included a 10-minute presentation on the basic concepts behind social, intelligent agents and 20–30 minute presentation with videos and examples regarding how to use FAtiMA-Toolkit. This presentation consisted of around 50 slides that provided viewers with the necessary know-how to use FAtiMA-Toolkit to create human–agent scenarios. Participants could spend as much time as they wanted in the tutorial section.

Tasks. For each task, participants were asked to recreate a story as a human–agent interaction scenario with social agents using FAtiMA-Toolkit. The stories were picked from the StoryCommonSense [49] dataset with some adaptations.

The first story, Story 1, should be a simple and straightforward exercise. While also quite relevant to the user study, it's secondary objective is to serve as a first introduction to the survey and the tasks within it. *Story 1* is as follows:

```
Ellie wanted to go to the playground.
Her mom told her they could go if Ellie ate all her lunch.
Ellie ate everything on her plate.
And after lunch, she got to go play in the playground.
```

Participants were asked to perform this first task in around 15 minutes. The second story, Story 2, was the most complex of the two. Here, participants were asked to create branches within the same scenario, *Story 2* is as follows, the branches are described by "||":

```
  John was hungry.
   He went into a restaurant and ordered a sandwich.
   The waiter quickly served him || The waiter took a long time
   John payed and left the waiter a large tip || John was quite angry
and left the restaurant.
```

Participants were asked to perform this second task in around 25 minutes. The time constraint was emphasised but not enforced. Participants were told to prioritise completing the scenario to submitting it before the deadline. The objective is to ensure authors, particularly the experienced authors, would not spend hours working in the same scenario. The objective was to have multiple scenarios created in similar time frames. Figure 15 captures the prompt provided to participants for the authoring tasks.

6.3 Results

The dimensions studied were authoring effort, artefacts created per time, model understandability, perceived competence, enjoyment, effort and feature usage.

6.3.1 Authoring Effort—Story 1. On average, participants from the AA-FT group performed the first authoring task in 20.6 minutes ($SD = 9.2$). Participants from FT took quite a bit longer. However, there is an increased variance between the values: Median: 30.7 ($SD = 17.6$).

In terms of overall artefacts created, the AA-FT group authored, on average, 27.7 artefacts ($SD = 10.1$) while authors from the FT group created 21.8 ($SD = 11.9$) artefacts. The high variance of results even within the same group makes the defined metric more useful.

A Mann-Whitney U independent samples test was performed, and AA-FT participants, on average ($M = 1.5$, $SD = 0.7$), had a significantly higher amount of artefacts created per minute compared to the FT Group ($M = 0.94$, $SD = 0.78$); [$U = 91.000$ $z = 2.045$ p $= 0.043$, $r = 0.43$]. Figure 16 displays the values across the different SIA components within FAtiMA-Toolkit.

The following is the distribution of the participant's level of artefacts produced per minute.

- *Characters and Their Internal States*: AA-FT group ($M = 0.42$, $SD = 0.21$) had a higher level of character-related artefacts per minute compared to FT group ($M = 0.3$, $SD = 0.24$).
- *Decision-Making*: AA-FT group ($M = 0.19$, $SD = 0.1$) had a significantly higher level of decision-making-related artefacts per minute compared to FT ($M = 0.14$, $SD = 0.14$). [$U = 90.500$, $z = 2.044$, p $= 0.043$, $r = 0.43$; $r > 0.3$, medium effect size].
- *Emotional Appraisal*: AA-FT group ($M = 0.29$, $SD = 0.29$) had a significantly higher level compared to FT ($M = 0.02$, $SD = 0.04$). [$U = 104.00$, $z = 3.17$, p $= 0.003$, $r = 0.67$; $r > 0.5$, large effect size].
- *World Model*: AA-FT group ($M = 0.26$, $SD = 0.19$) had a higher level compared to FT ($M = 0.28$, $SD = 0.22$).
- *Dialogues*: AA-FT group ($M = 0.07$, $SD = 0.08$) had a lower level compared to FT ($M = 0.18$, $SD = 0.17$).
- *Overall*: AA-FT participants, on average ($M = 1.5$, $SD = 0.7$), had a significantly higher amount of artefacts created per minute compared to the FT Group ($M = 0.94$, $SD = 0.78$). [$U = 91.000$ $z = 2.045$ p $= 0.043$, $r = 0.43$; $r > 0.3$, medium effect size].

In terms of comparing in-group experience levels, an interaction effect was found when comparing experienced participants' results across framework versions. Experienced participants that used AA-FT to create the first scenario had a significantly higher level of emotional appraisal artefacts per minute than experienced participants from the FT group. [$U = 26.500$; $z = 2.208$; p $= 0.030$, $r = 0.67$, $r > 0.5$ large effect size]. Figure 17 captures this.

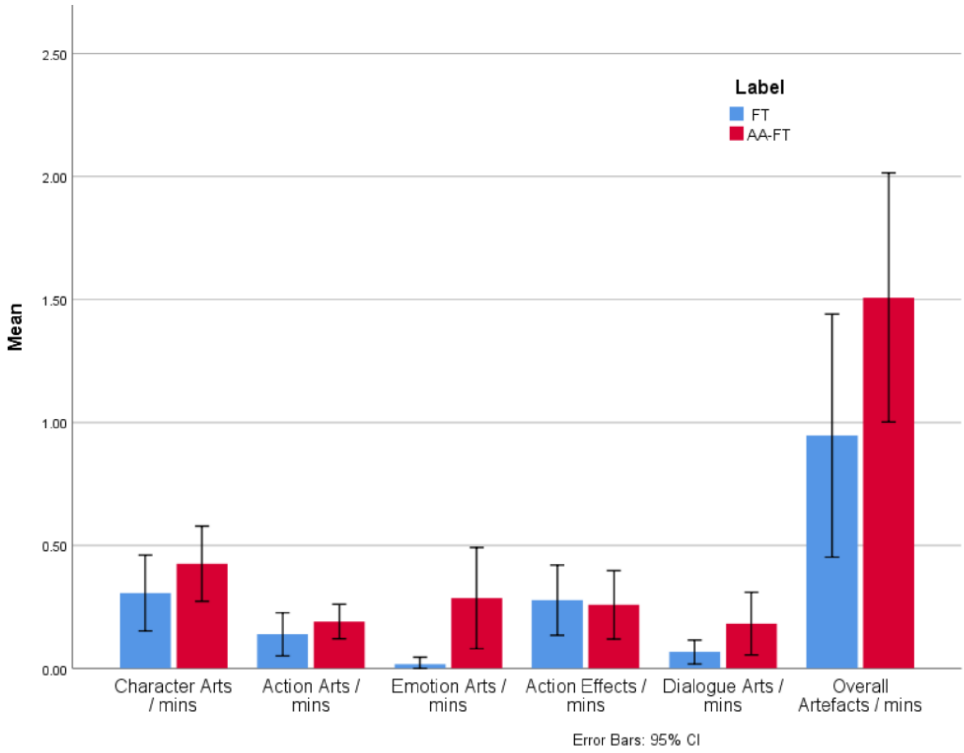

Fig. 16. Average artefacts per minute regarding the *Story 1* authoring task. AA-FT artefacts per minute were significantly higher than FT in terms of character, decision-making and emotional appraisal components.

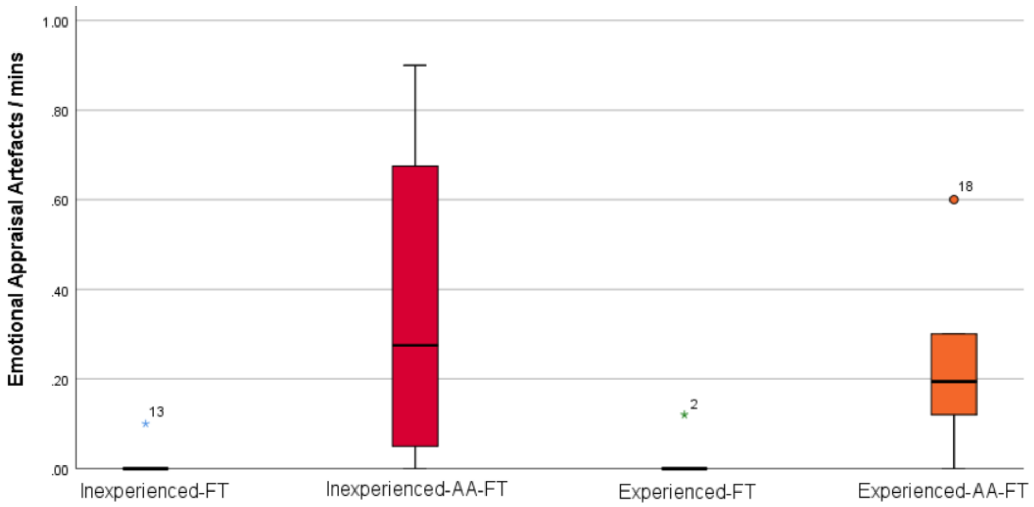

Fig. 17. Emotional appraisal artefacts per minute regarding the *Story 1* authoring task across different groups and experience levels. It is important to keep in mind the low sample size of each group.

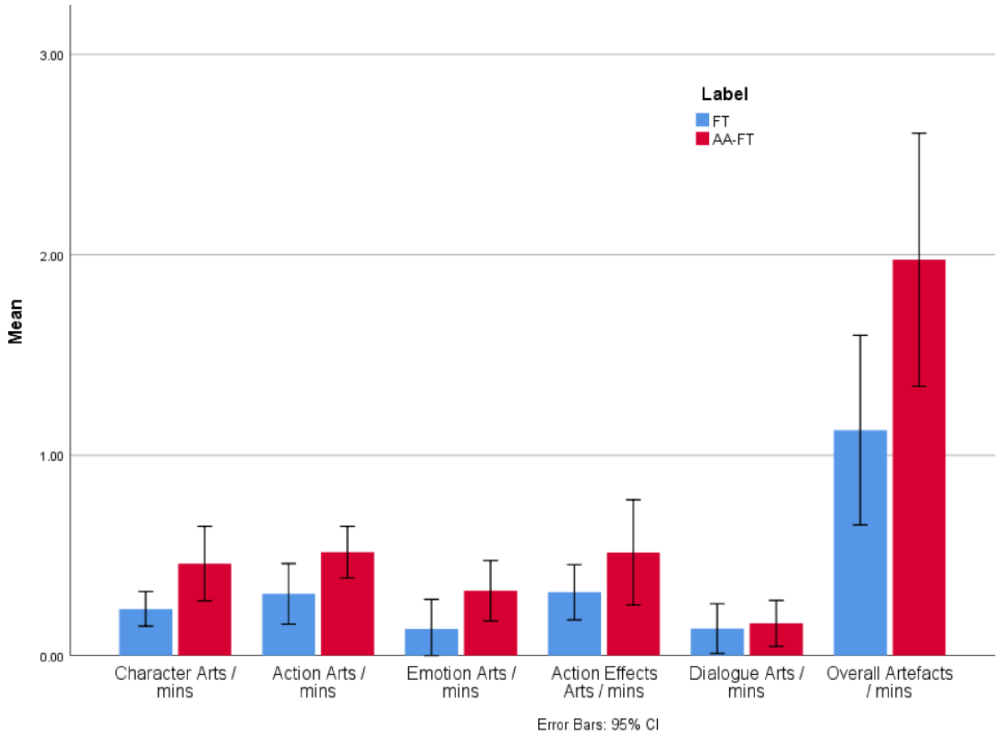

Fig. 18. Artefacts per minute regarding the *Story 2* authoring task across different SIA components. AA-FT artefacts per minute were significantly higher compared to FT in terms of character and internal state, decision-making and emotional appraisal-related and overall concepts.

6.3.2 Authoring Effort—Story 2. On average, participants from the AA-FT group performed the second authoring task in 25.2 minutes ($SD = 8.4$). Participants from FT took a bit longer, averaging 33 minutes ($SD = 11.7$).

In terms of overall artefacts created, the AA-FT group authored, on average, 45.3 artefacts ($SD = 13.9$), while authors from the FT group created 35.75 ($SD = 21.9$) artefacts. Once again, the high variance of the produced scenarios leads us to view the results in terms of artefacts per time.

A Mann-Whitney U independent samples test was performed, and AA-FT participants, on average ($M = 1.97$, $SD = 0.88$), had a significantly higher amount of artefacts created per minute compared to the FT Group ($M = 1.12$, $SD = 0.57$); [$U = 92.500$ $z = 2.144$ $p = 0.03$, $r = 0.46$; $r > 0.3$, medium effect size]. Figure 18 displays the values across the different components within FAtiMA-Toolkit.

The following is the distribution of the participant's level of artefacts produced per minute.

— *Characters and Their Internal States*: AA-FT group ($M = 0.24$, $SD = 0.1$) had a significantly higher level of character-related artefacts per minute compared to FT group ($M = 0.46$, $SD = 0.26$). $U = 97.000$ $z = 2.445$ $p = 0.015$, $r = 0.52$; $r > 0.5$, large effect size].
— *Decision-Making*: AA-FT group ($M = 0.51$, $SD = 0.18$) had a significantly higher level of decision-making-related artefacts per minute compared to FT ($M = 0.32$, $SD = 0.19$). [$U = 91.500$, $z = 2.079$, $p = 0.038$, $r = 0.44$; $r > 0.3$, medium effect size].
— *Emotional Appraisal*: AA-FT group ($M = 0.31$, $SD = 0.22$) had a significantly higher level of emotional appraisal-related artefacts per minute compared to FT ($M = 0.11$, $SD = 0.17$). [$U = 97.000$, $z = 2.466$, $p = 0.014$, $r = 0.52$; $r > 0.5$, large effect size].

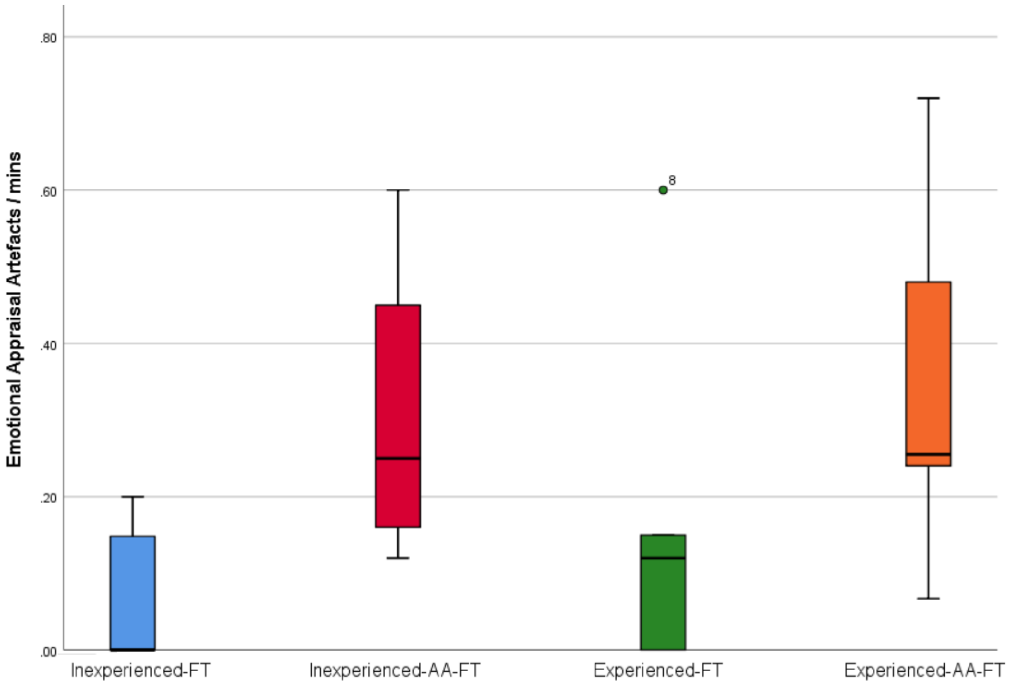

Fig. 19. Emotional appraisal artefacts per minute regarding the *Story 2* authoring task across different groups and experience levels. Once again, it is important to keep in mind the low sample size of each group.

— *World Model*: AA-FT group ($M = 052$, $SD = 0.37$) had a higher level compared to FT ($M = 0.34$, $SD = 0.2$).
— *Dialogues*: AA-FT group ($M = 0.16$, $SD = 0.16$) had a lower level compared to FT ($M = 0.1$, $SD = 0.15$).
— *Overall*: AA-FT participants, on average ($M = 1.97$, $SD = 0.88$), had a significantly higher amount of artefacts created per minute compared to the FT Group ($M = 1.12$, $SD = 0.57$). [$U = 92.500$ $z = 2.144$ p = 0.03, $r = 0.46$; $r > 0.3$, medium effect size].

Regarding the Story 2 task, despite the low amount of participants within each of the subgroups being quite low, an interaction effect was found when comparing inexperienced participants' results across framework versions. Inexperienced participants that used AA-FT to create the second scenario had a significantly higher level of emotional appraisal artefacts per minute compared to inexperienced participants from the FT group. [$U = 24.500$; $z = 2.036$; p = 0.042, $r = 0.61$; $r > 0.5$, large effect size]. Figure 19 captures this.

6.3.3 Model Understandability. Participants were asked to answer five different questions that required them to interpret, understand and use SIA concepts. Out of the five questions, participants could obtain a maximum score of 10 if all of the questions were correct.

Overall, no significant difference was found in the number of correct answers before ($M = 7.96$, $SD = 1.36$) and after ($M = 8.41$, $SD = 1.56$) completing the authoring tasks [p = 0.071]. Between the frameworks, no significant difference was found before; AA-FT ($M = 8.15$; $SD = 0.78$) and FT ($M = 7.92$; $SD = 0.5$), and after; AA-FT ($M = 8.7$; $SD = 0.95$) and FT ($M = 8.17$; $SD = 1.94$).

Looking at the results, in terms of the experience level of participants, as expected, experienced participants achieved higher scores before ($M = 8$; $SD = 1.42$) and after ($M = 8.55$; $SD = 1.5$) the

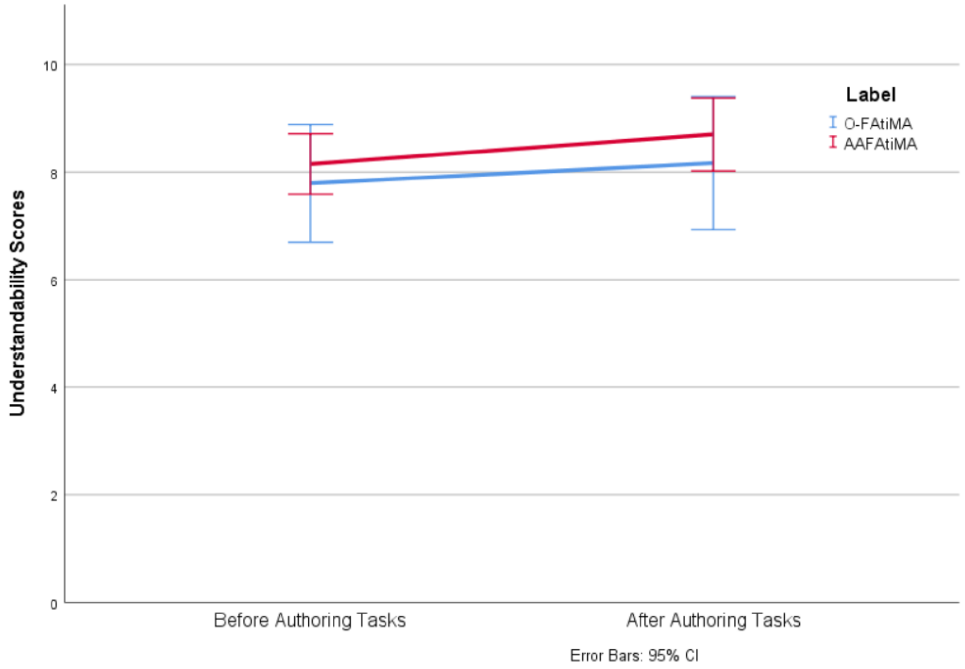

Fig. 20. Participants average overall model understandability scores before and after the authoring tasks.

authoring tasks compared to inexperienced participants pre ($M = 7.84$; $SD = 1.36$) and post ($M = 8.27$; $SD = 1.67$) authoring tasks. In both cases, the difference was not significant. Figure 20 displays these results.

6.3.4 IMI—Perceived Competence. The set of items to assess the perceived competence sub-scale of IMI was tested for their internal consistency, with a Cronbach Alpha of 0.926 for six questions. Each item prompted participants to state their level of agreement, using a Likert 1–7 scale (1-Not At all True, 4-Somewhat True and 7-Very True), with items such as "I think I did pretty well at this activity."

An independent t-test was run on the data with a 95% CI for the mean difference between the experienced and inexperienced groups. It was found that, before the task, experienced participants ($M = 5.36$, $SD = 0.83$) were significantly more confident than inexperienced authors ($M = 3.27$, $SD = 1.35$) [$t(20) = -4.37$, $p < 0.001$]. After the task, while the gap closed and the difference between the experienced ($M = 4.98$; $SD = 1.15$) and inexperienced ($M = 3.57$; $SD = 1.31$) participants were still significant $t(20) = -2.7$, $p = 0.014$).

No significant difference was found between groups that authored using different versions of the toolkit in the average perceived competence before and after the task.

6.3.5 IMI—Enjoyment and Effort. Regarding interest/enjoyment and effort, the corresponding sub-scales of IMI were used. Each used several items to capture specific self-reported measurements. Participants were asked to rate their level of agreement, using a Likert 1–7 scale (1-Not At all True, 4-Somewhat True and 7-Very True), with sentences about their enjoyment and effort. Enjoyment statements such as "This activity was fun to do" and "I would describe this activity as very interesting" and effort-related statements such as "I put a lot of effort into this" and "It was important to me to do well at this task."

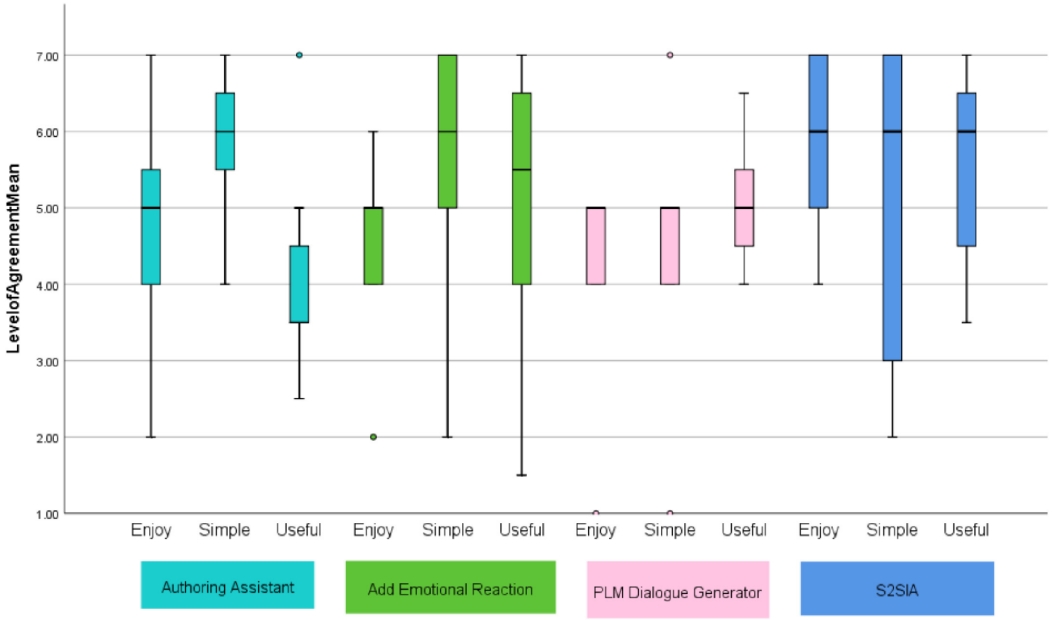

Fig. 21. Values regarding enjoyment, simple to use and usefulness across different AA-FT features.

For interest/enjoyment, the items obtained an internal consistency (Cronbach Alpha) of 0.868, while the effort sub-scale items obtained a lower consistency of 0.743. Still, the values allow us to combine the items (seven regarding interest/enjoyment and five about effort) into a single value.

On average, participants moderately agreed that they enjoyed the activity ($M = 5.06$; $SD = 1.09$) and placed a moderate to high amount of effort in the execution ($M = 5.37$; $SD = 1.07$) of the task. No difference between the subgroups was found.

6.3.6 Feature Usefulness, Enjoyment and Usability. Participants using AA-FT completed an additional round of questions pertaining to their level of agreement of specific statements regarding some of the newly implemented features of AA-FT, using a Likert 1–7 scale (1-Not At all True, 4-Somewhat True and 7-Very True). The items assessed the usefulness, "The _ feature was useful," "The _ feature was a waste of time," enjoyment: "I enjoyed using the _" and how simple the feature was to be used: "The _ Feature was simple to use."

We focused this survey on four key features added to the toolkit, the Story to Scenario Wizard (which uses S2SIA and the Authoring Assistant), the Add Emotional Reaction, the GPT Dialogue Generator and the User Interface Guide. Naturally, due to the time constraint and the nature of the tasks, which had multiple possible solutions, not all of the queried participants were able to use all of these particular features. Thus, in reporting these results, the number of authors is also included.

Regarding *Usefulness*, participants regarded S2SIA, on average, as the most useful feature (N = 9, $M = 5.5$; $SD = 1.27$), followed by the LLM Utterance Generator (N = 5, $M = 5.1$, $SD = 0.96$), along with the Emotional Reaction Helper (N = 9, $M = 5$, $SD = 1.84$) and finally, the UI Guide (N = 7, $M = 4.14$, $SD = 1.46$).

In terms of the statement regarding *enjoyment*, once again, S2SIA was the preferred feature (N = 9, $M = 5.89$, $SD = 1.17$), followed by LLM Utterance Generator (N = 5, $M = 5.1$, $SD = 1.73$), the Interface Guide (N = 9, $M = 4.71$, $SD = 1.46$) and at the end, the Add Emotional Reaction (N = 9, $M = 4.67$, $SD = 1.22$).

Finally, in terms of the most *Simple to use*, the Interface Guide was the top choice (N = 7, $M = 5.88$, $SD = 1.05$) followed closely by the Add Emotional Reaction (N = 9, $M = 5.44$, $SD = 1.67$), afterwards the S2SIA (N = 9, $M = 5.11$, $SD = 197$) with the Generate Utterances being the last one (N = 5, $M = 4.4$, $SD = 2.19$). Figure 21 describes the values and results presented here.

6.3.7 Feature Usage. AA-FT-Toolkit generates a log file that details which features were used and how often. The log is a simple.txt file that saves information when an author uses certain components. A crawler program was created that went through the log file and collected feature usage data. While there is high variance within these results, they still provide valuable insight into the overall authoring experience using the new features.

On average, each AA-FT participant (out of nine samples collected) clicked on the Story to Scenario Wizard 10.6 times ($M = 10.6$; $SD = 12.1$), computed a story 6.8 times ($M = 6.9$, $SD = 9.0$) and accepted its output 3.6 times ($M = 3.6$, $SD = 3.3$).

Regarding the Add Emotional Reaction, of the seven participants, each opened the helper 11.7 times ($M = 11.7$, $SD = 7.7$). The wide array of buttons surrounding the Interface embedded Authoring Assistant led to the highest clicks across nine samples; each participant used it 21.4 times ($SD = 14$). Finally, regarding the Utterances generation by LLMs, the crawler was only able to detect its use by three different participants. On average, each used clicked 12.7 times on the Wizard ($SD = 15.8$) and only accepted LLM's output three times ($SD = 2.2$).

6.4 Discussion

We consider the results presented here to be quite encouraging. Authoring Assisted FAtiMA-Toolkit authors, on average, were able to significantly create more artefacts per minute compared to participants who created similar scenarios using FAtiMA-Toolkit. The effects were more significant in areas where S2SIA and the Authoring Assistant provided more substantial support, suggesting that the developed solution does indeed help to ease the authoring burden.

In the first task, AA-FT authors generated 15 times more emotional appraisal artefacts per minute and almost four (3.7) times more decision-making artefacts per minute compared to the previous version of the toolkit. In the second task, overall, Authoring Assisted FAtiMA-Toolkit participants authored almost twice (1.75) the number of artefacts per minute compared to standard toolkit authors, in particular, three times more emotion and 1.5 action-related artefacts per minute.

The S2SIA feature proved to be quite proficient at automatically generating characters with beliefs and goals, decision-making and emotional appraisal content, especially the latter two components. S2SIA and Authoring-Assistant features such as the Add Emotional Reaction were quite used. One particular finding we would like to highlight is feature usage records suggest that, when using S2SIA, authors often backtrack and try the feature with adjustments to the story. This further proves the importance of auditability and validation steps when using these types of tools.

We looked into possible differences between the levels of experience between groups, and while significant differences were found, we consider that these were not consistent enough to extract meaningful conclusions. In the first task, a significant positive effect was found in terms of the experienced participants using the AA-FT framework, while in the second task, this effect was found in inexperienced participants. The lack of consistency, along with the low sample size, makes it difficult to draw any significant interpretations. However, it is possible that a study with more participants and a higher sample size could lead to interesting results in terms of the interaction effects of AA-FT with experienced and inexperienced authors.

Results on usefulness, enjoyment and simplicity to use suggest further avenues for future work. Furthermore, the levels of each of the IMI sub-scales and the level of understandability of participants solidify our findings. Participants had a high level of effort and moderately high enjoyment across

groups. Additionally, as expected, inexperienced participants' perceived competence levels rose after the authoring tasks while experienced participants' perceived competence decreased. Finally, participants learned from the experience. While the understandability levels were high before the task, using FAtiMA-Toolkit positively affected learning to working with SIA concepts.

6.5 Author Feedback

Along with the experimental results described above, participants were asked to provide feedback regarding their experience with using FAtiMA-Toolkit. Authoring Assisted FAtiMA-Toolkit users were also prompted with an additional question asking for feedback regarding the new features.

The responses were in line with our expectations, both positive and negative. One major issue users tend to have with the toolkit is its graphical interface. "Could use a more appealing UI" and "The UI could be more user-friendly" is the feedback we have received regarding the toolkit almost since its inception. To address this, a redesign that prioritises both visual appeal and user control could reduce the cognitive load associated with the toolkit's current interface [18, 50].

Authors also mentioned a problem with the lack of a direct link between beliefs and conditions of decision rules. "It wasn't always clear what I was doing wrong, some log of what is happening would be very helpful in this situation." and "Please link the variables in the kb and the actions between each panel since it is difficult to remember the exact name when creating the conditions." Emphasising the importance of user feedback, particularly when in designing conversational agents [16], the need for systems to clearly communicate the status of actions [62] and that lack of transparent feedback can lead to user frustration [17, 30].

Regarding the newest version of the toolkit, participants were most pleased with the Authoring Assistant and the Story to Scenario Wizard. S2SIA was particularly impactful when starting new scenarios even for inexperienced authors: "In my opinion, 'Story to Scenario Wizard' was the most useful feature since it easily laid the groundwork for the scenario with just a couple of sentences" and "Story to Scenario Description" was quite interesting and useful because it helped someone who didn't know how to use FAtiMA-Toolkit to at least compile and simulate something. "The Assistant. Complementing with the slides, it really helped to start using the Toolkit." Finally, participants presented some of their frustrations regarding components and features that were not working as intended. However, quite a few of the messages were encouraging and motivating: " speaking of FAtiMA-Toolkit as a whole, I would just like to say to keep up the good work since it's an amazing tool!" and "I really find this tool interesting, will definitely explore it more."

7 Conclusion

SIAs have been successfully and extensively used in a wide range of applications. To enhance the impact and quality of these experiences, it is necessary to rethink the process behind their design. In this work, we tackle human–agent interaction from the perspective of authors, instead of the usual end-user experience.

Our first contribution is an analysis of how different theory-driven frameworks handle and manipulate SIA-related concepts. This overview laid the groundwork for two innovative studies focusing on the authoring experience itself. We examined how authors, with varying levels of experience, perceive and manipulate SIA concepts, and how these concepts are used to create human–agent interaction scenarios through agent-modelling tools, particularly the FAtiMA-Toolkit.

Results suggest that SIA concepts are generally understandable and interpretable for both experienced and inexperienced authors. However, one crucial issue emerged relevant to the Affective Computing field. Despite extensive research on affective computational models, our findings revealed that emotional concepts are not as well understood as previously thought. Both experienced and inexperienced participants underuse or do not prioritise the emotional components of SIAs.

This indicates that improving the understandability of emotional components is essential for the future of agent-modelling tools.

By drawing from the lessons learned from the introspection above, we proposed an innovative approach to the creation of SIAs. Our framework, *Authoring Assisted FAtiMA-Toolkit* leverages the power of data-driven approaches over a theory-grounded agent modelling tool, facilitating the creation of human–agent interaction artefacts, *easing the authoring burden*, without sacrificing the frameworks inherit *understandability and authorial control.*

Authoring Assisted FAtiMA-Toolkit developers are able to use helpers available within the framework to leverage data-driven tools to facilitate the creation of SIA concepts in a multitude of components. *S2SIA—Story to SIA artefacts* allows authors to write a simple description of a story and automatically create a scenario with relevant SIA concepts such as characters, beliefs, goals and actions. The issues identified in our initial analysis led to the development of an *Authoring Assistant* that reinforces the framework's author-friendliness and authorial control.

An evaluation study was conducted to assess the effectiveness of the developed framework. Two groups used different versions of FAtiMA-Toolkit to create two human–agent interaction scenarios based on stories. On average, authors using AA-FT significantly created more artefacts per minute than the ones using the previous version of the tool. In particular, Authoring-Assisted authors generated, at least, three times more emotional appraisal artefacts and, at least, 1.5 times more decision-making-related artefacts per minute compared to the standard FAtiMA-Toolkit authors. Moreover, participants using the tool's new features were still able to understand the concepts and the tasks. In general, participants found these new features useful, simple to use and enjoyable.

The developed framework leverages data-driven approaches to facilitate the creation of affective social agents in a theory-driven agent modelling tool. Together, both S2SIA and the Authoring Assistant contribute towards easing the authoring burden without compromising the framework's understandability and authorial control. In addition to this, their effect, in particular, positively impacted decision-making and emotional components.

The continuation of the work presented in this manuscript has led to a larger focus on the potential of LLMs and use their inherit knowledge together with theory-driven prompting to extract social practices and identify appropriate social affordances for different scenarios [3]. The issues and questions addressed are crucial for advancing the next generation of SIAs and human–agent interaction. We challenge designers of agent modelling tools to reconsider the current state of the authoring experience and computational models of emotion.

References

[1] Rehaf Aljammaz, Elizabeth Oliver, Jim Whitehead, and Michael Mateas. 2020. Scheherazade's Tavern: A prototype for deeper NPC interactions. In *Proceedings of the International Conference on the Foundations of Digital Games*, 1–9.

[2] Prithviraj Ammanabrolu, Wesley Cheung, William Broniec, and Mark O. Riedl. 2020. Automated storytelling via causal, commonsense plot ordering. arXiv:2009.00829. Retrieved from https://doi.org/10.48550/arXiv.2009.00829

[3] Ana Antunes, Joana Campos, Manuel Guimarães, João Dias, and Pedro A. Santos. 2023. Prompting for socially intelligent agents with ChatGPT. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*, 1–9.

[4] Andre Spinola Antunes. 2021. *Creation of a Mental Health Virtual Assistant for College Students*. Master's thesis. Instituto Superior Técnico, Lisbon University.

[5] Ruth S. Aylett, Sandy Louchart, Joao Dias, Ana Paiva, and Marco Vala. 2005. FearNot!–An experiment in emergent narrative. In *Proceedings of the International Workshop on Intelligent Virtual Agents*. Springer, 305–316.

[6] Antoine Bechara, Hanna Damasio, and Antonio R. Damasio. 2000. Emotion, decision making and the orbitofrontal cortex. *Cerebral Cortex* 10, 3 (2000), 295–307.

[7] Christian Becker-Asano and Ipke Wachsmuth. 2010. Affective computing with primary and secondary emotions in a virtual human. *Autonomous Agents and Multi-Agent Systems* 20, 1 (2010), 32–49. DOI: https://doi.org/10.1007/s10458-009-9094-9

[8] Sara Bernardini, Kaśka Porayska-Pomsta, and Tim J. Smith. 2014. ECHOES: An intelligent serious game for fostering social communication in children with autism. *Information Sciences* 264 (2014), 41–60.
[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, Vol. 33, 1877–1901.
[10] Jerome Bruner. 1991. The narrative construction of reality. *Critical Inquiry* 18, 1 (1991), 1–21.
[11] Ferdinand De Coninck, Zerrin Yumak, Guntur Sandino, Remco C. Veltkamp, and B. CleVR. 2019. Non-verbal behavior generation for virtual characters in group conversations. In *Proceedings of the 2019 IEEE International Conference on Artificial Intelligence and Virtual Reality*, 41–49.
[12] Daniel DeKerlegand, Ben Samuel, and Mike Treanor. 2021. Pedagogical challenges in social physics authoring. In *Proceedings of the International Conference on Interactive Digital Storytelling*. Springer, 34–47.
[13] Daniel Clement Dennett. 1987. *The Intentional Stance*. MIT Press.
[14] Joao Dias, Samuel Mascarenhas, and Ana Paiva. 2014. Fatima modular: Towards an agent architecture with a generic appraisal framework. In *Emotion Modeling*. Springer, 44–56.
[15] Joao Dias and Ana Paiva. 2005. Feeling and reasoning: A computational model for emotional characters. In *Proceedings of the Portuguese Conference on Artificial Intelligence*. Springer, 127–140.
[16] Stephan Diederich, Alfred Benedikt Brendel, Stefan Morana, and Lutz Kolbe. 2022. On the design of and interaction with conversational agents: An organizing and assessing review of human-computer interaction research. *Journal of the Association for Information Systems* 23, 1 (2022), 96–138.
[17] Alan Dix. 2004. *Human-Computer Interaction*. Vol. 1. Pearson Education.
[18] Masyura Ahmad Faudzi, Zaihisma Che Cob, Sharul Azim Sharudin, Ridha Omar, and Masitah Ghazali. 2023. The effects of user interface design for mobile learning application on learner's extraneous cognitive load: A conceptual framework. In *Proceedings of the Asian HCI Symposium 2023*, 51–57.
[19] Patrick Gebhard. 2005. ALMA: A layered model of affect. In *Proceedings of the 4th International Joint Conference on Autonomous Agents and Multiagent Systems*, 29–36.
[20] Patrick Gebhard, Martin Klesen, and Thomas Rist. 2004. Coloring multi-character conversations through the expression of emotions. In *Tutorial and Research Workshop on Affective Dialogue Systems*. Springer, 128–141.
[21] Jonathan Gratch, Arno Hartholt, Morteza Dehghani, and Stacy Marsella. 2013. Virtual humans: A new toolkit for cognitive science research. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 35, 215–233.
[22] J. Gratch and S. Marsella. 2003. Modeling coping behavior in virtual humans: Dont worry, be happy. In *Proceedings of 2nd International Joint Conference on Autonomous Agents and Multiagent systems*, 313–320.
[23] J. Gratch and S. Marsella. 2004. A domain-independent framework for modeling emotion. *Journal of Cognitive Systems Research* 5, 4 (2004), 269–306.
[24] Manuel Guimarães, Samuel Mascarenhas, Rui Prada, Pedro A. Santos, and João Dias. 2019. An accessible toolkit for the creation of socio-emotional agents. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, 2357–2359.
[25] Manuel Guimarães, Rui Prada, Pedro A. Santos, João Dias, Arnav Jhala, and Samuel Mascarenhas. 2020. The impact of virtual reality in the social presence of a virtual agent. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, 1–8.
[26] Manuel Guimaraes, Pedro Santos, and Arnav Jhala. 2017. CiF-CK: An architecture for social NPCS in commercial games. In *Proceedings of the 2017 IEEE Conference on Computational Intelligence and Games (CIG '17)*. IEEE, 126–133.
[27] Manuel Guimarães, Pedro Santos, and Arnav Jhala. 2017. Prom week meets Skyrim. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, 1790–1792.
[28] Brent Harrison and Mark O. Riedl. 2016. Towards learning from stories: An approach to interactive machine learning. In *AAAI Workshop on Symbiotic Cognitive Systems*, 746–750.
[29] Arno Hartholt, David Traum, Stacy C. Marsella, Ari Shapiro, Giota Stratou, Anton Leuski, Louis-Philippe Morency, and Jonathan Gratch. 2013. All together now: Introducing the virtual human toolkit. In *Proceedings of the 13th International Conference on Intelligent Virtual Agents*, 368–381.
[30] Morten Hertzum and Kasper Hornbæk. 2023. Frustration: Still a common user experience. *ACM Transactions on Computer-Human Interaction* 30, 3 (2023), 1–26.
[31] Kristin M. Tolle, D. Stewart W. Tansley, and Anthony J. G. Hey. 2011. The Fourth Paradigm: Data-Intensive Scientific Discovery [Point of View]. In *Proceedings of the IEEE* 99, 8 (2011), 1334–1337. DOI: https://doi.org/10.1109/JPROC.2011.2155130
[32] Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing 7, 1 (2017).
[33] Constantin Houy, Peter Fettke, and Peter Loos. 2012. Understanding understandability of conceptual models–What are we actually talking about? In *Proceedings of the International Conference on Conceptual Modeling*. Springer, 64–77.

[34] Intrinsic Motivation Inventory. 1994. Intrinsic motivation inventory (IMI). *The Intrinsic Motivation Inventory, Scale description*, 1–3.
[35] R. Lazarus. 1991. *Emotion and Adaptation*. Oxford University Press. Retrieved from https://www.researchgate.net/publication/232438867_Emotion_and_Adaptation
[36] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al. 2023. Summary of ChatGPT/GPT-4 research and perspective towards the future of large language models. arXiv:2304.01852.
[37] Yukun Ma, Khanh Linh Nguyen, Frank Z. Xing, and Erik Cambria. 2020. A survey on empathetic dialogue systems. *Information Fusion* 64 (2020), 50–70.
[38] S. Marsella and J. Gratch. 2006. EMA: A computational model of appraisal dynamics. *In European Meeting on Cybernetics and Systems Research*.
[39] Stacy C. Marsella and Jonathan Gratch. 2009. EMA: A process model of appraisal dynamics. *Cognitive Systems Research* 10, 1 (2009), 70–90.
[40] Samuel Mascarenhas, Manuel Guimarães, Rui Prada, Pedro A. Santos, João Dias, and Ana Paiva. 2022. FAtiMA Toolkit: Toward an accessible tool for the development of socio-emotional agents. *ACM Transactions on Interactive Intelligent Systems* 12, 1 (2022), 1–30.
[41] Michael Mateas and Josh McCoy. 2013. An architecture for character-rich social simulation. In *Game AI Pro: Collected Wisdom of Game AI Professionals*. Steven Rabin (Ed.), A. K. Peters, Ltd., Natick, MA, 515–530.
[42] Edward McAuley, Terry Duncan, and Vance V. Tammen. 1989. Psychometric properties of the intrinsic motivation inventory in a competitive sport setting: A confirmatory factor analysis. *Research Quarterly for Exercise and Sport* 60, 1 (1989), 48–58. DOI: https://doi.org/10.1080/02701367.1989.10607413
[43] Josh McCoy, Mike Treanor, Ben Samuel, Michael Mateas, and Noah Wardrip-Fruin. 2011. Prom week: Social physics as gameplay. In *Proceedings of the 6th International Conference on Foundations of Digital Games*. ACM, 319–321.
[44] Joshua McCoy, Mike Treanor, Ben Samuel, Aaron A. Reed, Michael Mateas, and Noah Wardrip-Fruin. 2014. Social story worlds with comme il faut. *IEEE Transactions on Computational Intelligence and AI in Games* 6, 2 (2014), 97–112.
[45] Josh McCoy, Mike Treanor, Ben Samuel, Brandon Tearse, Michael Mateas, and Noah Wardrip-Fruin. 2010. Authoring game-based interactive narrative using social games and comme il faut. In *Proceedings of the 4th International Conference & Festival of the Electronic Literature Organization: Archive & Innovate*. Citeseer, Vol. 50.
[46] Joshua McCoy, Mike Treanor, Ben Samuel, Noah Wardrip-Fruin, and Michael Mateas. 2011. Comme il faut: A system for authoring playable social models. In *Proceedings of the 7th Artificial Intelligence and Interactive Digital Entertainment Conference*.
[47] Albert Mehrabian. 1996. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology* 14, 4 (1996), 261–292.
[48] Luís Morais, João Dias, and Pedro A. Santos. 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. Retrieved from https://dl.acm.org/doi/10.1145/3337722.3337746
[49] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and evaluation framework for deeper understanding of commonsense stories. arXiv:1604.01696. Retrieved from https://doi.org/10.48550/arXiv.1604.01696
[50] Don Norman. 2013. *The Design of Everyday Things: Revised and Expanded Edition*. Basic Books.
[51] Andrew Ortony, Gerald L. Clore, and Allan Collins. 1990. *The Cognitive Structure of Emotions*. Cambridge University Press.
[52] Xueni Pan, Marco Gillies, Chris Barker, David M. Clark, and Mel Slater. 2012. Socially anxious and confident men interact with a forward virtual woman: An experimental study. *PloS One* 7, 4 (2012), e32931.
[53] Guglielmo Papagni and Sabine Koeszegi. 2021. A pragmatic approach to the intentional stance semantic, empirical and ethical considerations for the design of artificial agents. *Minds and Machines* 31, 4 (2021), 505–534.
[54] Susanne Patig. 2008. Preparing meta-analysis of metamodel understandability. In *Proceedings of the 1st Workshop on Empirical Studies of Model-Driven Engineering*. Citeseer, 11–20.
[55] Rosalind W. Picard. 2000. *Affective Computing*. MIT Press.
[56] Alexandru Popescu, Joost Broekens, and Maarten Van Someren. 2014. Gamygdala: An emotion engine for games. *IEEE Transactions on Affective Computing* 5, 1 (2014), 32–44.
[57] Shrimai Prabhumoye, Chris Quirk, and Michel Galley. 2019. Towards content transfer through grounded text generation. arXiv:1905.05293. Retrieved from https://doi.org/10.48550/arXiv.1905.05293
[58] Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. 2018. Modeling naive psychology of characters in simple commonsense stories. arXiv:1805.06533. Retrieved from https://doi.org/10.48550/arXiv.1805.06533
[59] Najmeh Sadoughi, Yang Liu, and Carlos Busso. 2017. Meaningful head movements driven by emotional synthetic speech. *Speech Communication* 95 (2017), 87–99.

[60] Klaus R. Scherer. 1999. Appraisal theory. In *Handbook of Cognition and Emotion*. Tim Dalgleish and Mick J. Power (Eds.), John Wiley & Sons Ltd, 637–663. DOI: https://doi.org/10.1002/0470013494.ch30
[61] John R. Searle. 2008. *Mind, Language and Society: Philosophy in the Real World*. Basic Books.
[62] Ben Shneiderman and Catherine Plaisant. 2010. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Pearson Education India.
[63] Feng-Guang Su, Aliyah R. Hsu, Yi-Lin Tuan, and Hung-Yi Lee. 2019. Personalized dialogue response generation learned from monologues. In *INTERSPEECH*, 4160–4164. Retrieved from https://dl.acm.org/doi/10.5555/523237
[64] Pradyumna Tambwekar, Murtaza Dhuliawala, Lara J. Martin, Animesh Mehta, Brent Harrison, and Mark O. Riedl. 2018. Controllable neural story plot generation via reinforcement learning. arXiv:1809.10736. Retrieved from https://doi.org/10.48550/arXiv.1809.10736
[65] Mike Treanor, Josh McCoy, and Anne Sullivan. 2016. A framework for playable social dialogue. In *Proceedings of the 12th Artificial Intelligence and Interactive Digital Entertainment Conference*.
[66] Wim Westera, Rui Prada, Samuel Mascarenhas, Pedro A. Santos, João Dias, Manuel Guimarães, Konstantinos Georgiadis, Enkhbold Nyamsuren, Kiavash Bahreini, Zerrin Yumak, et al. 2020. Artificial intelligence moving serious gaming: Presenting reusable game AI components. *Education and Information Technologies* 25, 1 (2020), 351–380. DOI: https://doi.org/10.1007/s10639-019-09968-2
[67] Jing Xu, Arthur Szlam, and Jason Weston. 2021. Beyond goldfish memory: Long-term open-domain conversation. arXiv:2107.07567. Retrieved from https://doi.org/10.48550/arXiv.2107.07567
[68] Jian Yang, Gang Xiao, Yulong Shen, Wei Jiang, Xinyu Hu, Ying Zhang, and Jinghui Peng. 2021. A survey of knowledge enhanced pre-trained models. arXiv:2110.00269. Retrieved from https://doi.org/10.48550/arXiv.2110.00269

Received 31 December 2023; revised 27 December 2024; accepted 30 December 2024