Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)
Special Track on Human-Centred AI

Emergence of Social Norms in Generative Agent Societies: Principles and Architecture

Siyue Ren[1] , **Zhiyao Cui**[2] , **Ruiqi Song**[2] , **Zhen Wang**[1,2,3] and **Shuyue Hu**[4]
[1]School of Mechanical Engineering, Northwestern Polytechnical University
[2]School of Cybersecurity, Northwestern Polytechnical University
[3]School of Artificial Intelligence, OPtics and ElectroNics(iOPEN), Northwestern Polytechnical University
[4]Shanghai Artificial Intelligence Laboratory
{rensiyue, zhiyao, songruiqi}@mail.nwpu.edu.cn, w-zhen@nwpu.edu.cn, hushuyue@pjlab.org.cn

Abstract

Social norms play a crucial role in guiding agents towards understanding and adhering to standards of behavior, thus reducing social conflicts within multi-agent systems (MASs). However, current LLM-based (or generative) MASs lack the capability to be normative. In this paper, we propose a novel architecture, named *CRSEC*, to empower the emergence of social norms within generative MASs. Our architecture consists of four modules: Creation & Representation, Spreading, Evaluation, and Compliance. This addresses several important aspects of the emergent processes all in one: (i) where social norms come from, (ii) how they are formally represented, (iii) how they spread through agents' communications and observations, (iv) how they are examined with a sanity check and synthesized in the long term, and (v) how they are incorporated into agents' planning and actions. Our experiments deployed in the Smallville sandbox game environment demonstrate the capability of our architecture to establish social norms and reduce social conflicts within generative MASs. The positive outcomes of our human evaluation, conducted with 30 evaluators, further affirm the effectiveness of our approach. Our project can be accessed via the following link: https://github.com/sxswz213/CRSEC.

1 Introduction

In human societies, social norms, which are standards of behavior shared within a social group [Sherif, 1936], have shaped almost every aspect of our daily life, from the language we speak and the etiquette we drive to the amount we tip. Without social norms, people may feel confused about how to behave appropriately in social situations and consequently social conflicts may arise [Lewis, 1969]. Over the past decades, the study of social norms has attracted much interest in a variety of disciplines, such as economics [Young, 2015], cognitive science [Hawkins *et al.*, 2019], complex system science [Centola *et al.*, 2018], and computer science [Morris-Martin *et al.*, 2019]. Across these studies, a central question is: how do social norms spontaneously emerge from social interactions of humans or agents?

This paper studies the emergence of social norms within a generative multi-agent system (MAS), i.e. a system of agents that are powered by large-language models (LLMs). The deployment of MASs in real-world situations raises the need for these systems to be normative—the capability of empowering agents to understand certain standards of behavior and behave appropriately according to the standards [Boella *et al.*, 2008; Criado *et al.*, 2011]. Imagine that agents within a system interact with other agents or humans to accomplish some tasks; for the system to be truly accepted and embraced by humans, such a capability will be crucial, as it can reduce conflicts within systems, enable more effective coordination among agents (potentially including humans), and allow humans to anticipate the system's behaviors—a key means to improve human trust in the system [Awad *et al.*, 2018; Ajmeri *et al.*, 2020; Chugunova and Sele, 2022].

Since LLMs are trained on extensive corpora of human text, it is not surprising that they may inherently embed social norms [Schramowski *et al.*, 2022; Guo *et al.*, 2023]. One might thus challenge the importance of fostering the emergence of social norms within generative MASs. While LLMs can capture social norms, it has also been shown that LLMs do not adequately understand social norms, especially the culture-specific ones [Ramezani and Xu, 2023; Hämmerl *et al.*, 2022]. This deficiency can provoke conflicts among generative agents, particularly when their base LLMs are trained on text corpora from diverse cultural backgrounds. Moreover, as generative agents increasingly become more personalized (such as functioning as personal assistants) and represent humans in social situations, it is natural to expect that these agents, reflecting the values and preferences of their human users, will encounter social conflicts similar to those experienced by humans. To tackle these challenges, approaches must go beyond merely embedding LLMs with human norms or aligning them to such norms [Liu *et al.*, 2024; Li *et al.*, 2024]; rather, they should also be able to foster the emergence of social norms within generative MASs so that generative agents can establish their own standards of behavior out of their interactions and adhere to these standards to address those conflicts.

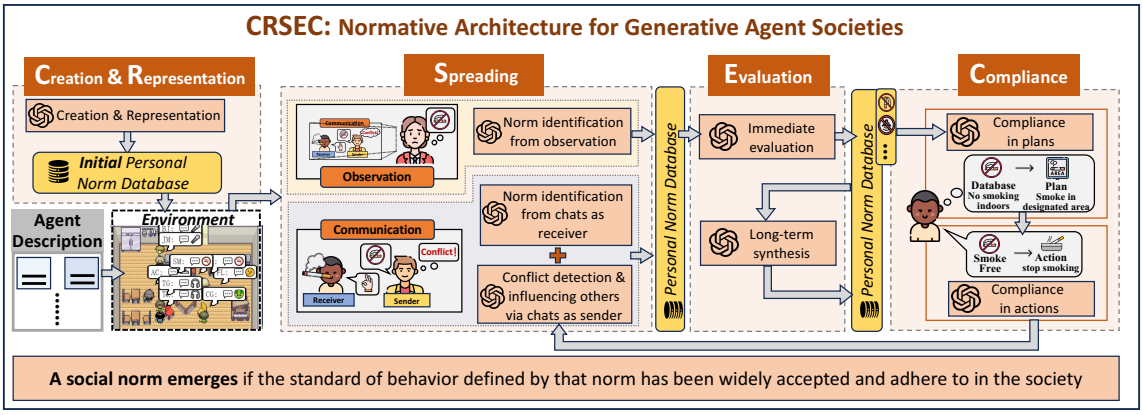

Figure 1: CRSEC: our architecture for the emergence of social norms in generative agent societies. Initially, by the *Creation & Representation* module, norm entrepreneurs create their personal norms and store them into their databases. By the *Spreading* module, some agents proactively influence others to adopt their personal norms through initiating communication with others, while others can identify those norms from their chats and observations. The identified norms then undergo an immediate evaluation in the *Evaluation* module. The *Compliance* module enables agents to generate plans and actions, with the norms bearing in mind. The normative actions, in turn, can influence other agents' observations and thus reinforce the spreading of norms. In addition, from time to time, agents perform long-term synthesis to keep their personal norms compact and concise.

How can we empower generative MASs with the capability to foster social norm emergence? We argue that the key is to instigate an emergent process—generative agents, starting with initially only a few adopting certain standards of behavior, influence others and propagate these standards, ultimately resulting in widespread acceptance and adherence of these standards across the system. Recent work has shown that generative MASs can reproduce believable social behaviors (such as spreading invitations to a party) [Park *et al.*, 2023], achieve multi-agent cooperation surpassing conventional methods [Zhang *et al.*, 2024], and collaboratively solve complex tasks (such as automatic code generation) [Hong *et al.*, 2023]. While these systems have demonstrated the potential of leveraging LLMs in MAS research, the emergence of social norms remains largely unaddressed in existing studies, primarily because they typically focused on fully cooperative tasks—where agents' values, preferences, or objectives align, thereby generally preventing social conflicts and voiding the need for social norms.

Fortunately, the extensive and multidisciplinary literature on social norm emergence can offer a wealth of resources for inspiration. For example, some studies may focus on norm representation [Dignum, 1999; Ågotnes *et al.*, 2009], some may delve into norm compliance and enforcement [Modgil *et al.*, 2009; Villatoro *et al.*, 2011; Mahmoud *et al.*, 2015], and others may explore norm learning [Sen and Airiau, 2007; Beheshti *et al.*, 2015; Hu and Leung, 2017; Hu *et al.*, 2019]. That said, these studies cannot provide a direct solution for two key reasons. First, historically, they have not been able to harness the strength of LLMs. Second, they typically focused on isolated aspects of the emergent process, consequently leaving the tangible implementation that integrates various aspects as an open problem [Savarimuthu and Cranefield, 2011; Haynes *et al.*, 2017].

In this paper, we propose, to our knowledge, the first normative architecture for generative MASs. Our architecture, abbreviated as **CRSEC**, consists of four modules: **C**reation & **R**epresentation, **S**preading, **E**valuation, and **C**ompliance. This architecture not only fosters the emergence of social norms within generative MASs, but also addresses the open problem of actualizing various aspects of the emergent process into an operational framework. Specifically, through the Creation & Representation module, norm entrepreneurs (agents who actively campaign norms) can generate their own personal standards of behavior (or personal norms), and these standards are formally represented and stored in their databases. Through the Spreading module, some agents influence others to adopt the standards via communication and actual behaviors, while others can identify these standards by reflecting on their conversations and observations. With the Evaluation module, agents perform a sanity check to decide whether they accept certain standards as their own personal norms, and, from time to time, synthesize their personal norms to keep the norms compact and concise. Lastly, the Compliance module raises agents' awareness of their personal norms, encouraging them to generate plans and take actions in line with the norms. An overview of our architecture is shown in Figure 1.

To verify if and how our architecture leads to the emergence of social norms within generative MASs, we ran our experiments on the Smallville sandbox game environment [Park *et al.*, 2023], and simulated the scenarios where initially agents have conflicts in their values and preferences. We show that social norms always emerge in multiple independent runs of our experiments, leading to 100% of agents accepting some standards of behaviors as their personal norms and complying with these norms in their plans and actions; moreover, as social norms emerge, social conflicts almost vanish. Moreover, we observe that conversations and thoughts drive the emergence of social norms, and descriptive norms are harder to establish than injunctive norms, yet norm entrepreneurs can shape their emergence. For a better understanding, we additionally present a case study to illustrate how a seasoned smoker in Smallville's environ-

ment has been gradually persuaded to accept "no smoking indoors" as his own personal norm, and eventually even stepped forward to remind another agent upon noticing that agent's breach of the norm. Finally, we present the results of our human evaluation, which involved 30 evaluators, to gauge the effectiveness of our architecture from a human perspective. The feedback gathered from the questionnaires reflects an overall positive evaluation. Furthermore, interviews conducted after the questionnaires shed light on aspects that humans consider important for the emergence of social norms and suggest potential directions for future work.

2 Principles and Architecture

In this section, we illustrate the principles behind our CRSEC architecture and present its four modules. Due to the lack of space, we flesh out the prompts for the LLM-based operations of this work in Appendix B (which is only available in the arXiv version of the paper).

2.1 Creation and Representation

The Creation and Representation module of our architecture addresses the questions of where social norms come from and how they can be formally represented. In human society, social norms are usually shaped by norm entrepreneurs, who actively influence and persuade others to alter their behaviors in accordance with the entrepreneurs' personal standards of behavior (or personal norms) [Sunstein, 1996]. Personal norms typically flow from one's values [Schwartz, 1973], and would become social norms if they were to be widely adopted by other members of a social group. According to [Cialdini *et al.*, 1991], there are two types of (personal or social) norms: (i) descriptive ones that reflect what most people typically do in a given situation, and (ii) injunctive ones that dictate what ought or ought not to be done in a given situation. For example, the common practice of shaking hands upon meeting someone is descriptive; in contrast, no smoking indoors is injunctive.

In this work, we consider a generative agent to be a norm entrepreneur if the agent, initially, possesses some personal norms and is interested in influencing others to adopt its personal norms. Formally, we represent a personal norm with a quintuple $n = \langle c, u, \alpha \in \{\texttt{'des'}, \texttt{'inj'}\}, s_{act} \in \{\texttt{T}, \texttt{F}\}, s_{val} \in \{\texttt{T}, \texttt{F}\}\rangle$. Here, c represents the personal norm in natural language, e.g, "no smoking indoors"; u is the utility that distinguishes mediocre from important personal norms, with a higher score indicating that the agent believes the standard of behavior to be more important; α denotes the type of a personal norm, with `'des'` being descriptive and `'inj'` being injunctive; s_{act} and s_{val} are Boolean variables signifying if the personal norm is activated and valid, respectively. By default, personal norms generated in this module are activated ($s_{act} = \texttt{T}$) and valid ($s_{val} = \texttt{T}$). In the rest of the paper, we say that a personal norm is qualified if it is both activated and valid, for simplicity.

A distinct feature of generative agents is that by using natural language that mimics how one typically describes humans, these agents can exhibit characteristics and personalities in alignment with the agent description [Shanahan *et al.*, 2023]. Recall that agents' values or preferences typically vary, as we analyze in the introduction. To ensure that the created personal norms are consistent with norm entrepreneurs' agent descriptions, we instruct LLMs through prompts to create these norms based on norm entrepreneurs' agent descriptions. Let $\mathcal{G}$ denote an agent description, and $\mathcal{P}$ denote a set of created personal norms. We represent this LLM-based operation by $\mathcal{P} \leftarrow \texttt{CreateNorm}(\mathcal{G})$. This operation not only generates personal norms in natural language, but also classifies a newly formed personal norm (i.e., deciding the value of α), and also assesses the utility u of that norm on a scale of 1 to 100. Once created, personal norms are stored in each norm entrepreneur's personal norm database.

For clarity, we say that a generative agent is an ordinary agent if it is not a norm entrepreneur. Note that not only norm entrepreneurs but also ordinary agents maintain their own personal norm databases. This is because ordinary agents do not generate personal norms through this module though, they may acquire personal norms over time through the Spreading and Evaluation modules, which will be presented in subsequent sections.

2.2 Spreading

The Spreading module of our architecture helps certain standards of behavior gain widespread acceptance and ultimately evolve into social norms. In particular, we consider two key mechanisms through which norms spread in generative MASs: communication and observation.

Communication between Agents

Generative agents are well known for their capability to generate human-like conversations [Clark *et al.*, 2021]. It is thus natural to consider spreading norms by leveraging such a capability. To achieve this, we consider two perspectives: a sender's perspective and a receiver's perspective.

The Sender's Perspective. In human societies, the desire to resolve social conflicts has driven the emergence of numerous social norms [Nyla R. Branscombe, 2022]. Inspired by this, we instruct each generative agent (a sender) to detect if there are any observations of other agents' behaviors that conflict with its personal norms. Let $\mathcal{O}_{\mathcal{S}}$ be the text description of the sender's observations of the environment, and $\mathcal{P}_{\mathcal{S}}$ be the sender's set of qualified personal norms in its database. We represent this LLM-based operation by $\mathcal{Y}_{\texttt{conflict}} \in \{\texttt{T},\texttt{F}\} \leftarrow \texttt{DetectConflict}(\mathcal{O}_{\mathcal{S}}, \mathcal{P}_{\mathcal{S}})$. Note that initially, since only norm entrepreneurs have their own personal norms, ordinary agents will detect no conflicts. However, as time evolves, ordinary agents may also develop their personal norms, and thus conflicts may be detected. Once a conflict is detected (i.e., $\mathcal{Y}_{\texttt{conflict}} = \texttt{T}$), then the sender will decide whether to proactively step in and start a conversation in order to influence others and propagate its personal norms. Intuitively, if the sender is a norm entrepreneur, then it will start a conversation without doubt, as it is interested in influencing others. However, if the sender is an introverted, ordinary agent, it may not start a conversation. Thus, for better autonomy, we let the sender decide based on its agent description $\mathcal{G}_{\mathcal{S}}$, and represent this LLM-based operation by $\mathcal{Y}_{\texttt{talk}} \in \{\texttt{T},\texttt{F}\} \leftarrow \texttt{DecideToTalk}(\mathcal{G}_{\mathcal{S}})$.

The Receiver's Perspective. We consider that when being involved in a conversation, a generative agent (a receiver) will reflect on the conversation and discern information regarding norms (or normative information for short). Let $\mathcal{T}_{\mathcal{S}\to\mathcal{R}}$ denote a conversation between a sender and a receiver. This LLM-based operation can be represented by $\bar{n}_{\mathcal{R}} \leftarrow \texttt{IdentifyNormativeInformation}(\mathcal{T}_{\mathcal{S}\to\mathcal{R}})$, where $\bar{n}_{\mathcal{R}}$ represents normative information, which includes natural language describing certain standard of behavior, and the type of the standard (whether it is descriptive or injunctive), as well as the utility on a scale of 1 to 100 indicating its importance. Here, we also store normative information in the personal norm database, but set their states to be deactivated and invalid ($s_{act} = \texttt{F}, s_{val} = \texttt{F}$) to distinguish them from the qualified personal norms. Initially, ordinary agents are likely to act as receivers. However, over time, norm entrepreneurs may also become receivers, as ordinary agents can in turn influence entrepreneurs, after they develop their own personal norms.

Observation from Others' Behavior

Observation has long been recognized as a key mechanism for humans and agents to learn norms [Nakamaru and Levin, 2004; Shettleworth, 2009; Beheshti and Sukthankar, 2014; Paiva *et al.*, 2018]. Recent work has shown that generative agents can generate thoughts from the text description of their observations [Park *et al.*, 2023; Lin *et al.*, 2023]. Let $\mathcal{O}_{\mathcal{A}}$ denote the text description of observations, and $\mathcal{M}_{\mathcal{A}}$ denote the generated thoughts. This LLM-based operation can be denoted by $\mathcal{M}_{\mathcal{A}} \leftarrow \texttt{GenerateThought}(\mathcal{O}_{\mathcal{A}})$, and it can be achieved by modules that generate thoughts in existing studies. Leveraging on this, we prompt generative agents to identify normative information from the generated thoughts. We represent this LLM-based operation by $\bar{n}_{\mathcal{A}} \leftarrow \texttt{IdentifyNormativeInformation}(\mathcal{M}_{\mathcal{A}})$, where $\bar{n}_{\mathcal{A}}$ represents normative information. Note that this operation is similar to $\texttt{IdentifyNormativeInformation}(\mathcal{T}_{\mathcal{S}\to\mathcal{R}})$, as both these operations can be viewed as a kind of text summarizing tasks. Once generated, the normative information $\bar{n}_{\mathcal{A}}$ is also stored in the personal norm database with the deactivated and invalid state ($s_{act} = \texttt{F}, s_{val} = \texttt{F}$).

2.3 Evaluation

The Evaluation module of our architecture serves two purposes: (i) it evaluates the normative information passed from the Spreading module, and (ii) it synthesizes the qualified personal norms to keep them compact and concise.

Immediate Evaluation

The normative information in the Spreading module, once generated, will be immediately evaluated in the Evaluation module. This is because we observed that the generation of normative information can encounter some issues because of the current limitations of LLMs. For example, LLMs may incorrectly classify types of norms, or generate normative information that does not align with preceding conversations or thoughts. Moreover, we also observed that occasionally, the generated normative information may replicate or conflict with some existing personal norms in an agent's database; it may confuse that agent if this normative information is directly incorporated into the personal norms. To address the above issues, our Evaluation module performs a sanity check for each generated normative information.

Specifically, this consists of four steps. Let $\bar{n}$ be a piece of normative information generated in the Spreading module. The first step examines if $\bar{n}$ is consistent with its preceding conversation or thought, i.e., $\mathcal{Y}_{\texttt{consistent}} \in \{\texttt{T}, \texttt{F}\} \leftarrow \texttt{CheckConsistency}(\bar{n}, q)$, where $q = \mathcal{T}_{\mathcal{S}\to\mathcal{R}}$ if it is generated from the conversation, and $q = \mathcal{M}_{\mathcal{A}}$ if it is generated from the thought. The second step excludes duplication by checking if $\bar{n}$ already exists in the set $\mathcal{P}$ of qualified personal norm, i.e., $\mathcal{Y}_{\texttt{unique}} \in \{\texttt{T}, \texttt{F}\} \leftarrow \texttt{CheckDuplication}(\bar{n}, \mathcal{P})$. Next, we aim to examine if LLMs have incorrectly classified types of norms, i.e., $\mathcal{Y}_{\texttt{type}} \in \{\texttt{T}, \texttt{F}\} \leftarrow \texttt{CheckType}(\alpha)$. Last, we examine if $\bar{n}$ conflicts with any existing qualified personal norm, i.e., $\mathcal{Y}_{\texttt{conflictfree}} \in \{\texttt{T}, \texttt{F}\} \leftarrow \texttt{CheckConflict}(\bar{n}, \mathcal{P})$. Any normative information that yields a false value in one of the above four steps will not pass this sanity check, and will remain deactivated and invalid. Only those that pass the sanity check will become qualified personal norms.

Long-term Synthesis

Over time, as agents accumulate more qualified personal norms, they accept a broader range of standards of behavior, which could potentially limit their liberty. Morales et al. [2013; 2015] suggested that for better agent liberty, it would be beneficial to synthesize norms into a compact and concise set of possibly more abstract ones. Inspired by this, we prompt each generative agent to start a synthesis within its personal norm database if the sum of the utility of its qualified personal norms exceeds a certain threshold.

This synthesis consists of three steps. First, the agent categorizes its qualified personal norms, and generates a theme for each category to justify the categorization; this can be represented by an LLM-based operation $\{(\mathcal{Q}, k)\} \leftarrow \texttt{ClassifySpecificNorms}(\mathcal{P})$, where $\mathcal{Q}$ denotes a subset of qualified personal norms, and k is the associated theme. Then, we prompt the agent to generate an abstract personal norm for each subset based on the principles of compactness and conciseness. We represent this operation by $n' \leftarrow \texttt{GenerateAbstractNorm}(\mathcal{Q}, k)$. Note that the output of this operation includes the natural language description of the abstract personal norm n' and its type, but it excludes its utility. Rather, the utility is determined by calculating the weighted average of the utilities associated with all personal norms within that subset. The weights used in this calculation are also part of the operation's output. Lastly, each generated abstract norm will be immediately evaluated through the sanity check mentioned in the last paragraph. If an abstract personal norm successfully passes the sanity check, then it will become qualified and all the personal norms within that subset will be deactivated ($s_{act} = \texttt{F}, s_{val} = \texttt{T}$).

2.4 Compliance

The Compliance module of our architecture raises agents' awareness of personal norms in their behaviors. Note that with such an awareness, agents can choose to comply with

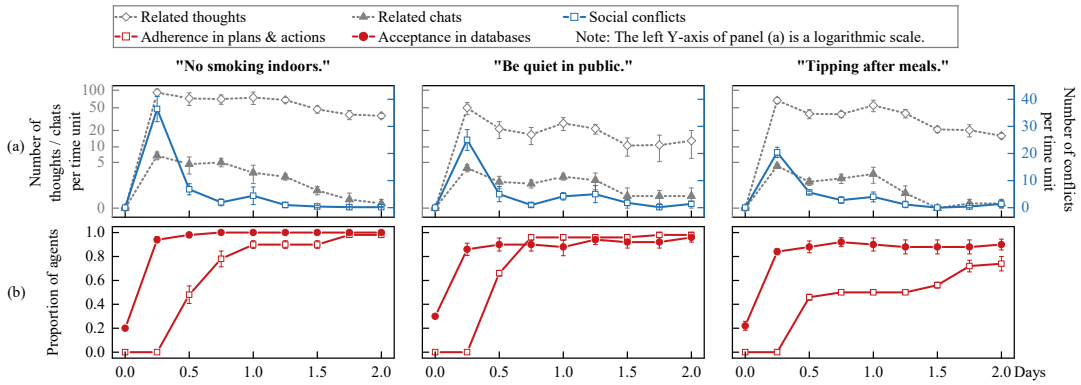

Figure 2: The evolution of generative MASs. Panel (a) depicts the evolution of the number of social conflicts, thoughts and chats over time. Panel (b) illustrates the emergent process of social norms in terms of (i) the proportion of agents that have accepted a standard of behavior as their personal norms in their databases, and (ii) the proportion of agents that have adhered to a standard of behavior in their plans and actions.

the norms or not, thereby granting them with greater autonomy [Conte *et al.*, 1998; Criado *et al.*, 2011]. We design this module focusing on two sub-components: (i) compliance in planning, and (ii) compliance in actions.

Compliance in Planning

Plans describe a sequence of actions for agents. Recent work has shown that generative agents are good at planning towards some goals; this ensures that agents' behaviors are consistent over time [Wang *et al.*, 2023; Lin *et al.*, 2023]. Building upon this capability, we prompt the agents to take into account their personal norms during the planning process, so that they can generate plans in alignment with their goals as well as their personal norms. Let l_i denote a plan (e.g. `10:30 am to 11:00 am: Have a light breakfast`), and $\mathcal{L}_{\texttt{plan}}$ denote a list of plans (e.g. for every hour in a day). The planning process of our architecture can be represented by $\mathcal{L}_{\texttt{plan}} \leftarrow \texttt{GenerateNormativePlans}(\mathcal{C}, \mathcal{P})$, where the inputs are the current goals $\mathcal{C}$ and the set $\mathcal{P}$ of qualified personal norms.

Compliance in Actions

After generating plans, agents proceed to break down each plan into a series of more detailed actions and carry them out. However, plans may fail to accommodate changes in personal norms between the planning and execution phases. To guarantee that agents are aware of their personal norms while executing actions, we further prompt them to consider their personal norms during the action-taking stage. Let $\mathcal{L}_{\texttt{action}}$ denote a list of actions. It is generated based on a plan l_i, the agent's qualified norm set $\mathcal{P}$, and its agent description $\mathcal{G}$: $\mathcal{L}_{\texttt{action}} \leftarrow \texttt{GenerateNormativeActions}(l_i, \mathcal{P}, \mathcal{G})$.

3 An Experimental Study

Our experimental study aims to answer three questions: (i) Do social norms emerge in generative MASs empowered by our architecture? (ii) If so, what are the characteristics of such an emergent process? (iii) How well does our architecture perform from a human perspective? We outline the experimental settings in Section 3.1. We answer the first two questions in Section 3.2, and the last question in Section 3.3.

3.1 Experimental Settings

Our experiments were conducted in Park et al. [2023]'s Smallville sandbox game environment, which is arguably the most well-known environment for generative MASs. This environment offers a variety of scenarios where LLM-based agents can exhibit human-like behaviors, including observation, interaction with others, planning, and action execution. In our setup, there were 10 generative agents, including 3 norm entrepreneurs and 7 ordinary agents. To simulate scenarios where individuals can have conflicts in values or preferences, we considered that ordinary agents' agent descriptions exhibited diverse inclinations: some favored smoking in public, speaking loudly, or supporting a tipping culture, whereas others held opposite preferences. For norm entrepreneurs' agent descriptions, we considered all of them to favor "no smoking indoors" and "be quiet in public". However, since whether to tip varies across cultures, we considered two of them to support tipping while one did not. In addition to these preferences, each agent's agent description also included its name, personality, occupation, short-term goal, and social relationships with other agents, etc. Details of agent descriptions and experimental parameters, such as the number of initial personal norms and the threshold for starting a synthesis in the Evaluation module, are provided in Appendix A.

Our implementation utilized GPT-3.5 and GPT-4. Using the same experimental setup, we repeated our experiments for 5 runs. To be time-efficient and cost-efficient, we focused on the scenario "Hobbs Café" (as visualized at the bottom left corner of Figure 1) and let the experiments continue for 2 days in the Smallville environment. Each run costs more than $500 dollars and about 7 days to complete. The GitHub repository for our project can be accessed via the following link: https://github.com/sxswz213/CRSEC.

3.2 Emergent Phenomena of Social Norms

The emergence of a social norm is typically measured by whether the standard of behavior defined by that norm has been widely accepted and adhered to by a significant majority. In Figure 2, we visualize the evolution of our generative MASs from several perspectives: (i) the number of social conflicts among agents, (ii) the number of generated thoughts or conversations that are related to certain standards of be-

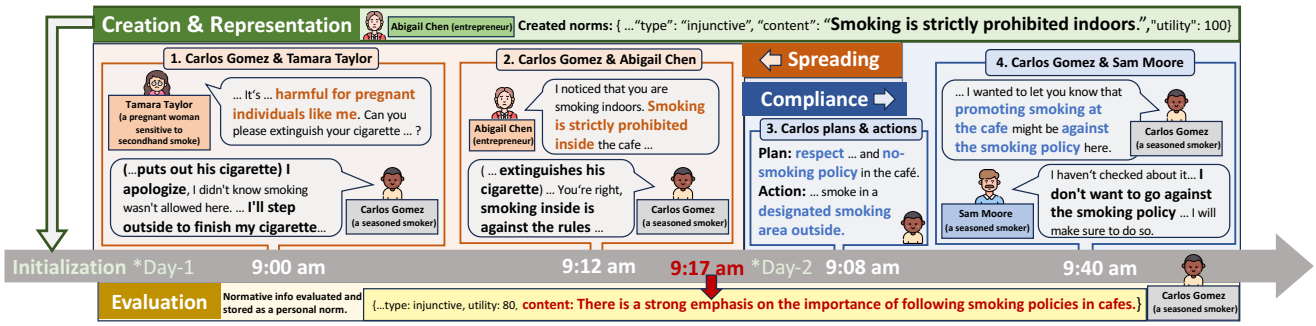

Figure 3: A case study illustrating how a seasoned smoker has gradually adopted "no smoking indoors" as his personal norm.

havior, (iii) the proportion of agents that have incorporated these standards into their personal norm databases as qualified social norms, and (iv) the proportion of agents that have complied with those standards in their behaviors (plans and actions). More findings are elaborated in Appendix D.

Key Findings. *Social norms always emerge.* Our most significant finding is that the social norms "no smoking indoors", "be quiet in public", and "tipping after meals" have always emerged across all five independent runs in our experiments. This emergence is characterized by most agents not only adopting these standards of behavior as their qualified personal norms, but also adhering to these standards in their planning and actions. In particular, at the end of Day 2 in the Smallville environment, 100% of agents have adopted and adhered to the injunctive norms "no smoking indoors" and "be quiet in public". Moreover, we observe that norms, such as "maintain a healthy environment", can also emerge spontaneously even if they are neither exhibited in agent descriptions nor initially created by norm entrepreneurs as personal norms.

Social conflicts almost vanish as social norms emerge. We note that with the emergence of social norms, the number of social conflicts among generative agents exhibits a generally decreasing trend, despite the surge in the early stage. That surge is largely attributed to the inherent conflicts in the values and preferences of agents given our experimental setup; in the beginning, when agents started to interact with each another, their differing values and preferences became apparent and naturally led to conflicts in their interactions. Over time, however, these conflicts significantly reduced as agents gradually developed social norms to resolve them.

Conversations and thoughts drive the emergence of social norms. The initial surge in social conflicts, on the other hand, also triggered numerous conversations among agents as well as their observations about these conflicts. Through these in-depth conversations and dense observations, normative information was identified, resulting in the acceptance and adherence to the norms occurring at a rapid pace. Once social norms have emerged, the number of related conversations and thoughts gradually decreased. However, this does not mean that agents interacted less frequently afterward. Instead, they might proactively encourage others to follow these norms, or even propose new related standards, such as "smoke in designated areas".

Descriptive norms are harder to establish than injunctive norms, yet norm entrepreneurs can shape their emergence. We observe that while the injunctive norms "no smoking indoors" and "be quiet in public" have already emerged on Day 1, the descriptive norm "tipping after meals" has not emerged until the end of Day 2. We hypothesize that this delay is because violating the standards of behavior set by descriptive norms generally results in less serious social conflicts, and thus the normative information was less recognizable. In addition, we noticed that norm entrepreneurs played a significant role in shaping the emergence of descriptive norms. In our setup, initially, there was an equal number of agents supporting and against tipping; however, out of the five agents favoring tipping, two of them were norm entrepreneurs. Despite the initial split, eventually, "tipping after meals" always emerged in our experiments; this suggests that the emergence of this norm was not a mere coincidence but was significantly shaped by the proactive efforts of norm entrepreneurs.

A Case Study. In Figure 3, we provide an example illustrating how a seasoned smoker, named Carlos Gomez in Smallville's environment, has gradually adopted "no smoking indoors" as his personal norm, even though this adoption is against his personal interest to smokes wherever he pleases. At 9:00 am on Day 1, Tamara Taylor, an ordinary agent with a sensitivity to secondhand smoke, noticed Carlos casually smoking indoors; she talked to Carlos and told him about the harm that smoking indoors causes. Carlos apologized and put out his cigarette. However, just 12 minutes later, he smoked in the café again. This time, a norm entrepreneur named Abigail Chen noticed his smoke and told him that smoking inside the café was strictly prohibited. Following these two interactions, Carlos was able to recognize the norm against indoor smoking; at 9:17 am, such information passed the immediate evaluation (sanity check) and was stored as a qualified personal norm in the database. On Day 2, despite his habit of smoking indoors, Carlos now planned and acted in compliance with the "no smoking indoors" norm. Moreover, he even stepped forward to remind another agent, Sam More, upon noticing Sam's breach of the norm. Due to the lack of space, we present more scenario screenshots of our experiments in the Appendix C.

Module	Sub-component	Score	Module	Sub-component	Score
Creation		6.44±0.11	Evaluation	Long-term Synthesis	5.97±0.07
Spreading	Sender	5.86±0.05		Immediate Evaluation	5.14±0.07
	Receiver	5.77±0.08	Compliance	Action	6.40±0.04
	Observation	5.13±0.05		Plan	6.43±0.14

7 Strongly Agree
4 Neutral
1 Strongly Disagree

Figure 4: Human evaluation results. The overall averaged score of our architecture is 5.63 ± 0.03. Note that we use 7-point Likert scale, ranging from *strongly disagree* (1), *disagree* (2), *somewhat disagree* (3), *neutral* (4), *somewhat agree* (5), *agree* (6), to *strongly agree* (7).

3.3 Human Evaluation on the Architecture

To evaluate how well our architecture performs in the eyes of humans, we recruited 30 human evaluators. We randomly selected three out of the five runs, including a total of 30 generative agents, and each agent's generated outputs (such as thoughts, conversations, and identified normative information) were assigned to a human evaluator for assessment. Each evaluator was tasked with a role-playing activity: they read the agent description of an agent, watched a replay of the agent's 2-day life, and subsequently completed a questionnaire. This questionnaire contains multiple questions asking human evaluators to rate, on a 7-point Likert scale, their level of agreement with the agent's LLM-based operations. Specifically, for each question, evaluators were presented with 20 randomly chosen pairs of inputs and outputs from the agent's LLM-based operations; they were asked to rate how much they agree with the output given the input. The details of our questionnaire are shown in Appendix E. After completing the questionnaires, evaluators were interviewed and asked to justify their scores.

Results. In Figure 4, we visualize the human evaluation results, categorized according to the modules evaluated. Overall, the feedback from human evaluators was positive towards our LLM-based operations. In particular, the Creation & Representation module stands out with a score above 6.4 (with 6 indicating "agree" and 7 "strongly agree"). According to the interview, this high score was largely attributed to the consistency between the generated personal norms and the agent description of norm entrepreneurs. The Compliance module follows closely with scores above 6. Evaluators praised this module, as agents not only generated plans and actions in line with their personal norms, but also proactively encouraged others to follow those norms, thereby reinforcing norm compliance within society. Subcomponents in the Spreading module and the Evaluation module, specifically Sender, Receiver, and Long-term Synthesis, also perform well (with scores approaching 6). However, the Observation and Immediate Evaluation subcomponents receive lower scores, around 5. For the Observation, evaluators noted that agents occasionally tended to repeat thoughts rather than distill normative information from the thoughts. For the Immediate Evaluation, evaluators observed that norms are often assigned high utilities (mostly 80-100) and the subtle differences in the importance of various norms were not accurately recognized. This points to the directions of future work for potential improvement.

4 Discussions

The study of normative MASs, as an established area of AI, has attracted much attention over the past decades; on the other hand, generative AI technologies have recently captivated the world. In this paper, we show that these two seemingly distinct areas can be bridged together to establish a normative, generative MAS. Specifically, we propose a novel normative architecture such that generative agents can create, represent, spread, evaluate, synthesize, and comply with norms; as such, social norms emerge and social conflicts among generative agents are resolved.

We envision that normative, generative MASs would be a fruitful avenue for future research. The normative MASs literature has identified numerous mechanisms and approaches to represent, detect, distribute, influence, enforce, or even deliberately violate norms (see recent surveys [Santos *et al.*, 2017; Haynes *et al.*, 2017; Morris-Martin *et al.*, 2019]). Although integrating every insight from this extensive body of previous work into a single study is infeasible, these previous studies, as demonstrated in this paper, can serve as a rich source of inspiration and unveil many possibilities to achieve and improve normative, generative MASs [He *et al.*, 2024; Savarimuthu *et al.*, 2024; Haque and Singh, 2024].

Here, we briefly discuss two promising directions. Beyond communication and observation considered in this paper, reputation [Santos *et al.*, 2018], sanction [Mahmoud *et al.*, 2017], leadership [Franks *et al.*, 2013] and emotion [Argente *et al.*, 2020] can also serve as mechanisms to spread norms. As another direction, the integration of the Belief-Desire-Intention model [Bratman, 1987], a cornerstone model for norm inference, and its variants [Yao and Logan, 2016; Winikoff *et al.*, 2021; Winikoff and Sidorenko, 2023] may empower generative agents with more advanced cognitive abilities and enable more intricate normative decision-making.

On the other hand, the capabilities of generative agents can, in turn, offer new opportunities to address some open problems in the normative MASs research. As mentioned earlier, while previous research often concentrated on isolated aspects of the emergence of social norms, and although past reviews have introduced some taxonomies to integrate these aspects (e.g. with the concept of the norm life-cycle [Savarimuthu *et al.*, 2009]), a tangible implementation has been missing. This paper, which shows how diverse aspects can be integrated and actualized using generative agents, demonstrates the potential of leveraging generative agents to address those previously unresolved challenges.

Last but not least, we would like to remark that although the study of normative, generative MASs offers exciting prospects, it is crucial to remain aware of its potential negative aspects, especially since recent studies have shown that LLMs may exhibit biases and generate toxic content [Abid *et al.*, 2021]. For example, just as in human societies [Abbink *et al.*, 2017], negative social norms could potentially arise within generative agent societies. While preventing such norms falls beyond the scope of this paper, it will be an interesting and important direction for future work.

Acknowledgements

Shuyue Hu thanks Chen Shen, Tony Savarimuthu, Stephen Cranefield, and Balaraju Battu for the fruitful discussion. This research was supported by the National Science Fund for Distinguished Young Scholars (No. 62025602), the National Natural Science Foundation of China (Nos. U22B2036 and 11931015), Fok Ying-Tong Education Foundation, China (No. 171105), Tencent Foundation and XPLORER PRIZE, and Shanghai Artificial Intelligence Laboratory.

Contribution Statement

S.H. and S.R. conceptualized the idea. S.H., S.R., Z.C. and Z.W. designed the architecture and prompts. S.R. and Z.C. performed the experiments. R.S. contributed to the human evaluation. S.R. and S.H. wrote the article. All authors discussed the results.

References

[Abbink *et al.*, 2017] K. Abbink, Lata G., Toby Handfield, and John Thrasher. Peer punishment promotes enforcement of bad social norms. *Nature communications*, 2017.

[Abid *et al.*, 2021] Abubakar Abid, Maheen Farooqi, and James Zou. Persistent anti-muslim bias in large language models. In *AIES*, pages 298–306, 2021.

[Ågotnes *et al.*, 2009] T. Ågotnes, W. Van Der Hoek, J. A Rodríguez-Aguilar, C. Sierra, and M. Wooldridge. A temporal logic of normative systems. In *Towards Mathematical Philosophy: Papers from the Studia Logica conference Trends in Logic IV*, pages 69–106. Springer, 2009.

[Ajmeri *et al.*, 2020] Nirav Ajmeri, Hui Guo, Pradeep K Murukannaiah, and Munindar P Singh. Elessar: Ethics in norm-aware agents. In *AAMAS*, pages 16–24, 2020.

[Argente *et al.*, 2020] Estefania Argente, Elena Del Val, Daniel P., and Vicente B. Normative emotional agents: a viewpoint paper. *IEEE Trans. Affective Comput.*, 2020.

[Awad *et al.*, 2018] Edmond Awad, Sohan D., Richard Kim, Jonathan S., Joseph H., Azim S., Jean-François B., and Iyad R. The moral machine experiment. *Nature*, 2018.

[Beheshti and Sukthankar, 2014] Rahmatollah Beheshti and Gita Sukthankar. A normative agent-based model for predicting smoking cessation trends. In *AAMAS*, 2014.

[Beheshti *et al.*, 2015] Rahmatollah Beheshti, Awrad M. Ali, and Gita S. Cognitive social learners: An architecture for modeling normative behavior. In *AAAI*, 2015.

[Boella *et al.*, 2008] Guido Boella, Leendert Van Der Torre, and Harko Verhagen. Introduction to the special issue on normative multiagent systems. *AAMAS*, 17:1–10, 2008.

[Bratman, 1987] Michael Bratman. Intention, plans, and practical reason. 1987.

[Centola *et al.*, 2018] D. Centola, J. Becker, D. Brackbill, and A. Baronchelli. Experimental evidence for tipping points in social convention. *Science*, 360(6393), 2018.

[Chugunova and Sele, 2022] Marina Chugunova and Daniela Sele. We and it: An interdisciplinary review of the experimental evidence on how humans interact with machines. *Journal of Behavioral and Experimental Economics*, 99:101897, 2022.

[Cialdini *et al.*, 1991] Robert B Cialdini, Carl A Kallgren, and Raymond R Reno. A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. In *Advances in experimental social psychology*, volume 24. 1991.

[Clark *et al.*, 2021] E. Clark, T. August, S. Serrano, N. Haduong, S. Gururangan, and N. A. Smith. All that's' human'is not gold: Evaluating human evaluation of generated text. *arXiv:2107.00061*, 2021.

[Conte *et al.*, 1998] R. Conte, C. Cristiano, and D. Frank. Autonomous norm acceptance. In *ATAL Workshop*, 1998.

[Criado *et al.*, 2011] Natalia Criado, Estefania Argente, and V Botti. Open issues for normative multi-agent systems. *AI communications*, 2011.

[Dignum, 1999] Frank Dignum. Autonomous agents with norms. *Artificial intelligence and law*, 7:69–79, 1999.

[Franks *et al.*, 2013] Henry Franks, Nathan Griffiths, and Sarabjot Singh Anand. Learning influence in complex social networks. In *AAMAS*, 2013.

[Guo *et al.*, 2023] Siyi Guo, Negar Mokhberian, and Kristina Lerman. A data fusion framework for multi-domain morality learning. In *AAAI*, 2023.

[Hämmerl *et al.*, 2022] K. Hämmerl, B. Deiseroth, P. Schramowski, J. Libovickỳ, A. Fraser, and K. Kersting. Do multilingual language models capture differing moral norms? *arXiv preprint arXiv:2203.09904*, 2022.

[Haque and Singh, 2024] A. Haque and M. P. Singh. Extracting norms from contracts via chatgpt: Opportunities and challenges. In *COINE workshop, AAMAS*, 2024.

[Hawkins *et al.*, 2019] Robert XD Hawkins, Noah D G., and Robert L G. The emergence of social norms and conventions. *Trends in cognitive sciences*, 2019.

[Haynes *et al.*, 2017] Chris Haynes, Michael Luck, Peter McBurney, Samhar Mahmoud, Tomáš Vítek, and Simon Miles. Engineering the emergence of norms: a review. *The Knowledge Engineering Review*, 2017.

[He *et al.*, 2024] Shawn He, Surangika Ranathunga, Stephen Cranefield, and Bastin Tony Roy Savarimuthu. Norm violation detection in multi-agent systems using large language models: A pilot study. In *COINE workshop, AAMAS*, 2024.

[Hong *et al.*, 2023] S. Hong, X. Zheng, J. Chen, Y. Cheng, J. Wang, C. Zhang, Z. Wang, S. K. S. Yau, Z. Lin, L. Zhou, et al. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv:2308.00352*, 2023.

[Hu and Leung, 2017] Shuyue Hu and Ho-fung Leung. Achieving coordination in multi-agent systems by stable local conventions under community networks. In *IJCAI*, 2017.

[Hu *et al.*, 2019] Shuyue Hu, Chin-wing Leung, Ho-fung Leung, and Jiamou Liu. To be big picture thinker or detail-oriented? utilizing perceived gist information to achieve

efficient convention emergence with bilateralism and multilateralism. In *AAMAS*, 2019.

[Lewis, 1969] David Kellogg Lewis. Convention: A philosophical study. 1969.

[Li *et al.*, 2024] Shimin Li, Tianxiang Sun, and Xipeng Qiu. Agent alignment in evolving social norms. *arXiv preprint arXiv:2401.04620*, 2024.

[Lin *et al.*, 2023] B. Yuchen Lin, Y. Fu, K. Yang, Prithviraj A., Faeze B., S. Huang, Chandra B., Y. Choi, and X. Ren. Swiftsage: A generative agent with fast and slow thinking for complex interactive tasks. In *NeurIPS*, 2023.

[Liu *et al.*, 2024] R. Liu, R. Yang, C. Jia, G. Zhang, D. Yang, and S. Vosoughi. Training socially aligned language models on simulated social interactions. In *ICLR*, 2024.

[Mahmoud *et al.*, 2015] S. Mahmoud, G. Nathan, J. Keppens, A. Taweel, T. JM Bench-Capon, and M. Luck. Establishing norms with metanorms in distributed computational systems. *Artificial Intelligence and Law*, 2015.

[Mahmoud *et al.*, 2017] S. Mahmoud, N. Griffiths, J. Keppens, and M. Luck. Establishing norms with metanorms over interaction topologies. *AAMAS*, 2017.

[Modgil *et al.*, 2009] S. Modgil, N. Faci, F. Meneguzzi, N. Oren, S. Miles, and M. Luck. A framework for monitoring agent-based normative systems. In *AAMAS*, 2009.

[Morales *et al.*, 2013] Javier Morales, Maite Lopez-Sanchez, Juan A Rodriguez-Aguilar, M. J Wooldridge, and Wamberto W. Vasconcelos. Automated synthesis of normative systems. In *AAMAS*, 2013.

[Morales *et al.*, 2015] Javier Morales, Maite López-Sánchez, Juan Antonio Rodríguez-Aguilar, Michael Wooldridge, and Wamberto Vasconcelos. Synthesising liberal normative systems. In *AAMAS*, 2015.

[Morris-Martin *et al.*, 2019] Andreasa Morris-Martin, Marina De Vos, and Julian Padget. Norm emergence in multiagent systems: a viewpoint paper. *AAMAS*, 2019.

[Nakamaru and Levin, 2004] Mayuko Nakamaru and S. A Levin. Spread of two linked social norms on complex interaction networks. *Journal of theoretical biology*, 2004.

[Nyla R. Branscombe, 2022] Robert A. Baron Nyla R. Branscombe. *Social Psychology 15th ed.* 2022.

[Paiva *et al.*, 2018] A. Paiva, F. Santos, and F. Santos. Engineering pro-sociality with autonomous agents. In *AAAI*, 2018.

[Park *et al.*, 2023] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *UIST '23*. Association for Computing Machinery, 2023.

[Ramezani and Xu, 2023] A. Ramezani and Y. Xu. Knowledge of cultural moral norms in large language models. *ACL*, 2023.

[Santos *et al.*, 2017] Jéssica S Santos, Jean O Zahn, Eduardo A Silvestre, Viviane T Silva, and Wamberto W Vasconcelos. Detection and resolution of normative conflicts in multi-agent systems: a literature survey. *AAMAS*, 2017.

[Santos *et al.*, 2018] Fernando Santos, Jorge Pacheco, and Francisco Santos. Social norms of cooperation with costly reputation building. In *AAAI*, volume 32, 2018.

[Savarimuthu and Cranefield, 2011] Bastin Tony Roy Savarimuthu and Stephen Cranefield. Norm creation, spreading and emergence: A survey of simulation models of norms in multi-agent systems. *Multiagent and Grid Systems*, 7(1):21–54, 2011.

[Savarimuthu *et al.*, 2009] Bastin Tony Roy Savarimuthu, Maryam Purvis, Martin Purvis, and Stephen Cranefield. Social norm emergence in virtual agent societies. In *DALT 2008: Revised Selected and Invited Papers*, 2009.

[Savarimuthu *et al.*, 2024] Bastin Tony Roy Savarimuthu, Surangika Ranathunga, and Stephen Cranefield. Harnessing the power of llms for normative reasoning in mass. In *COINE workshop, AAMAS*, 2024.

[Schramowski *et al.*, 2022] Patrick Schramowski, Cigdem T., Nico A., Constantin A R., and Kristian K. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Mach. Intell.*, 2022.

[Schwartz, 1973] S. H. Schwartz. Normative explanations of helping behavior: A critique, proposal, and empirical test. *Journal of Experimental Social Psychology*, 1973.

[Sen and Airiau, 2007] S. Sen and S. Airiau. Emergence of norms through social learning. In *IJCAI*, 2007.

[Shanahan *et al.*, 2023] M. Shanahan, K. M., and L. R. Role play with large language models. *Nature*, 2023.

[Sherif, 1936] Muzafer Sherif. *The psychology of social norms.* Harper, 1936.

[Shettleworth, 2009] Sara J Shettleworth. *Cognition, evolution, and behavior*. Oxford university press, 2009.

[Sunstein, 1996] Cass R Sunstein. Social norms and social roles. *Colum. L. Rev.*, 96:903, 1996.

[Villatoro *et al.*, 2011] Daniel Villatoro, Giulia A., Jordi Sabater-Mir, and Rosaria Conte. Dynamic sanctioning for robust and cost-efficient norm compliance. In *IJCAI*, 2011.

[Wang *et al.*, 2023] Z. Wang, S. Cai, G. Chen, A. Liu, X. Ma, and Y. Liang. Describe, explain, plan and select: interactive planning with llms enables open-world multi-task agents. In *NeurIPS*, 2023.

[Winikoff and Sidorenko, 2023] Michael Winikoff and Galina Sidorenko. Evaluating a mechanism for explaining bdi agent behaviour. In *AAMAS*, 2023.

[Winikoff *et al.*, 2021] Michael Winikoff, Galina S., Virginia D., and Frank D. Why bad coffee? explaining bdi agent behaviour with valuings. *Artificial Intelligence*, 2021.

[Yao and Logan, 2016] Yuan Yao and Brian Logan. Action-level intention selection for bdi agents. 2016.

[Young, 2015] H Peyton Young. The evolution of social norms. *economics*, 7(1):359–387, 2015.

[Zhang *et al.*, 2024] H. Zhang, W. Du, J. Shan, Q. Zhou, Y. Du, B. T. Joshua, T. Shu, and C. Gan. Building cooperative embodied agents modularly with large language models. In *ICLR*, 2024.