

# Do Large Language Models Solve the Problems of Agent-Based Modeling? A Critical Review of Generative Social Simulations

Maik Larooij<sup>1\*</sup> and Petter Törnberg<sup>1</sup>

<sup>1</sup>University of Amsterdam.

\*m.k.larooij@uva.nl.

## Abstract

Recent advancements in AI have reinvigorated Agent-Based Models (ABMs), as the integration of Large Language Models (LLMs) has led to the emergence of “generative ABMs” as a novel approach to simulating social systems. While ABMs offer means to bridge micro-level interactions with macro-level patterns, they have long faced criticisms from social scientists, pointing to e.g., lack of realism, computational complexity, and challenges of calibrating and validating against empirical data. This paper reviews the generative ABM literature to assess how this new approach adequately addresses these long-standing criticisms. Our findings show that studies show limited awareness of historical debates. Validation remains poorly addressed, with many studies relying solely on subjective assessments of model ‘believability’, and even the most rigorous validation failing to adequately evidence operational validity. We argue that there are reasons to believe that LLMs will exacerbate rather than resolve the long-standing challenges of ABMs. The black-box nature of LLMs moreover limit their usefulness for disentangling complex emergent causal mechanisms. While generative ABMs are still in a stage of early experimentation, these findings question of whether and how the field can transition to the type of rigorous modeling needed to contribute to social scientific theory.

**Keywords:** Systematic literature review, Agent-Based Modeling, Generative Agents, Autonomous Agents, Large Language Models, Multi-Agents, Validation

## 1 Introduction

Human societies emerge from countless interactions of individuals – yet, social scientific methods tend to focus on either aggregate patterns or individual decision-making, while neglecting the central question of the relationship between the two [1]. Agent-Based Models (ABMs) promise a means of changing this, by enabling researchers to simulate how macro-level patterns emerge from micro-level interactions. By modeling individuals as autonomous agents with heterogeneous characteristics and adaptive behaviors, ABMs offer a way to capture the decentralized, dynamic, and non-linear nature of social systems. This approach is particularly appealing for studying phenomena such as collective action, the diffusion of innovations, political polarization, segregation dynamics, and financial market fluctuations – domains where the emergent dynamics for which traditional analytical methods struggle to account play a central role. ABMs moreover enable exploring how the world could be different, testing how different policy interventions or structural changes might shape societal outcomes.

Despite their theoretical advantages, however, ABMs have historically struggled with widespread adoption within the social sciences. Critics have in particular raised two key objections.

First, the models have been criticized for oversimplifying human behavior by representing individuals as simple rule-followers or optimizers [2, 3]. While this approach has been necessary to make computational implementation feasible, it often fails to capture the complexity of human decision-making, which is characterized by more complex reasoning, story-telling, as well as learning, emotions, social norms, and cognitive biases [4]. As each model is essentially distinct and a tailor-made representation of a specific social phenomenon [5], comparisons across models tend to be challenging, limiting the possibility for cumulativity of research findings [6–9].

Second, the models have been criticized for lacking clear connections to the empirical world, often reflecting the assumptions of the modeler more than the realities of human behavior [7]. Rigorously calibrating and validating the models against real-world data has proven challenging due to the models' complexity and lack of standardization [5, 7]. Unlike traditional models with clear parameter estimation techniques, ABMs often involve a large number of assumptions about agent behavior, making empirical validation difficult. Scholars have argued that simulation models must be subjected to rigorous validation if they are to contribute to our understanding of the simulated system [10]. The lack of standardized methods for calibrating these models with real-world data has moreover led to concerns about their reliability and reproducibility [11]. Because ABMs can generate highly detailed and complex outputs, distinguishing between meaningful patterns and overfitting to specific assumptions became a major challenge, as the models can be fit to match nearly any data – in von Neumann's famous phrasing, “with four parameters I can fit an elephant, with five I can make him wiggle his trunk”. Without rigorous empirical grounding, many ABM studies struggled to move beyond illustrative toy-models, limiting their influence in policy and empirical social science research.

The models have hence existed in a fundamental tension between the contradictory aims of realism and explainability, and between their inherent versatility and

the demands of validation, calibration, and comparability. As a result of these challenges, the data-driven approach of computational social science came to overshadow ABMs from the 2010s onward, following advancement in machine learning and growing availability of digital social data [12].

However, recently, ABMs have made an unexpected comeback. Large Language Models (LLMs) have gained significant traction in recent years, enabling computers to mimic human creativity, reasoning, and language-generation. The rise of LLMs quickly led to the idea that LLMs' capacity to mimic human behavior could be leveraged to simulate social systems, by integrating LLMs with ABMs [13, 14]. This promises to solve the challenge of ABMs' lacking realism, by drawing on LLMs' human-like reasoning, language generation, and behavior [15]. Such "Generative ABMs" offer several advantages, allowing agents to be equipped with distinct personalities and human-like capabilities. Generative ABMs have already been harnessed in both cooperative – achieving a shared goal – and adversarial – debate and competition – settings [2], with applications including debating [16], policy making [17, 18], economy [19, 20], epidemic modeling [21], (online) society simulation [14, 22, 23], psychology [24, 25], gaming [26, 27], software development [28] and embodied agents [29].

However, while LLMs promise to address the first key challenge of ABMs by making agents more realistic, their implications for the second – rigorous validation and calibration – remain an open question that is central to the future potential of generative ABMs. To address the question of whether generative ABMs enable us to move past the problems that have historically limited the adoption of ABMs, this paper carries out a systematic review of the rapidly growing field of generative ABMs to examine whether and how modelers have dealt with the central challenge of validation. The paper situates generative ABMs in the longer history of ABMs, and the long-standing challenges of validation and calibration. We carry out a systematic review of current research in the field and their validation practices. We find that the issue of validation remains insufficiently addressed in the field, and that it represents a key weakness of the generative ABM literature. Discussing the challenge, we argue that there are reasons to believe that integrating LLMs in ABMs may, in fact, aggravate rather than resolve these challenges, as LLMs are highly complex 'black-box' models that intensify the challenges of interpretability, standardization, and replicability [30, 31].

## 2 Agent-Based Modeling in the Social Sciences

Imagine a murmuration of starlings. Collectively, they form a fluid cloud, moving as a single organism. Yet, there is no "group mind" or leader orchestrating the flight. Each bird simply responds to its neighbors, who in turn respond to it – producing a fluid, nonlinear pattern that appears both choreographed and organic. Modeling this movement from a global perspective – as, say, a set of interacting variables – would be impossible. But, as early computer modelers discovered in the 1980s, if we model the individual birds, then the dynamics of the murmuration will simply *emerge* from the bottom-up through the aggregation of local interactions [32].

ABMs take the same approach to studying human social dynamics, treating social systems as consisting of 'agents' that perceive their environment, interact and take

actions – and casting society as a form of murmuration: a global pattern that emerges through individual interaction [33]. Traditional sociological methods view society as a hierarchical system where institutions and norms shape individual behavior from the top down, captured through a set of variables. ABMs, by contrast, explore how complex social patterns emerge from decentralized interactions among many agents. Instead of imposing top-down explanations, ABMs view social structures, norms, and institutions as emerging from the interactions of individuals. This shift views human groups as being similar to ant-hills or bird flocks: they are *complex systems*; non-linear, path-dependent, and self-organizing. and enables studying how these dynamics shape social phenomena – which is often impossible to achieve using traditional methods [4]. ABMs allow examining how macro-patterns may unexpectedly emerge from underlying micro-level mechanisms. At the same time, they only provide *sufficient* explanations, not *necessary* ones, as the same macro-phenomenon can be explained by several different micro-level mechanisms.

ABM as a social scientific method traces back to the early 1970s, with Thomas Schelling’s influential models of social segregation [34]. The method however saw rapid growth as the rising availability of computers made modeling accessible to a broad range of researchers. In the 1990s and early 2000s, ABMs grew substantially, together with the broader field of complexity science. The method developed its own journals (e.g., the Journal of Artificial Societies and Social Simulation, JASSS) and scientific association (The European Social Simulation Association, ESSA).

From 2010 onward, however, ABM and complexity science stalled, challenged by the growth of Computational Social Science and the broader emphasis on data analysis over generative explanation [12].

ABMs faced growing criticism for being empirically untethered to the systems being represented, reflecting the *ad hoc* intuitions of the modeler rather than the dynamics of the real world. The underlying rules governing the actions of the agents would often be simply assumed by the modeler based on what seemed to them to make sense, often without a clear link to either empirical data nor existing social scientific theory [7]. While such an approach may have been acceptable in a situation of data scarcity, the growing availability of digital data in the 2010s intensified demands that ABMs should be calibrated and validated against the real world [8, 10, 35, 36]. Critics revived long-standing arguments that models that have not been subjected to proper validation are “void of meaning” and contribute nothing to the understanding of the simulated system [10]. Without rigorous empirical grounding, ABMs remain mere toy-models or illustrations, limiting their usefulness for contributing to social science research, and their capacity to influence policy.

As journals increasingly began requiring ABMs to be calibrated and validated against real-world data, this often turned out to be a tall order. As ABMs seek to capture emergent outcomes from large numbers of interacting agents, they are themselves complex systems [37]. While this property is necessary for the models to capture social complexity, it also means that the models display many of the pernicious properties that make complex systems challenging to study: the models are often mathematically chaotic, and hence sensitive to initial conditions [38], making them challenging to reproduce. They are moreover often high-dimensional, with their many degrees of

freedom resulting in the so-called ‘curse of dimensionality’ [39]. This dimensionality furthermore makes means that ABMs are computationally costly: the agent interactions tend to scale quadratically with the number of agents, whereas sensitivity analyses scale exponentially with the number of parameters.

The flexibility and lack of standardization means that few design constraints are imposed by ABMs. As a result, every model is specifically tailored for its intended purpose, making models difficult to compare [6]. The lack of comparability represents an impediment to the type of universal methodologies and benchmarks that are crucial for achieving cumulativity of research findings, as it is otherwise unclear how the findings of one model speak to those of another model, let alone to dynamics in the real world. As a result, the field of ABMs faced a “replication crisis”, as studies found that the models could almost never be replicated – and that some of the findings in the field turned out to be the results of software bugs [40].

The lack of empirical grounding implies the need for the models to be highly convincing representations of the systems they simulate. However, social scientists have remained skeptical of the capacity of the often simple rule-based models to represent human behavior. The theoretical descriptions of ABMs as consisting of intelligent, adaptive, and goal-seeking agents that dynamically shape and respond to their environment while taking into account past events [41] stood in often stark contrast to their realities as simple rule-followers based on a series of “if-then” statements or simple optimization. Using such simple models, it is hard to convincingly argue that the models encompass the full complexity of human decision-making, in particular as the role of beliefs, memories, story-telling, and past interactions are rarely accounted for [42]. While ABMs do represent improvements over the atomized perfect optimizers of neoclassical theory, they remain part of the same broader methodologically individualist paradigm [1, 43].

Many social scientists have furthermore found it challenging to integrate ABMs with existing social scientific theory. In part due to their simplicity, the explanations offered by ABMs often seemed to be fundamentally at odds with existing social scientific explanations, as they cast any social phenomenon as simply ‘emerging’ from bottom-up interactions [36]. Such reductive explanations tend to come with a hint of the same political biases as neoclassical models, through the lens of which the market appears as ‘natural’, while the top-down role of institutions are viewed as imposing unnatural constraints [44]. The widely lauded Schelling [34] segregation model, for instance, seemed to suggest that segregation naturally “emerged” from an innate mutual dislike of the opposing group, while disregarding dominant explanations like structural racism, white flight, and red-lining – thereby also seeming to eradicate the possibility of collective solutions to address the problem.

As a result of these challenges, together with the growing capacities of machine learning and AI to produce insights from the increasingly abundant digital social data, the social data analytics and computational social science came to overshadow ABMs from the 2010s onward [12].

## 2.1 Generative ABMs

The emergence of Large Language Models (LLMs) has recently marked an unexpected shift in the fortunes of ABMs. While ABM research had already experimented with using deep learning-based approaches to modeling human behavior [2], LLMs offer an entirely new approach to represent human behavior. As LLMs are trained to mimic human language and reasoning, they appear in many ways to be out-of-the-box models of humans. In contrast to traditional ABMs [45], such “generative agents” are trained on vast amounts of information, giving them an internal world-model that allows them to generalize to new problems and mimic human reasoning. Generative agents can furthermore understand and generate natural language in ways indistinguishable from humans [46].

These capacities were first explored in a 2023 paper using LLMs to simulate human behavior in an The Sims-like artificial world, finding that the generative agents would begin to reproduce human-like behavior and social dynamics [22]. Following this, the field of generative ABMs exploded in growth, making it subject to several reviews seeking to survey the emerging field [2, 15, 47–49].

The agents used in the ABMs are often provided with distinct personas, consisting of, for example, demographic, personal and social information, that is either hand-crafted, generated by AI, or based on real-world datasets. The agents are often given memory modules in which meaningful information can be stored and retrieved. Planning modules allow agents to reason and act according to self-generated plans. Based on these modules, agents are able to take actions to change the state of the environment (e.g. changing the state of a toilet to ‘occupied’), alter their internal states, or communicate with other agents.

Generative agents thus seem to promise to resolve the problem of ABMs’ lacking realism, by making it trivial to create simulations with agents that can remember, reason, argue, engage in realistic conversations, and achieve “believable human behavior” [22]. The substantial excitement surrounding generative ABMs is hence understandable.

At the same time, it would seem that the question of calibration and validation that has long haunted ABMs remains potentially exacerbated by the introduction of LLMs. We can point to three interrelated ways in which LLMs may aggravate the existing challenges of ABM validation.

First, LLMs are ‘black-box’ models: their capacities are emergent, and it is virtually impossible to understand why a particular model gives a particular output. Language models rely on large-scale neural networks, and their output is the result of billions of parameters [50]. The more advanced the network gets, the less we know exactly what is going on in its decision-making process. LLMs are furthermore stochastic – the same input can lead to different outputs – making validating and reproducing results potentially challenging.

Second, the models cannot be trusted to accurately represent social groups or impersonate human behavior. While models are given descriptions of individuals to impersonate, studies have found that the models are poor at representing groups and their attributes, often engaging in exaggerated stereotypes rather than accurate representations. This is intimately associated with the hotly debated issue of LLM bias

[51]. Specifically, we can point to two types of bias that pose challenges to validation of generative ABMs: *social bias* and *selection bias* [52]. Social bias refers to the discrimination, stereotypes, or prejudices against certain groups of people, which may mean that models do not accurately represent external groups, but rather engage in problematic stereotypes. Selection bias refers to the choice of texts that make up a training corpus, which means that LLMs may reproduce historical events or complete social network interactions, basing simulations on prior knowledge instead of dynamically producing the dynamics as intended by the modelers.

Third, as LLMs are probabilistic next-word predictors, they possess no internal mechanism to validate the correctness of their outputs. They may hence generate outputs that are factually incorrect or nonsensical – a phenomenon referred to as ‘hallucination’ [53, 54]. This could potentially lead to unexpected effects, or hard-to-solve problems in the internal workings of models, as the outputs may not be coherent with agents’ memories or personas (e.g. an agent ‘recalling’ events that it never experienced), leading to unexpected outcomes.

## 2.2 Validation in ABMs

Before turning to reviewing how validation has been handled in the field of Generative ABMs, we will first draw on the literature on validation in ABMs to provide a classification of types of validation that have been discussed and applied. Validation is the process of assessing the degree to which a model is an accurate representation of the real world [8]. Sargent [42] argues that the aim of validation should be ‘operational validity’: “determining whether the simulation model’s output behavior has the accuracy required for the model’s intended purpose over the domain of the model’s intended applicability”. Models are always abstractions, and it is necessary to bracket part of the system being modeled. ‘Realism’ is, in other words, neither achievable nor desirable.

First, the literature distinguishes between *internal* and *external* validation methods. External validation makes use of external resources for validation, such as real-world data, well-studied behavior, or humans for judgment. Internal validation is carried out to validate the internal consistency or to do internal observations of the model, for example by monitoring the system and reasoning about the observations or performing a sensitivity analysis [55].

Second, we can distinguish between ‘subjective’ and ‘objective’ validation. The former includes techniques like human judgment or basic observations of the system in the form of traces, graphics and performance indicators. One influential example of subjective validation is the Turing Test, in which individuals are asked if they can discriminate between human and model outputs [56]. Objective validation encompasses techniques such as comparing with real-world data or similar models with the use of statistical tests. One of the earliest papers on model validation proposes a multi-stage validation approach of first developing the model’s assumptions, then validating where possible by empirically testing them, and lastly comparing the input-output relations of the model to the real system [10].

These two dimensions allow us to divide validation techniques into four categories. For example, using human judges would be *subjective* and *external*, whereas a sensitivity analysis would be *objective* and *internal*.

### 3 Method

To identify papers, we used as a starting-point a series of previous survey studies [2, 15, 47–49], and employed backward snowballing as described in [57]. We furthermore carried out a search to identify relevant new papers using Scopus, a widely used bibliographic database. The search aimed to capture key contributions across disciplines, including computer science, artificial intelligence, and computational social science. We developed a Boolean search query using keywords and phrases that reflect the core topics of interest. The query was designed to be broad enough to capture diverse research while ensuring specificity by targeting the title, abstract, and keywords fields:

```
TITLE-ABS-KEY ( ( "generative social simulation" ) OR ( "generative agent-based model*" ) OR ( "agent-based simulation" AND "generative AI" ) OR ( "LLM*" AND "agent-based model*" ) OR ( "large language model*" AND "ABM" ) OR ( "foundation model*" AND "ABM" ) OR ( "multi-agent system*" AND "generative AI" ) OR ( "generative agent*" ) OR ( "social simulation" AND "LLM*" ) OR ( "large language model-based agents" ) )
```

The final search was conducted on 2025-03-27, yielding 35 results. To ensure relevance, we followed a two-step screening process:

1. Title and abstract screening: We removed irrelevant papers based on a preliminary review of titles and abstracts.
2. Full-text review: The remaining papers were assessed for methodological rigor and direct relevance to the survey's research questions.

We focus on papers focused on using generative ABMs for social simulation. Studies were included if they met the following eligibility criteria:

- Papers that use Large Language Models (LLMs) as the basis of a simulated agent
- Papers that leverage multiple agents
- Papers that in which agents interact, meaning that their actions depend on the state of the system or on the actions of other agents.
- Papers in which the model is seen as representing human behavior

Studies were excluded if they:

- Focus on task completion rather than on modeling a social phenomenon.
- Do not seek to model or achieve similarity with human behavior.
- Were not published in English.
- Surveyed existing studies rather than presenting new research.

These criteria meant that we excluded papers in which agents made individual decisions, but with these never being affected by other agents, such as Feng et al's [58] study of human characteristics on electric vehicle charging behavior, or Ren et al's [59] study using agents with different personalities to replicate human web search behavior. We also excluded studies using agents to produce synthetic survey data [60], and papers aimed at creating frameworks for generative ABMs rather than replicating human behavior, like CAMEL [61], MetaGPT [62], CGMI [63] and AutoGen [64].

To extract the needed data, the papers were read in-depth, in particular sections containing the relevant information, such as the introduction, evaluation, and limitations sections. The identified records were inserted into a structured database containing all the classifications.

We are seeking to provide an overview of what systems researchers are seeking to reproduce using generative ABMs, and how they approach validation. We will then draw on this to discuss the current role of the issue of validation within generative ABM research. Our review hence focuses on the following questions:

1. RQ1: What social phenomena are generative ABMs seeking to reproduce?
2. RQ2: What strategies are employed to validate the models against the studied phenomena?

### 3.1 The target system

To answer RQ1, we distinguish between simulating *individual* and *group* behavior. Papers that seek to reproduce individual behavior focus on the actions of a single agent, for example if decisions are congruent with a given profile. Other papers search for group behavior at the macro-level of the simulation, like the formation of the network through connections, or the propagation of information through the network. For individual behavior, we further distinguish between the following aspects of the simulated phenomenon, which were identified inductively through an initial analysis:

1. Alignment with profile: the alignment of the agent's actions with their personality, memories and profile, as well as the specific situation.
2. Emotion: the emotions expressed by the agent given a specific situation or action.
3. Conversations and content generation: the communication and dialogue between agents.
4. Social awareness: the awareness of others and of social norms.
5. Decision making and reasoning: the decision making and reasoning capabilities.
6. Opinion and attitude: the opinion or attitude towards some event, and possibly the shift in opinion or attitude.

For group behavior, we used the following three categories:

1. Network propagation: the propagation of something through the network, such as (mis)information, emotion, attitudes, or cultural elements.
2. Network structure: the structure of the network, for example the network degree distribution, density or some other network structure attribute.
3. Social dynamics: phenomena that emerge from individuals actions, for example a 'herd effect'

Some papers focus on multiple phenomena and are thus given multiple categorizations. A paper is classified as focusing on a certain type of behavior only if the paper sets out to model that behavior, thus excluding observed phenomena that are noted only as an aside. For example, Kaiya et al [65] set out to measure social reasoning, but note some human-like emotions and reactions.

### 3.2 Validation Techniques

For RQ2, we draw on existing categorizations in the literature on ABM validation. We hence separate between *internal* and *external* validation, as well as between *subjective* and *objective* validation. We expand on the resulting four categories based on the initial analysis, and arrive at the following five categories of validation:

1. Validation based on human(-like) judgment [External / Subjective]: using input from researchers, experts, crowd-workers or even LLMs to judge the alignment of the model with human behavior
2. Validation against well-known social patterns [External / Subjective]: comparing observations of the model to well-established phenomena in human behavior, although not quantitatively
3. Validation against similar models [External / Subjective or Objective]: comparing subjectively or objectively with the results and behavior of earlier models, such as traditional ABMs.
4. Validation against human-generated data [External / Objective]: comparing to collected human behavioral data, such as social media interactions, human annotations or earlier experiments
5. Validation based on internal consistency [Internal / Objective]: evaluating the internal coherence and consistency of the system, for example through sensitivity analysis.

We distinguish between primary and secondary validation techniques. Primary techniques are the most important in the validation process of the paper, while secondary techniques are used merely to lend supportive evidence.

## 4 Results

We identify 35 papers that fulfill the eligibility criteria. The full classification of all 35 papers is presented in table 1. The generated content and conversations, social dynamics and network propagation are the most popular areas of study. 21 papers focus on a single category, the others measure multiple categories in one paper. On average, a paper studies 1.63 distinct categories.

### 4.1 Target systems of simulations

We will begin by reviewing the literature by focusing on what social phenomena that generative ABMs are used to reproduce. We will here discuss each category in turn, thereby providing an overview of the studied phenomena and their validation.

	Profile Alignment	Emotion	Conversation / Content	Social Awareness	Decisions / Reasoning	Opinion / Attitude	Network Propagation	Network Structure	Social Dynamics
Generative Agents [14]	●						●	●	
WarAgent [18]	●							●	
Social Simulacra [22]			●						
S3 [23]	●	●		●	●	●			
De Marzo et al. [66]								●	
Humanoid Agents [67]	●	●				●			
Lyfe Agents [65]				●					
Zhang et al. [68]								●	
AgentVerse [69]								●	
Williams et al. [21]				●					
Project Sid [70]	●		●	●	●	●			
Li et al. [71]				●					
Li et al. [72]						●			
SpeechAgents [73]		●							
FPS [74]						●			
SOTOPIA [75]				●					
Zhang et al. [76]			●						
MetaAgents [77]	●								
Xie et al. [78]				●					
CRSEC [79]				●			●		
He et al. [80]								●	
Wu et al. [81]				●					
Digital Representatives [82]	●								
Affordable Generative Agents [83]	●		●				●		
Chuang et al. [84]								●	
FUSE [85]							●		
HiSim [86]		●						●	
OASIS [87]							●	●	
Sreedhar & Chilton [88]				●					
LMAgent [89]	●		●	●	●				
Gu et al. [90]			●					●	
Orlando et al. [91]								●	
Wang et al. [92]	●							●	
Ferraro et al. [93]			●					●	
Zheng et al. [94]								●	

Table 1 Classification of the reproduced social phenomena for the 35 papers

## Profile Alignment

An important aspect of human behavior which modelers seek to reproduce is the way human actions are shaped by their memories, personality traits, and experiences. We may refer to this as *profile alignment*. Systematically validating profile alignment however represents a challenge, as it is not clear what constitutes the ‘correct’ agent action based on their profiles and specific situations. Ground truth is often challenging or nearly impossible to access, as it would require comparing the model agent’s behavior against the specific person being modelled. Park et al [14] assess the validity by interviewing agents on their self-knowledge, memory, plans, reactions, and reflections, and judging the answers ‘believability’ based on the alignment with the personality and memories of the agents as well as with the environment. As Yu et al. [83] build upon these Generative Agents, they also measure the believability of their Affordable Generative Agents. Wang et al. [92] also search for believability of their agents’ behavior and functioning of the memory system. Hua et al [18], which use LLM agents to simulate the interactions between conflicting countries, evaluate the congruence, stability, and rationality of actions with respect to the country profile in a war situation using expert assessment. Altera.ai [70] deployed LLM agents in a Minecraft session, and sought to validate profile alignment by monitoring the specialization into distinct roles that should align with the profile of the agent, which in turn should align with the agent’s actions. Jarrett et al [82] create ‘digital representatives’ of individuals, which are validated based on their alignment with the preferences of the individuals they substitute in collective decision-making scenarios. Liu et al. [89] validate the alignment of the agents’ shopping behavior with the profile, context and established rules and expectations.

## Emotional alignment

Two papers sought to use LLMs to simulate human emotions. Gao et al [23] sought to predict the emotions of social media users towards a particular event based on three levels: calm, moderate, and intense. Wang et al [67] develop a platform for modeling human behavior, including basic needs, social relationships, and emotions. The paper validates the emotions expressed by agents for different activities through comparison with human annotation.

## Conversations and content generation

In many papers, the agents can interact with each other through conversation. Simulations of, for instance, social media platforms often seek to validate that the content produced matches the content produced by human participants. Park et al [22] create ‘SimReddit’ – a synthetic version of the social media platform Reddit – and evaluate whether human participants can distinguish the synthetic conversations from human conversations. [83] leveraged the same strategy to validate the conversations of LLM agents interacting in a virtual town. Gu et al. [90] took a turn to quantitative text measures when comparing synthetic and real-world conversations, measuring the cosine similarity between the generated text and human text embeddings, and Gao et al. [23] extends this with the Perplexity score of the generated text as well. Ferraro et al.

[93] try to find out whether generative agents are able to produce similar keywords, interests and content compared to humans on social media. Mou et al [86] focuses instead on the agents' stances and the alignment of the generated content with real-world Twitter data. They furthermore categorize different types of content – e.g., 'call for action', 'sharing of opinion', or 'reference to a third party' – and compare whether these match the Twitter data.

Others are more concerned with the humanness of the conversation itself, as opposed to the similarity of the content. Wang et al [67] assessed conversational realism based on whether engaging in a conversation brought agents closer to one another. Zhang et al [73] design a system to simulate human communication, validating the performance based on consistency with the scenario and characters, and the quality and logical coherence of the script content. Zhang *et al.* [76] tried to validate whether their model is able to generate human-like dialogues by scoring it on the naturalness, empathy, interestingness and humanness of the conversations, as well as the agent's ability to choose a fitting dialogue strategy. Similarly, Liu et al. [89] compared the content of their agents with human content on naturalness and expressiveness.

### Social awareness and social intelligence

Social awareness refers to the collective consciousness shared by individuals in a society [95]. Validation may focus on the agents' ability to adhere to social norms and rules. Simulations may also focus on the related phenomenon of social intelligence, in which agents seek to reproduce the human ability to interpret actions of others.

Altera.ai [70] validate the capacities of the agents to understand both themselves and others. The authors assess whether agents were able to accurately deduce the sentiment of others and react to changing social cues. Li et al. [71] test whether agents can display a 'Theory of Mind', that is, whether they can reason about the concealed mental states of other agents. Zhou et al [75] simulate interactions in various social scenarios to measure the social intelligence of agents based on a newly created benchmark, SOTPIA-EVAL, which scores social intelligence based on believability, knowledge, keeping or revealing secrets, relationships, social rules, and financial and material benefits. Xie *et al.* [78] measured the extent to which agents exhibit human trust behavior, based on anticipating the actions and thoughts of other players in games like the Trust Game and the Dictator Game. Ren et al [79] focus on the emergence of social norms and how they are incorporated in the agents' planning and actions in a sandbox society.

### Decision-making and reasoning

Reproducing human decision making or reasoning is a common focus of generative ABMs. Gao et al [23], for instance, compare agents' decisions to take specific actions on social media – such as sharing, posting new content, or doing nothing – against real user decisions. Kaiya et al [65] evaluated agents based on their autonomy and social reasoning capabilities in three experimental scenarios: a murder mystery, a high school activity fair, and a patient-in-help scenario. Williams et al. [21] examined whether agents were able to make the decision to stay at home or go into quarantine in a COVID-19 like epidemic scenario. Altera.ai [70] assessed agents' ability to reason

about societal rules. Wu et al. [81] explore whether agents were able to make adaptive decisions without explicit directions, focusing on three scenarios where agents can only reach an optimal outcome when working together in a competitive context. Sreedhar and Chilton [88] simulate the Ultimatum Game with LLM agents to simulate and replicate human strategic reasoning. Liu et al. [89] simulate user shopping behavior and evaluate on the shopping decisions made by the agents.

### Opinions and attitudes

Two papers focused on using generative agents to simulate human attitude or opinions. Validations of this category can be focused on replicating the change of attitude or opinion in reaction to a certain event. Gao et al [23] validate their agents' attitudes (positive/negative) towards posts on social media against real-world social media data. Wang et al [67] tested whether they are able to predict if certain activities satisfy basic needs of fullness, social, fun, health, and energy, thus measuring the attitude towards events and the effect on the individual agents.

### Network propagation

Network propagation was a key focus of study, with 8 out of 29 papers carrying out some form of validation of network propagation. The validation of these systems are often concerned with measuring the spread of something, like information or attitudes, and comparing this with real-world human behavioral patterns.

Park et al [14] focused on the diffusion of information through the network. Gao et al [23] measure three forms of propagation on social media: information, emotion, and attitudes. Li et al. [72] and Liu et al [74, 85] simulate the spread of misinformation in a social media network structure. [87] reported on the spread of all forms of information. Altera.ai [70] track the spread of cultural memes and religion in a network, and Ren et al [79] seek to model the spread of social norms through a social network.

### Network structure

The structure of the network is mainly determined by the formation of relationships between agents. Here, the goal of the validation is to make sure that the resulting network characteristics match real-world network properties. De Marzo et al [66], for instance, evaluate whether the resulting network structure from their simulation was scale-free, which is a near-universal property of networks emerging on social media. He et al [80] examine whether the networks produced on their synthetic social media network, Chirper.ai, show the property of homophily, i.e., the tendency of agents to be more connected with others that are similar to themselves. Park et al [14] and Yu et al [83] both examine the evolution of social networks in their sandbox town simulation, using metrics such as the degree of nodes, and the density of the network. Hua et al. [18] assess the historical correctness of the formed alliances and war declarations between countries in a war situation. Gu et al. [90] try to measure the formation of echo chambers with measures like the modularity, network density, average path length and the clustering coefficient of the network.

## Social dynamics

This final category focuses on papers that seek to reproduce social dynamics, like a flock of starlings collectively forming a fluid cloud. This category encompasses group behavior that emerges from individual actions.

Zhang et al [68] claim to find conformity, consensus reaching, and group dynamics behavior in their agent society. The agents in Chen et al [69] exhibit volunteering, conformity, and destructive behavior when trying to survive in Minecraft. Wang et al. [92] also find conformity behavior, next to signs of an 'information cocoon'. A similar phenomenon, the formation of echo chambers, was also validated in various papers [90, 93, 94]. Chuang et al. [84] simulate the phenomenon of the Wisdom of Partisan Crowds, where the group average moves closer to the ground truth when exposed to the average belief of the partisan group, despite differences in initial belief. Mou et al [86] measure the attitude distribution and average attitude of users on a simulated Twitter platform, HiSim, using the deviation from the mean (bias), the standard deviation from the mean (diversity) and the similarity measured with Dynamic Time Warping and Pearson correlation. Yang et al [87] replicate the 'herd effect' on Reddit, by testing whether an initial like or dislike of a comment results in other agents following the behavior. They managed to reproduce the effect, but found that the effect was stronger among their generative agents than among human, likely as humans possess a stronger critical mind that reduces the herd effect. Orlando et al. [91] measure the replication of the 'friendship paradox', the phenomenon that individuals on average have fewer friends than their friends.

## 4.2 Validation Techniques

We turn now to the question of the validation approach taken in the papers. Table 2 provides an overview of the validation techniques used. For the classification of the papers, we distinguish between primary and secondary techniques. The primary techniques are those that are most important in the validation process, marked in green in the table. The secondary techniques are supporting evidence of lesser importance, marked in orange in the table.

We will here discuss the techniques used in order, drawing on examples of their use.

### Validation based on human or human-like judgment

One of the most widely used techniques for validation of generative agents is to draw on external or internal judges. The judges can be 1) experts (often in the form of the authors themselves), 2) crowd-workers, or 3) LLMs. For the first category, Hua et al [18] use field experts to evaluate the congruence and consistency of actions taken by countries with respect to their profile. Li et al [77] use the authors' own judgment to assess the performance of their MetaAgents to identify suited job-seekers, create a workflow and align with the given role. Wang et al. [92] make use of humans to asses the believability of the chatting behavior and memory module of the agents.

While the use of crowd-workers has been widely criticized both on ethical grounds, for low performance, and for their now wide-spread use of LLMs, crowd-workers

	Human(-like) Judgment	Social Patterns	Other Models	Human Generated	Internal Consistency	Other Models
	Subjective			Subjective		Objective
Generative Agents [14]	●	○				
WarAgent [18]	●			●	○	
Social Simulacra [22]	●		○			
S3 [23]				●		
De Marzo et al. [66]		●			○	
Humanoid Agents [67]				●		
Lyfe Agents [65]	●					
Zhang et al. [68]	●					
AgentVerse [69]	●					
Williams et al. [21]	●					
Project Sid [70]	●					
Li et al. [71]	●	○				
Li et al. [72]	●					
SpeechAgents [73]	●				○	
FPS [74]	●					
SOTOPIA [75]	●			○		
Zhang et al. [76]	●			●		
MetaAgents [77]	●					
Xie et al. [78]	●					
CRSEC [79]	●	○				
He et al. [80]		○			●	
Wu et al. [81]		●		○		
Digital Representatives [82]	○			●		
Affordable Generative Agents [83]	●		●			○
Chuang et al. [84]				●		
FUSE [85]	●					
HiSim [86]		○		●		●
OASIS [87]		○		●		
Sreedhar & Chilton [88]				●	○	
LMAgent [89]	●		○	●		
Gu et al. [90]			○	●	●	
Orlando et al. [91]	●					
Wang et al. [92]	○	●	○		○	
Ferraro et al. [91]	●			●		
Zheng et al. [94]	●					
Total (main technique)	12	14	1	12	1	1
Total (secondary technique)	2	6	3	4	3	2

**Table 2** Classification of the validation techniques for the 35 papers.

remain widely used for validation, often hired through Amazon MTurk. Park et al [14] recruited crowd-workers to score the agents on ‘believability’. Crowd-workers were also used in Park et al [22], to distinguish between synthetic conversations and real conversations for repopulated subreddits. Human annotators were used to evaluate whether generative agents answer Theory of Mind questions correctly in Li et al [71]. Several other papers also employ crowd-workers to evalaute whether humans can distinguish human and generative agents on various tasks [75, 76, 79].

While using LLMs to validate the realism of LLMs appears self-evidently problematic, the practice has become increasingly widespread as a means of validation. Zhou et al [75] use GPT-4 to validate their model realism. Zhang et al [73] use ChatGPT to assess the consistency and coherence of the conversational content. Jarrett et al [82] use an LLM to compare generated critique to a ground-truth. Liu et al. [89] use both humans and GPT-4 to evaluate the behavior of agents on among other believability, social influence and the naturalness and expressiveness of the content. Yu et al [83] use GPT-4 to discern whether responses originated from AI or a human, despite previous research suggesting that LLMs may be inappropriate for such tasks [96].

### Validation against well-known social patterns

Validation can also be carried out by matching the dynamics of the model with well-established real-world patterns. This means that no specific empirical data need to be collected for comparison, as the comparison is made against known social patterns.

De Marzo et al. [66] examine whether the networks generated show the property of having scale-free degree distributions, which is a well-known and easily measurable property of social media networks. Li et al [72] show that the effect of different personalities and agent attributes on agents’ propensity to share fake news matches earlier studies. Liu et al [74] conclude that their finding that fake political news propagates faster than other fake news topics is consistent with previous research. In addition, they compare the relationship between the Big Five personalities on agents’ propensity to spread misinformation with prior empirical research on the spread of misinformation. Orlando et al. [91] use the degrees of agents to validate the known phenomenon of the ‘friendship paradox’.

While some social patterns, such as degree distributions, are easily quantified, many known social patterns are qualitative rather than quantitative, meaning that authors often need to argue for the plausibility of the model in a more narrative form – with varying levels of persuasiveness and rigor. For instance, Zhang et al. [68] observe dynamics in their model that they argue corresponds to well-studied behavior in social psychology, such as conformity, consensus-reaching, and group dynamics. Chen et al [69] similarly observe social dynamics that they argue match established theories within social psychology. Williams et al [21] argue that the actions taken by agents in response to disease outbreak – such as staying at home and quarantining – share some similarities to the behavior observed during the COVID-19 pandemic. Yang et al [87] observe the emergence of group polarization in their model, which they argue matches behavior observed in humans.

Kaiya et al [65] and Altera.ai [70] take an even more loose approach, merely observing the simulated system and claiming that it matches human behavior. Ren et al [79]

take a case study approach, observing their system and interpreting the dynamics as representing the adoption of norms and formation of social conflicts, claiming that it thereby matches real-world social dynamics. Wu et al [81] find that agents agree to cooperate without explicit directions and report on performance metrics, which were compared against a study on humans in only one of the three experiments. Yu et al [83] and Park et al [14] argue that the formation of relationships and the spread of information through the network matches expected social dynamics. Mou et al [86], Zheng et al [94] and Ferraro et al [93] remark that their systems reproduce the known social pattern of echo chambers. Wang et al. [92] similarly find signs of an 'information cocoon' and conformity behavior.

### **Validation against similar models**

Generative ABMs can also be validated through comparison with previous models. The catch here is that it is implied that the previous models are already rigorously validated, which circles back to the validation problem that this paper addresses. As a consequence, authors can conclude on the performance compared to the other model, but the alignment with human behavior depends heavily on the quality of the other model. These comparisons again vary in whether it is carried out in a rigorous quantitative way or in a more subjective way that is argued for in narrative form.

[73] compare the generated conversations of their SpeechAgents with a previous single-agent model. Mou et al [86] and Gu et al [90] systematically compare their social media simulations with conventional non-generative ABMs. Liu et al [89] compare the purchasing behavior of their agents to different filtering algorithms and two other agent based approaches. This system is in turn used by Wang et al [92] in a comparison with their approach to recommender agents.

Yu et al [83] offers an example of a more subjective comparison, comparing their model with Park et al's [14] Generative Agents. The main mode of validation in both papers was the 'believability' of the agents' responses to a set of interview questions. Yu et al [83], however, offers no quantitative comparison between the two models, but merely includes the responses of agents, concluding that their version did not obviously impair the believability of the agents compared to the original.

### **Validation against human-generated data**

One common technique for validation is to quantitatively compare the synthetic data produced by the model with 'real-world' data, for instance collected from digital media or from human annotation. This technique is particularly common within social media simulations. Gao et al [23], Mou et al [86], Yang et al [87], Gu et al [90] and Ferraro et al [93] all based their validation on social media datasets on comparison with data collected from platforms such as Twitter or Reddit. Used metrics include the speed and scale of the spread of information through the network, the stance and attitude of users towards certain events and the similarity of the generated content. In a non-social media system, Hua et al [18] use historical data to qualitatively assess the accuracy of their war simulation, such as war declarations or alliance formation. Liu et al [89] use a dataset of Amazon products and reviews, use a part of the purchase history for

the initialization of agents and keep the rest as a ground truth for comparison to see to what degree agents make the same purchases.

Another common approach is to use simulations to reproduce experiments that have previously been carried out with human participants. Papers have been found to compare the results of generative agents on trust games [78], the Ultimatum Game [88] and tests to measure the effect of the Wisdom of Partisan Crowds [84] to previous studies with human subjects.

Finally, Wang et al [67] compare classifications of human annotators with classifications made by the generative ABM on emotions. Zhang et al [76] use human annotations as a ground truth to validate generative agent choices of dialogue strategies. Jarrett et al [82] validate the behavior of agents as digital representatives against an annotated dataset of human opinions and critiques in combination with demographic information.

### Validation based on internal consistency

Some researchers seek to validate their models based on their internal consistency, coherence, and stability, for example, through sensitivity analysis. The goal may be to show that the agents' behavior is stable to changes in system parameters and perturbations. For example, Hua et al [18] experiment with injecting counterfactual information and de-anonymization of countries to compare with the base approach to see if the results were as expected. De Marzo et al [66] use different prompts for their language model to see if the network structure instead converged to an expected structure, such as a random graph. He et al [80] use statistical tests on network metrics to show that the network data can be divided into distinct communities. Sreedhar and Chilton [88] use simulations of agents with different personality traits to evaluate whether the differences between the models align with expectations. While these form of internal tests are relatively common, the extent to which they can claim to represent a plausible validation of the model, rather than just a verification of its basic functioning, is however debatable.

## 5 Discussion

Based on our review of the way generative ABMs have been employed and validated in the literature, we now turn now to our central question of how and whether the field is sufficiently addressing the long-standing challenges of validation in ABM, as well as the new challenges that emerge from the use of LLMs.

As we have seen, the most common validation technique is to simply employ on-the-face ‘believability’: 15 out of 35 papers are validated solely through subjective validation, and 22 out of 35 use subjective validation as their only primary technique. 14 papers use the judgment of human ‘experts’, crowd-workers, or LLMs as evaluators. At its best, this consists of the researchers themselves looking at the simulated system and arguing – with varying rigour – that it shares some similarity with human behavior or a social system. This form of validation can also involve using MTurk workers, despite the fact that the low quality of such validation data at this point has been well established [97]. At its worst, this form of validation consists of simply asking an

LLM to evaluate its own believability. Various papers have doubted the qualities of LLMs for evaluation purposes [98–100]. On the other hand, some have argued that LLMs as evaluators are as good as humans [16, 75], this, however, should be taken as an argument against using human judgment in validation, rather than as an argument for using LLMs. Humans – in particular those without training – are famously poor at judging whether a material is produced by AI, and it is arguably relatively simple to set up an experiment in such a way as to generate a positive outcome with even low-quality results – for instance by asking whether a given text is generated by AI or by humans, rather than by allowing side-by-side comparison between the synthetic and real-world data.

When studies do carry out more rigorous quantitative comparison between human-generated and synthetic data, it often quickly becomes clear that the two are substantially less similar than the on-the-face believability may suggest. The style of writing, which tends to be the focus of such comparisons, between zero-shot LLMs and human conversations tends to be quite different: LLM responses are longer, more polite, articulate, and respectful [76, 86]. While humans struggle to tell the difference, there are generally substantial syntactical differences between human-generated and synthetic text.

Yet, from the perspective of the literature on ABM validation, even the few studies that carry out more rigorous quantitative validation against real-world data show some fundamental issues. As mentioned, the notion of operational validity highlights that the purpose of validation is for the model to match the underlying mechanisms of the object over the domain of the model’s intended applicability [42]. In the case of generative ABMs, it is rarely clear that the model aspects being validated are at all relevant for the model’s mechanisms. Take, for instance, the models seeking to simulate social media platforms. The purpose of such simulation tends to be to test the effect of affordances or algorithms on social dynamics on social media. Yet, the validation tends to focus on whether the synthetic text produced shares some superficial syntactical features with real-world social media text. Arguably, unless the algorithms being tested happen to leverage syntactical features of the text, such aspects are irrelevant for operational validity, which would rather depend on how and when users react to one another, and how these actions in turn interact with the affordances and algorithms of the platform. The sufficiency of the aspects that are being validated for the dynamics of the model must be clearly evidenced.

Operational validity would moreover often require validating profile alignment for the individuals being simulated. In other words, it would be necessary not only to test that the agent is acting consistently with itself through the simulation, but that it is acting in the same way as the person being simulated – generally, in the included studies, a person described through a prompt persona. How to carry out such validation is however an open question, as rigorous persona alignment would require calibrating and validating against a specific individual and their actions within situation being captured by the model. Doing so would in most cases imply either collecting specific data on individuals through linked surveys or human-participant experiments, or seeking to create what we may call ‘digital human twins’ by calibrating on digital trace data – in turn implying challenging ethical questions.

In existing studies, however, nearly all studies use zero-shot models without fine-tuning. ‘Calibration’ hence consists of using prompt engineering to achieve on-the-face believability. As a result, it is simply assumed that LLMs will accurately model individuals – despite that it has been shown that the models are poor at representing social groups, not least due to often problematic social bias [101]. While some papers do mention the problems of social bias in LLMs, for instance finding that the agents display gender biases in the simulation [78], the more fundamental question of how this impacts calibration of the model against human behavior is not raised.

The central feature of ABMs – their inherent flexibility that simultaneously affords their great versatility but also makes them challenging to calibrate, validate, reproduce, and compare, and that make it nearly impossible to achieve cumulativity of findings – does not appear to have in any way been alleviated by the introduction of LLMs, but – if anything – exacerbated. While some of the papers do recognize the challenge implied by the lack of a shared validation framework or benchmarks [65, 66, 69, 81, 83], and some even propose new frameworks [70, 74, 77, 84], such suggestions have been long-standing features of the ABM literature, and the challenge of negotiating the tradeoffs between versatility and standardization is not made easier by the introduction of LLMs. Historically, such frameworks have either seen limited re-use, or do not constraint ABMs enough to achieve standardization.

The introduction of LLMs also supercharges the long-standing challenge of the high computational costs associated to ABMs. While large simulations were always costly, the costs of conventional models appear modest compared to those of LLM-based social simulation. Some studies attempt to mitigate these costs by optimizing agent implementations [65, 83], leveraging distributed processing [87] or replacing less critical agents with traditional ABMs [86]. However, these strategies do not change the fundamental reality: generative ABMs are orders of magnitude more computationally demanding than conventional ABMs—already criticized for their high costs. As a result, none of the existing generative ABM studies conduct the necessary sensitivity analysis to assess model reliability and identify key parameters. Even more concerning, most rely on a *single model run*. Given the inherent stochasticity of generative ABMs, this approach is highly problematic – akin to drawing conclusions from a single-case study.

## 6 Conclusion

Generative ABM is an exciting and quickly growing field, making use of the capacity of LLMs to mimic human behavior. This paper has situated generative ABMs in the longer history of, and debate surrounding, agent-based modeling. The paper has reviewed the field to assess whether and how the now-rejuvenated field deals with the long-standing challenges that have historically limited the use of ABMs in the social sciences.

Our review has shown that the rapidly growing field is drawing on LLMs to model a wide range of human behavior, ranging from network dynamics to social media simulations. LLMs enable models to appear much more ‘realistic’, and open up for exploring aspects of human behavior that have previously been nearly impossible to model.

However, based on the review of the recent literature, we have also found that the addition of LLMs to ABMs does little to address the challenges with which the field has struggled, often exacerbating rather than resolving the long-standing issues limiting the use of ABMs.

The lack of rigor in relation to issues such as validation is partly excused by the fact that the field of generative ABMs is still in a stage of early experimentation, exploring the potential of using LLMs to reproduce aspects of human behavior through proof-of-concept models. This begs the question of whether field will be able to transition from such toy models to the type of rigorous modeling needed to contribute productively to social scientific theory.

To do so, it would be necessary to evidence operational validity [42]: the model must be convincingly shown to capture some existing social mechanism. For conventional ABMs, there have been two chief ways of achieving operational validity.

The first way is to employ carefully calibration and validation to achieve model realism [102]. LLMs appear to naturally afford this route, as they enable seemingly more realistic representations of human reasoning and decision-making. However, the introduction of LLMs do not appear to resolve the issues that made such validation and calibration challenging for ABMs, such as their inherent flexibility and lack of standardization and comparability. LLMs furthermore add to this the challenging task of developing calibration and validation procedures to culturally align the agents with real-world individuals, for instance showing that the model is not representing minorities by reproducing problematic stereotypes.

The second way to achieve operational validity is to focus on capturing highly simple – and therefore robust and generalizable – emergent phenomena. Many of the most insightful ABMs are highly simplistic thought-experiments that throw light on a minimal representation of a phenomenon. As such, these models do not necessarily “aim to provide an accurate representation of a particular empirical application. Instead, the goal of agent-based modeling is to enrich our understanding of fundamental processes that may appear in a variety of applications” [103, 25]. Scholars who take this approach to ABMs are “much more concerned with theoretical development and explanation than with prediction” [104, 21]; ABMs offer a form of computational thought-experiments that enable exploring whether particular micro-mechanisms can produce an observed pattern. Such models are believable not because of their realism but because the emergent phenomenon that they capture is so simple and general as to be applicable to a wide range of systems. The famous Schelling segregation model [34], for example, is insightful not because the model accurately represents cities, but because the dynamics that it describes can offer insights into phenomena ranging from residential segregation to why oil separates from water.

However, it is not clear how generative ABMs would be useful for either of these two approaches. For the former, generative ABMs are, as we have seen, exceedingly challenging to calibrate and validate against real-world data. It has yet to be shown that the models can be made to reproduce human actions in such a way as to make them social scientifically productive. For the latter, generative ABMs rely on highly complicated black-box models whose behaviors are poorly understood even on their own, and

it is hence not clear how these models may be used to reveal simple emergent mechanisms. While LLMs are exceptionally simple to use, as one can instruct the agents through natural language rather than employing the often complex mathematical and highly parameterized work of conventional ABMs, they can yet be highly challenging to interpret. As Axelrod [103, 27] puts it, “If the goal is to deepen our understanding of some fundamental process, then simplicity of the assumptions is important and realistic representation of all the details of a particular setting is not”. The introduction of LLMs into these models inevitably adds complexity that may undermine their usefulness as tools for theoretical research, as it makes it more challenging to figure out why and how the model produces a given result [36]. Thought-experiments are productive because they are so simple as to constitute explanations, and it is not clear whether generative ABMs will afford such simplicity.

The issues raised by this paper beg the more fundamental question of whether there is an added value to introducing LLMs in ABMs, beyond the undeniable excitement of novelty. While generative ABMs may exacerbate rather than resolve the long-standing challenges of ABMs, there are still reasons to be optimistic about their possibilities. A range of scholars have long criticized ABMs on epistemic grounds, arguing that complexity in the social world is fundamentally different from type of complexity that ABMs enable us to study [1, 4, 105, 106]. Human society is not akin to a murmuration of starlings: it is characterized not by emergent patterns that stem from the mass-interaction of simple agents, but from agents that *recognize* and *interact* with the social context within which they are operating. In the social world, emergent patterns are not just self-organizing structures; they become named, institutionalized, and capable of exerting downward causation, shaping individual and collective behavior. A potential answer to the question of the novelty brought by generative ABMs is hence that they open the door to studying yet unexplored aspects of social emergence, such as the central role of narrative [107] and social construction [108] in the human world. However, doing so requires addressing the challenges raised in this paper, to avoid the door opening merely to a new replication crisis.

## References

- [1] Byrne, D., Callaghan, G.: Complexity Theory and the Social Sciences: The State of the Art. Routledge, London (2022)
- [2] Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., et al.: The rise and potential of large language model based agents: A survey. arXiv preprint arXiv:2309.07864 (2023)
- [3] Epstein, J.M.: Generative Social Science: Studies in Agent-based Computational Modeling. Princeton University Press, ??? (2012)
- [4] Törnberg, P., Uitermark, J.: Seeing Like a Platform: An Inquiry Into the Condition of Digital Modernity. Taylor & Francis, New York (2025)
- [5] Heath, B., Hill, R., Ciarallo, F.: A survey of agent-based modeling practices

(january 1998 to july 2008). Journal of Artificial Societies and Social Simulation **12**(4), 9 (2009)

- [6] Fagiolo, G., Moneta, A., Windrum, P.: A critical guide to empirical validation of agent-based models in economics: Methodologies, procedures, and open problems. Computational Economics **30**, 195–226 (2007)
- [7] Windrum, P., Fagiolo, G., Moneta, A.: Empirical validation of agent-based models: Alternatives and prospects. Journal of Artificial Societies and Social Simulation **10**(2), 8 (2007)
- [8] Ormerod, P., Rosewell, B.: Validation and verification of agent-based models in the social sciences. In: International Workshop on Epistemological Aspects of Computer Simulation in the Social Sciences, pp. 130–140 (2006). Springer
- [9] Epstein, J.M., Axtell, R.: Growing Artificial Societies: Social Science from the Bottom Up. Brookings Institution Press, New York (1996)
- [10] Naylor, T.H., Finger, J.M.: Verification of computer simulation models. Management science **14**(2), 92 (1967)
- [11] Helbing, D.: Social Self-organization: Agent-based Simulations and Experiments to Study Emergent Social Behavior. Springer, London (2012)
- [12] Conte, R., Paolucci, M.: On agent-based modeling and computational social science. Frontiers in psychology **5**, 668 (2014)
- [13] Törnberg, P., Valeeva, D., Uitermark, J., Bail, C.: Simulating social media using large language models to evaluate alternative news feed algorithms. arXiv preprint arXiv:2310.05984 (2023)
- [14] Park, J.S., O'Brien, J., Cai, C.J., Morris, M.R., Liang, P., Bernstein, M.S.: Generative agents: Interactive simulacra of human behavior. In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, pp. 1–22 (2023)
- [15] Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N.V., Wiest, O., Zhang, X.: Large language model based multi-agents: A survey of progress and challenges. arXiv preprint arXiv:2402.01680 (2024)
- [16] Chan, C.-M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., Fu, J., Liu, Z.: Chat-eval: Towards better llm-based evaluators through multi-agent debate. arXiv preprint arXiv:2308.07201 (2023)
- [17] Xiao, B., Yin, Z., Shan, Z.: Simulating public administration crisis: A novel generative agent-based simulation system to lower technology barriers in social science research. arXiv preprint arXiv:2311.06957 (2023)

- [18] Hua, W., Fan, L., Li, L., Mei, K., Ji, J., Ge, Y., Hemphill, L., Zhang, Y.: War and peace (waragent): Large language model-based multi-agent simulation of world wars. arXiv preprint arXiv:2311.17227 (2023)
- [19] Horton, J.J.: Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research (2023)
- [20] Li, N., Gao, C., Li, Y., Liao, Q.: Large language model-empowered agents for simulating macroeconomic activities. Available at SSRN 4606937 (2023)
- [21] Williams, R., Hosseinichimeh, N., Majumdar, A., Ghaffarzadegan, N.: Epidemic modeling with generative agents. arXiv preprint arXiv:2307.04986 (2023)
- [22] Park, J.S., Popowski, L., Cai, C., Morris, M.R., Liang, P., Bernstein, M.S.: Social simulacra: Creating populated prototypes for social computing systems. In: Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology, pp. 1–18 (2022)
- [23] Gao, C., Lan, X., Lu, Z., Mao, J., Piao, J., Wang, H., Jin, D., Li, Y.: S3: Social-network simulation system with large language model-empowered agents. arXiv preprint arXiv:2307.14984 (2023)
- [24] Aher, G.V., Arriaga, R.I., Kalai, A.T.: Using large language models to simulate multiple humans and replicate human subject studies. In: International Conference on Machine Learning, pp. 337–371 (2023). PMLR
- [25] Kovač, G., Portelas, R., Dominey, P.F., Oudeyer, P.-Y.: The socialai school: Insights from developmental psychology towards artificial socio-cultural agents. arXiv preprint arXiv:2307.07871 (2023)
- [26] (FAIR)†, M.F.A.R.D.T., Bakhtin, A., Brown, N., Dinan, E., Farina, G., Flaherty, C., Fried, D., Goff, A., Gray, J., Hu, H., *et al.*: Human-level play in the game of diplomacy by combining language models with strategic reasoning. Science **378**(6624), 1067–1074 (2022)
- [27] Xu, Y., Wang, S., Li, P., Luo, F., Wang, X., Liu, W., Liu, Y.: Exploring large language models for communication games: An empirical study on werewolf. arXiv preprint arXiv:2309.04658 (2023)
- [28] Qian, C., Cong, X., Yang, C., Chen, W., Su, Y., Xu, J., Liu, Z., Sun, M.: Communicative agents for software development. arXiv preprint arXiv:2307.07924 **6**, 3 (2023)
- [29] Mandi, Z., Jain, S., Song, S.: Roco: Dialectic multi-robot collaboration with large language models. In: 2024 IEEE International Conference on Robotics and Automation (ICRA), pp. 286–299 (2024). IEEE

- [30] Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., Du, M.: Explainability for Large Language Models: A Survey (2023). <https://arxiv.org/abs/2309.01029>
- [31] Luo, H., Specia, L.: From understanding to utilization: A survey on explainability for large language models. arXiv preprint arXiv:2401.12874 (2024)
- [32] Reynolds, C.W.: Flocks, herds and schools: A distributed behavioral model. In: Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques, pp. 25–34 (1987)
- [33] Russell, S.J., Norvig, P.: Artificial Intelligence: a Modern Approach. Pearson, ??? (2016)
- [34] Schelling, T.C.: Dynamic models of segregation. *Journal of mathematical sociology* **1**(2), 143–186 (1971)
- [35] Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., Yang, D.: Can large language models transform computational social science? *Computational Linguistics* **50**(1), 237–291 (2024)
- [36] Macy, M.W., Willer, R.: From factors to actors: Computational sociology and agent-based modeling. *Annual review of sociology* **28**(1), 143–166 (2002)
- [37] Silverman, B., Bharathy, G., Kim, G.: The new frontier of agent-based modeling and simulation of social systems with country databases, newsfeeds, and expert surveys. *Agents, Simulation and Applications*, A. Uhrmacher and D. Weyns,(eds.) Taylor and Francis (2009)
- [38] Bertolotti, F., Locoro, A., Mari, L.: Sensitivity to initial conditions in agent-based models. In: Multi-Agent Systems and Agreement Technologies: 17th European Conference, EUMAS 2020, and 7th International Conference, AT 2020, Thessaloniki, Greece, September 14-15, 2020, Revised Selected Papers 17, pp. 501–508 (2020). Springer
- [39] De Marchi, S.: Computational and Mathematical Modeling in the Social Sciences. Cambridge University Press, ??? (2005)
- [40] Wilensky, U., Rand, W.: Making models match: Replicating an agent-based model. *Journal of Artificial Societies and Social Simulation* **10**(4), 2 (2007)
- [41] Epstein, J.M., Axtell, R.: Artificial societies and generative social science. *Artificial Life and Robotics* **1**, 33–34 (1997)
- [42] Sargent, R.G.: Verification and validation of simulation models. In: Proceedings of the 2010 Winter Simulation Conference, pp. 166–183 (2010). IEEE
- [43] Törnberg, P.: Complex realist economics: toward an ontology for an interested

- pluralism. *Review of Social Economy* **76**(4), 509–534 (2018)
- [44] Baker, E.: The ultimate think tank: The rise of the santa fe institute libertarian. *History of the Human Sciences* **35**(3-4), 32–57 (2022)
- [45] Zhao, P., Jin, Z., Cheng, N.: An in-depth survey of large language model-based artificial intelligence agents. arXiv preprint arXiv:2309.14365 (2023)
- [46] Jones, C.R., Bergen, B.K.: People cannot distinguish gpt-4 from a human in a turing test. arXiv preprint arXiv:2405.08007 (2024)
- [47] Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., et al.: A survey on large language model based autonomous agents. *Frontiers of Computer Science* **18**(6), 186345 (2024)
- [48] Cheng, Y., Zhang, C., Zhang, Z., Meng, X., Hong, S., Li, W., Wang, Z., Wang, Z., Yin, F., Zhao, J., et al.: Exploring large language model based intelligent agents: Definitions, methods, and prospects. arXiv preprint arXiv:2401.03428 (2024)
- [49] Mou, X., Ding, X., He, Q., Wang, L., Liang, J., Zhang, X., Sun, L., Lin, J., Zhou, J., Huang, X., et al.: From individual to society: A survey on social simulation driven by large language model-based agents. arXiv preprint arXiv:2412.03563 (2024)
- [50] Zini, J.E., Awad, M.: On the explainability of natural language processing deep models. *ACM Computing Surveys* **55**(5), 1–31 (2022)
- [51] Ferrara, E.: Should chatgpt be biased? challenges and risks of bias in large language models. arXiv preprint arXiv:2304.03738 (2023)
- [52] Navigli, R., Conia, S., Ross, B.: Biases in large language models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality* **15**(2), 1–21 (2023)
- [53] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems* (2023)
- [54] Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., et al.: Siren’s song in the ai ocean: a survey on hallucination in large language models. arXiv preprint arXiv:2309.01219 (2023)
- [55] Christopher Frey, H., Patil, S.R.: Identification and review of sensitivity analysis methods. *Risk analysis* **22**(3), 553–578 (2002)
- [56] Turing, A.M.: Computing Machinery and Intelligence. Springer, ??? (2009)

- [57] Wohlin, C.: Guidelines for snowballing in systematic literature studies and a replication in software engineering. In: Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering, pp. 1–10 (2014)
- [58] Feng, J., Cui, C., Zhang, C., Fan, Z.: Large language model based agent framework for electric vehicle charging behavior simulation. arXiv preprint arXiv:2408.05233 (2024)
- [59] Ren, R., Qiu, P., Qu, Y., Liu, J., Zhao, W.X., Wu, H., Wen, J.-R., Wang, H.: Bases: Large-scale web search user simulation with large language model based agents. arXiv preprint arXiv:2402.17505 (2024)
- [60] Argyle, L.P., Busby, E.C., Fulda, N., Gubler, J.R., Rytting, C., Wingate, D.: Out of one, many: Using language models to simulate human samples. *Political Analysis* **31**(3), 337–351 (2023)
- [61] Li, G., Hammoud, H., Itani, H., Khizbullin, D., Ghanem, B.: Camel: Communicative agents for “mind” exploration of large language model society. *Advances in Neural Information Processing Systems* **36**, 51991–52008 (2023)
- [62] Hong, S., Zheng, X., Chen, J., Cheng, Y., Wang, J., Zhang, C., Wang, Z., Yau, S.K.S., Lin, Z., Zhou, L., et al.: Metagpt: Meta programming for multi-agent collaborative framework. arXiv preprint arXiv:2308.00352 (2023)
- [63] Jinxin, S., Jiabao, Z., Yilei, W., Xingjiao, W., Jiawen, L., Liang, H.: Cgmi: Configurable general multi-agent interaction framework. arXiv preprint arXiv:2308.12503 (2023)
- [64] Wu, Q., Bansal, G., Zhang, J., Wu, Y., Zhang, S., Zhu, E., Li, B., Jiang, L., Zhang, X., Wang, C.: Autogen: Enabling next-gen llm applications via multi-agent conversation framework. arXiv preprint arXiv:2308.08155 (2023)
- [65] Kaiya, Z., Naim, M., Kondic, J., Cortes, M., Ge, J., Luo, S., Yang, G.R., Ahn, A.: Lyfe agents: Generative agents for low-cost real-time social interactions. arXiv preprint arXiv:2310.02172 (2023)
- [66] De Marzo, G., Pietronero, L., Garcia, D.: Emergence of scale-free networks in social interactions among large language models. arXiv preprint arXiv:2312.06619 (2023)
- [67] Wang, Z., Chiu, Y.Y., Chiu, Y.C.: Humanoid agents: Platform for simulating human-like generative agents. arXiv preprint arXiv:2310.05418 (2023)
- [68] Zhang, J., Xu, X., Zhang, N., Liu, R., Hooi, B., Deng, S.: Exploring collaboration mechanisms for llm agents: A social psychology view. arXiv preprint arXiv:2310.02124 (2023)

- [69] Chen, W., Su, Y., Zuo, J., Yang, C., Yuan, C., Qian, C., Chan, C.-M., Qin, Y., Lu, Y., Xie, R., *et al.*: Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents. arXiv preprint arXiv:2308.10848 **2**(4), 5 (2023)
- [70] Altera.al, Ahn, A., Becker, N., Carroll, S., Christie, N., Cortes, M., Demirci, A., Du, M., Li, F., Luo, S., et al.: Project sid: Many-agent simulations toward ai civilization. arXiv preprint arXiv:2411.00114 (2024)
- [71] Li, H., Chong, Y.Q., Stepputtis, S., Campbell, J., Hughes, D., Lewis, M., Sycara, K.: Theory of mind for multi-agent collaboration via large language models. arXiv preprint arXiv:2310.10701 (2023)
- [72] Li, X., Xu, Y., Zhang, Y., Malthouse, E.C.: Large language model-driven multi-agent simulation for news diffusion under different network structures. arXiv preprint arXiv:2410.13909 (2024)
- [73] Zhang, D., Li, Z., Wang, P., Zhang, X., Zhou, Y., Qiu, X.: Speechagents: Human-communication simulation with multi-modal multi-agent systems. arXiv preprint arXiv:2401.03945 (2024)
- [74] Liu, Y., Chen, X., Zhang, X., Gao, X., Zhang, J., Yan, R.: From skepticism to acceptance: Simulating the attitude dynamics toward fake news. arXiv preprint arXiv:2403.09498 (2024)
- [75] Zhou, X., Zhu, H., Mathur, L., Zhang, R., Yu, H., Qi, Z., Morency, L.-P., Bisk, Y., Fried, D., Neubig, G., et al.: Sotopia: Interactive evaluation for social intelligence in language agents. arXiv preprint arXiv:2310.11667 (2023)
- [76] Zhang, Q., Naradowsky, J., Miyao, Y.: Self-emotion blended dialogue generation in social simulation agents. arXiv preprint arXiv:2408.01633 (2024)
- [77] Li, Y., Zhang, Y., Sun, L.: Metaagents: Simulating interactions of human behaviors for llm-based task-oriented coordination via collaborative generative agents. arXiv preprint arXiv:2310.06500 (2023)
- [78] Xie, C., Chen, C., Jia, F., Ye, Z., Shu, K., Bibi, A., Hu, Z., Torr, P., Ghanem, B., Li, G.: Can large language model agents simulate human trust behaviors? arXiv preprint arXiv:2402.04559 (2024)
- [79] Ren, S., Cui, Z., Song, R., Wang, Z., Hu, S.: Emergence of social norms in large language model-based agent societies. arXiv preprint arXiv:2403.08251 (2024)
- [80] He, J.K., Wallis, F.P.S., Rathje, S.: Homophily in an artificial social network of agents powered by large language models. PsyArXiv (2023)
- [81] Wu, Z., Peng, R., Zheng, S., Liu, Q., Han, X., Kwon, B., Onizuka, M., Tang, S.,

- Xiao, C.: Shall we team up: Exploring spontaneous cooperation of competing llm agents. In: Findings of the Association for Computational Linguistics: EMNLP 2024, pp. 5163–5186 (2024)
- [82] Jarrett, D., Pislar, M., Bakker, M.A., Tessler, M.H., Koster, R., Balaguer, J., Elie, R., Summerfield, C., Tacchetti, A.: Language agents as digital representatives in collective decision-making. In: NeurIPS 2023 Foundation Models for Decision Making Workshop (2023)
- [83] Yu, Y., Zhang, Q., Li, J., Fu, Q., Ye, D.: Affordable generative agents. arXiv preprint arXiv:2402.02053 (2024)
- [84] Chuang, Y.-S., Harlalka, N., Suresh, S., Goyal, A., Hawkins, R., Yang, S., Shah, D., Hu, J., Rogers, T.T.: The wisdom of partisan crowds: Comparing collective intelligence in humans and llm-based agents. In: Proceedings of the Annual Meeting of the Cognitive Science Society, vol. 46 (2024)
- [85] Liu, Y., Song, Z., Zhang, X., Chen, X., Yan, R.: From a tiny slip to a giant leap: An llm-based simulation for fake news evolution. arXiv preprint arXiv:2410.19064 (2024)
- [86] Mou, X., Wei, Z., Huang, X.: Unveiling the truth and facilitating change: Towards agent-based large-scale social movement simulation. arXiv preprint arXiv:2402.16333 (2024)
- [87] Yang, Z., Zhang, Z., Zheng, Z., Jiang, Y., Gan, Z., Wang, Z., Ling, Z., Chen, J., Ma, M., Dong, B., et al.: Oasis: Open agents social interaction simulations on one million agents. arXiv preprint arXiv:2411.11581 (2024)
- [88] Sreedhar, K., Chilton, L.: Simulating human strategic behavior: Comparing single and multi-agent llms. arXiv preprint arXiv:2402.08189 (2024)
- [89] Liu, Y., Liu, W., Gu, X., Rui, Y., He, X., Zhang, Y.: LMAgent: A Large-scale Multimodal Agents Society for Multi-user Simulation (2024). <https://arxiv.org/abs/2412.09237>
- [90] Gu, C., Luo, L., Zaidi, Z.R., Karunasekera, S.: Large Language Model Driven Agents for Simulating Echo Chamber Formation (2025). <https://arxiv.org/abs/2502.18138>
- [91] Orlando, G.M., Gatta, V.L., Russo, D., Moscato, V.: Can Generative Agent-Based Modeling Replicate the Friendship Paradox in Social Media Simulations? (2025). <https://arxiv.org/abs/2502.05919>
- [92] Wang, L., Zhang, J., Yang, H., Chen, Z.-Y., Tang, J., Zhang, Z., Chen, X., Lin, Y., Sun, H., Song, R., et al.: User behavior simulation with large language model-based agents. ACM Transactions on Information Systems **43**(2), 1–37

(2025)

- [93] Ferraro, A., Galli, A., La Gatta, V., Postiglione, M., Orlando, G.M., Russo, D., Riccio, G., Romano, A., Moscato, V.: Agent-based modelling meets generative ai in social network simulations. Lecture Notes in Computer Science (including sub-series Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **15211 LNCS**, 155–170 (2025) [https://doi.org/10.1007/978-3-031-78541-2\\_10](https://doi.org/10.1007/978-3-031-78541-2_10) . Cited by: 0
- [94] Zheng, W., Tang, X.: Simulating social network with llm agents: An analysis of information propagation and echo chambers. Communications in Computer and Information Science **2269 CCIS**, 63–77 (2025) [https://doi.org/10.1007/978-981-96-0178-3\\_5](https://doi.org/10.1007/978-981-96-0178-3_5) . Cited by: 0
- [95] Schlitz, M.M., Vieten, C., Miller, E.M.: Worldview transformation and the development of social consciousness. Journal of Consciousness Studies **17**(7-8), 18–36 (2010)
- [96] Bhattacharjee, A., Liu, H.: Fighting fire with fire: can chatgpt detect ai-generated text? ACM SIGKDD Explorations Newsletter **25**(2), 14–21 (2024)
- [97] Karpinska, M., Akoury, N., Iyyer, M.: The perils of using mechanical turk to evaluate open-ended text generation. arXiv preprint arXiv:2109.06835 (2021)
- [98] Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Liu, Q., Liu, T., Sui, Z.: Large language models are not fair evaluators. arXiv preprint arXiv:2305.17926 (2023)
- [99] Panickssery, A., Bowman, S.R., Feng, S.: Llm evaluators recognize and favor their own generations. arXiv preprint arXiv:2404.13076 (2024)
- [100] Poláková, P., Ivenz, P., Klímová, B.: Examining the reliability of chatgpt as an assessment tool compared to human evaluators. Procedia Computer Science **246**, 2332–2341 (2024)
- [101] Boelaert, J., Coavoux, S., Ollion, É., Petev, I., Präg, P.: Machine bias. how do generative language models answer opinion polls? OSF (2024)
- [102] Edmonds, B., Moss, S.: From kiss to kids—an ‘anti-simplistic’ modelling approach. In: International Workshop on Multi-agent Systems and Agent-based Simulation, pp. 130–144 (2004). Springer
- [103] Axelrod, R.: Advancing the art of simulation in the social sciences. In: Simulating Social Phenomena, pp. 21–40. Springer, ??? (1997)
- [104] Gilbert, N.: A simulation of the structure of academic science. Sociological research online **2**(2), 91–105 (1997)

- [105] Bhaskar, R.: *The Possibility of Naturalism: A Philosophical Critique of the Contemporary Human Sciences*. Routledge, London (2014)
- [106] Bateson, G.: *Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology*. University of Chicago press, Chicago (2000)
- [107] Polkinghorne, D.: *Narrative Knowing and the Human Sciences*. Suny Press, ??? (1988)
- [108] Berger, P., Luckmann, T.: *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*. Penguin Books, New York (1966)



