Bojić *et al.*
European Journal of Futures Research (2024) 12:15
https://doi.org/10.1186/s40309-024-00238-0

European Journal
of Futures Research

RESEARCH ARTICLE

CERN for AI: a theoretical framework for autonomous simulation-based artificial intelligence testing and alignment

Ljubiša Bojić[1,3*], Matteo Cinelli[2], Dubravko Ćulibrk[3] and Boris Delibašić[4]

Abstract

This paper explores the potential of a multidisciplinary approach to testing and aligning artificial intelligence (AI), specifically focusing on large language models (LLMs). Due to the rapid development and wide application of LLMs, challenges such as ethical alignment, controllability, and predictability of these models emerged as global risks. This study investigates an innovative simulation-based multi-agent system within a virtual reality framework that replicates the real-world environment. The framework is populated by automated 'digital citizens,' simulating complex social structures and interactions to examine and optimize AI. Application of various theories from the fields of sociology, social psychology, computer science, physics, biology, and economics demonstrates the possibility of a more human-aligned and socially responsible AI. The purpose of such a digital environment is to provide a dynamic platform where advanced AI agents can interact and make independent decisions, thereby mimicking realistic scenarios. The actors in this digital city, operated by the LLMs, serve as the primary agents, exhibiting high degrees of autonomy. While this approach shows immense potential, there are notable challenges and limitations, most significantly the unpredictable nature of real-world social dynamics. This research endeavors to contribute to the development and refinement of AI, emphasizing the integration of social, ethical, and theoretical dimensions for future research.

Keywords AI Alignment, Social Science in Artificial Intelligence, Theoretical Framework, Digital City Simulation, Autonomy in AI

Introduction

The rapid evolution and expansion of artificial intelligence (AI), especially in the domain of natural language processing (NLP), has proven to be a promising frontier in technological development. AI-driven applications, particularly those based on Generative Pretrained Transformers (GPT), possess the potential to revolutionize various sectors of society by transforming processes, interactions and services, presenting many possibilities that were previously unimaginable [32]. The burst of innovative technologies based on AI, such as recommendation algorithms, chatbots, autonomous vehicles, and even complex financial trading strategies, have somewhat become part of our daily lives, integrating their functionalities globally across numerous industries and sectors.

With the increased reliance and adoption of such AI systems, numerous challenges pertaining to the alignment with human ethics and values, controllability, transparency and predictability of these models arise, therefore warranting further attention and investment in terms of research and development. While AI has made significant strides in decision-making and

*Correspondence:
Ljubiša Bojić
ljubisa.bojic@ivi.ac.rs
[1] Digital Society Lab, University of Belgrade, Institute for Philosophy and Social Theory, Belgrade, Fruskogorska 1, 21000 Novi Sad, Serbia
[2] Department of Social Sciences and Economics, University of Rome La Sapienza, Piazzale Aldo Moro 5, 00185 Rome, Italy
[3] The Institute for Artificial Intelligence Research and Development of Serbia, Fruskogorska 1, 21000 Novi Sad, Serbia
[4] Faculty of Organizational Sciences, University of Belgrade, Jove Ilica 154, 11000 Belgrade, Serbia

task-completion competencies, achieving alignment with human values, predictability and full controllability, especially for large-scale neural networks, remains a stumbling block in the evolution of this powerful technology [75].

Recent contributions in AI research have focused on Generative Pretrained Transformers (GPT), a model that employs machine learning algorithms to improve the generation of human-like text [92]. However, the rise of these highly-potent AI models has amplified concerns about their capacity for ethical alignment, controllability, and the unpredictability often inherent in large-scale neural networks [63].

There is insufficient understanding of how these models would behave in complex social dynamics and unfolding scenarios mirroring the real-world. This has put pressure on AI stakeholders to explore better testing and mitigation strategies [32]. Simultaneously, the need to ensure AI models' security is an imperative concern which is made paramount due to the profound potential impacts of implementing them for societal and commercial purposes [32, 63].

Definitions

Large language models (LLMs) are artificial intelligence neural networks capable of language generation, translation, question answering and summarization [92]. Aside from text generation, LLMs also exhibit the capacity to simulate understanding of inquiries and perform complex cognitive tasks [98]. LLMs have demonstrated relatively high performance in a wide range of tasks and languages without any task-specific training [92].

Artificial General Intelligence (AGI) is a futuristic concept, which refers to a type of AI with cognitive capabilities that can successfully understand, learn, and implement wide range of intellectual tasks equivalent to those of a human being [53].

AI Alignment represents the proposition of ensuring that the behavior of AI system is congruent with human intentions and values. The development of AI might lead to an intelligence explosion where AI surpasses human intelligence. If such a situation arises, it is important to ensure that AI is beneficially aligned and promotes the interests of humanity [21]. Thus, research is needed to ensure that AI development is carried out with necessary precautions.

Among other platforms, OpenAI's LLMs stand out due to their potential for fine-tuning, making them compatible with a wide range of use-cases. This adaptability sets the stage for their comprehensive influence and application across diverse fields. OpenAI develops AGI while attempting to devise strategies that ensure its safety and alignment with human values [3].

Practical applications leading towards simulation-based AI testing

LLMs can be given various degrees of autonomy while creating multiple agents with different prompts capable of interacting with each other. There are three notable applications of LLMs in the direction of simulations and autonomy: Auto-GPT [78], Interactive LLM Powered NPC [2] and AI Town [38].

Auto-GPT is an experimental, open-source autonomous AI agent, created on the underlying principles of the GPT-4 language model [78]. It is designed to autonomously chain together various tasks in order to achieve a bigger, overarching goal as set by the user. Unlike traditional chatbots like ChatGPT that require multiple prompts to function, Auto-GPT operates by automating this multi-step prompting procedure. The user merely has to provide a single initial prompt or a set of instructions coded in natural language, and Auto-GPT handles the rest, breaking down the provided goal into a series of manageable subtasks to accomplish its objective. Thus, Auto-GPT can be employed in similar ways as ChatGPT, but with the added advantage of automation, thereby ensuring quicker task completion. It also boasts internet integration, thereby allowing it to access and use real-time data. Logo of Auto-GPT is depicted on Fig. 1.

Interactive LLM Powered NPCs is an open-source project aiming to improve player interaction with non-player characters (NPCs) in video games [2]. The project transforms static conversations with NPCs to dynamic ones, allowing players to engage in immersive dialogues, recognizing their voice and showing lifelike animations of the characters (Fig. 2). The technology used includes

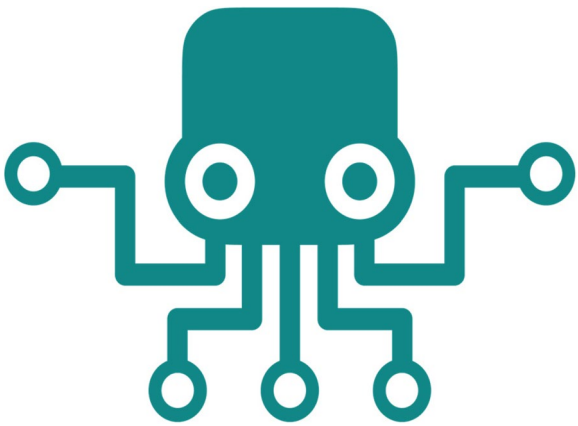

Fig. 1 A simple digital vector art of an octopus like creature, used as the logo of Auto GPT [6]

Fig. 2 Illustration of an interaction of non-playable characters in a game using Interactive LLM Powered NPCs technology [37]

facial animation to sync character lip movements, facial recognition to differentiate characters, and vector stores to give NPCs limitless memory capacity. It also uses a pre-conversation file to shape the dialogue style of each character, making the interaction more lifelike. The NPC adjusts based on its specific personality, knowledge, and communication style, and is capable of perceiving the player's facial expressions via webcam, adding depth to the interaction.

AI Town, developed by Convex.dev, is an innovative virtual town populated by AI characters who interact, chat, and socialize just like human characters [38]. It combines artificial intelligence technology and creative programming to create a dynamic virtual community. In this simulation, every resident is an artificial intelligence entity with specific traits and behaviors, capable of engaging in conversations and social interactions. These AI residents inhabit a digital landscape that mimics real-life towns, complete with architectural and environmental elements (Fig. 3). Users can visit AI Town and interact with its inhabitants in real-time by joining the conversation and immersing themselves in this digital environment. This interface provides a platform to observe AI behavior and language abilities in a social setting and explore advances in chatbot technology. AI Town offers an example into how research and testing of artificial intelligence could be done in a safe, controlled environment.

AI alignment

The insurgence of AI options, while offering significant opportunities for both private and public sectors, have

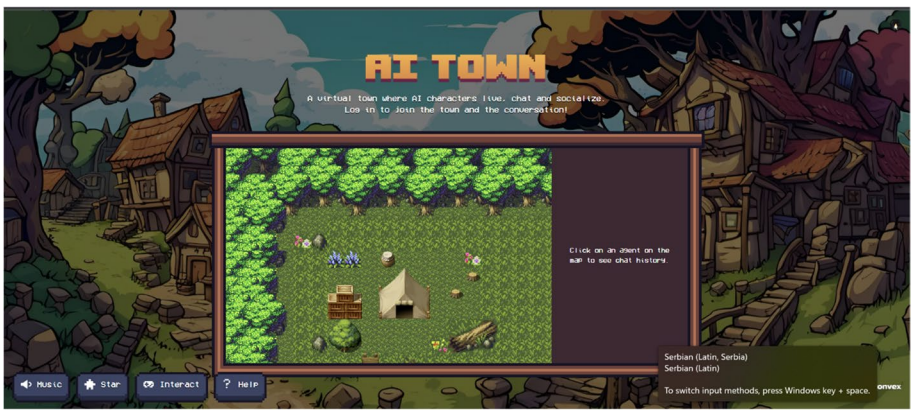

Fig. 3 Screenshot of an AI Town web application [38]

also presented a complex package of associated risks and ethical challenges. When considering previously developed AI systems, traditional methodologies of risk and error management have proven insufficient to mitigate potential harm efficiently and adequately [23, 96]. As a result, the growing body of research on AI safety necessitates an exploration of a multidimensional testing framework for these transformative technologies.

Prominent scholars have emphasized the risks associated with the unpredictability of AI systems. For instance, [63] pointed out that large-scale language models such as GPT can produce behaviors that, although desirable under controlled conditions, may present contentious outputs in unforeseen scenarios. On another note, [5] expressed concerns on the "value alignment problem" with AI, stressing the necessity to hardwire ethical boundaries and human values into AI systems to avoid potentially catastrophic effects of an alignment failure.

Further exploring the domain of AI ethics, [119] and Whittlestone et al. [123] declared the importance of transparency, accountability, and public involvement in AI systems' design and creation processes. Their work accentuates the need for democratized technology and calls for multiple stakeholders' contribution from different industries, sectors, and social-demographic backgrounds.

The incorporation of social science theories has gained ground in recent years due to their potential application in AI development. Rafols [93] and Cave et al. [28] suggested that taking lessons from social research methodologies could be a key ingredient to instill essential social sensibilities into AI systems, thus enriching their potential for more humane design. This marriage of technology and social sciences presents a promising avenue towards a socially-responsible AI future.

Previous research

The use of AI and specifically social simulation models in advancing scientific knowledge is an emerging area of interest in the field of AI. Social simulations as refuted machines have shown a significant contribution towards refuting prevailing theories in science, thus promoting scientific advancements [81].

The capacity of arguments in driving issue polarization has also been studied, indicating a potential role of AI in shaping public discourse [70]. AI's role in polarizing debates among artificial agents can have implications for understanding political polarization in society.

The use of LLMs is also a promising direction in autonomous agent research. Wang et al., [121] have shown that LLMs can acquire vast amounts of web knowledge and demonstrate human-level intelligence, indicating a potential role in diverse areas of social science, natural science, and engineering.

Simultaneously, the use of LLMs as a potential substitute for human participants in psychological research has also been evaluated [43]. Despite concerns about replicating human judgment, some studies show that LLMs can exhibit strong alignment with human judgments, thus suggesting that AI can play a role in replicating human subject studies in certain scenarios [1].

Furthermore, Generative AI is showing potential in improving social science research, online experiments, agent-based models, and content analyses [8]. AI models can help in performing routine tasks, advancing programming skills, and writing more effective prose, which could transform the way social science research is conducted.

The potential of generative AI has been explored in strategic game experiments as well, with results suggesting that AI can generate realistic outcomes, and exhibit human-like behavior under appropriate conditions [56]. AI's potential in mirroring human behavior has been studied through generative agents, interactive simulacra of human behavior [88]. Along these lines, AI has been utilized to create populated prototypes for social computing systems, showing potential in simulating real-world social interactions and behaviors [89].

Research objectives

In light of latest developments in LLMs, the need for responsible AI, recent practical simulation-related applications and previous inquiries, this study aims to answer the following research question (RQ1): which theories and approaches derived from multiple disciplines could be useful as foundations for a coherent, multi-faceted simulation-based testing approach for AI?

This approach aims to create a more comprehensive and stringent process for assessing AI agents, particularly LLMs and within the specific context of a virtual reality framework that simulates life in a digital city. The increased stakes brought on by the sophistication of AI are seen as a pressing call to action for the design of more dependable and human-aligned models through this simulation-based approach [21, 42], which could be useful for refining and fixing these systems [81].

We aim to delve into the depth of the opportunities and limitations provided by the merger of AI and various theoretical approaches, to nurture an effective, socially-responsible and comprehensive approach for testing and aligning LLMs within a digital environment [27, 96].

This paper tackles the novel and complex challenge of setting up a theoretical framework for testing and aligning LLMs. The subsequent sections will explore selected theories from various fields, their application to AI

behavior and alignment, and their relevance in the context of a simulation-based approach. The latter part of the paper will delve into the practical aspects of applying these theories into a simulation-based approach to AI development. This includes the design and operation of a digital city populated by AI entities or 'digital citizens', their interactions and decision-making processes, and the resultant insights and applications. The paper concludes with reflections on the limitations of this approach and possibilities for future research. This exploration is vital in our relentless quest to ensure that AI technologies are effective, secure, and uphold the values of human society.

Methods

In this research, two key methods were used to investigate the utility of a variety of theories from different scientific fields in the creation of a Digital City for AI testing and alignment. These methods include literature review and theoretical analysis.

The first method utilized in this research was an extensive literature review [57]. A broad spectrum of scholarly literature was explored to obtain comprehensive insights into the field under study. These included academic articles, books, and reports covering Artificial Intelligence, LLMs, AGI, social robotics, social simulations, computer science, physics, and economics amongst others. The reviewed literature was selected based on its relevance, the prominence of the authors in the field, and the impact it has had in the scholarly community.

Literature on simulation-based approaches in AI testing and development was thoroughly explored to understand the current methods adopted by researchers and the challenges they face [21, 67, 82]. The insights derived from the reviewed literature were instrumental in formulating the simulation design used in this research.

The goal of the literature review process was not only to understand the existing body of knowledge but also to identify gaps in the current research that our study could address.

The second method used in this study was theoretical analysis [40]. An extensive analysis was conducted to understand how theories from different fields could be applied to the development, testing and alignment of LLMs in a simulated digital city. This integrative and interdisciplinary perspective helped to develop the basis for the 'digital city' concept, highlighting the role these theories can play in AI alignment.

Procedures

The process of identifying, assessing and narrowing down relevant disciplines and theories for inclusion in this research was a meticulous task. From the outset, the undertaking was centered on identifying those theories and constructs with the greatest relevance and application to the field of AI development and testing. The filtering process involved the application of a selection of critical criteria.

Firstly, the theories were assessed for compatibility with the overarching AI concept. A strong correlation with the principles of AI and the ability to enhance its understanding was a prerequisite. The theories also were evaluated based on their potential to contribute a unique perspective on AI behavior or its alignment, considering also cross-disciplinary connections.

The selection process was driven by a comprehensive review of literature in fields like computer science, psychology, sociology, and economics.

The critical aspect of this selection process was pairing the theories with appropriate methodologies or techniques. For instance, Matching the Big Five personality theory with computational personality recognition techniques for the creation of digital citizens or the application of game theory in concert with reinforcement learning approaches for AI decision-making.

The hierarchical clustering process based on theoretical importance, relevance, and application in the AI context was also utilized for selecting the theories and approaches. Hierarchical cluster analysis is a statistical method that builds a hierarchy of clusters iteratively. Starting with individual theories in separate clusters, in each iteration, we merged the two most similar ones until only one cluster was left [49]. Similarity was determined based on theoretical underpinnings, field of study, potential application in AI context, and overall compatibility with AI development philosophies. The dendrogram obtained from this analysis gave us an understanding of the interconnection and relatedness between theories, thus guiding our selection process.

Potential ethical and practical considerations relating to the application of theory to AI development were integrated into the discretionary process. Theories that highlighted ethical dimensions in AI development were seen as providing a cornerstone for the digital city simulation.

The efficacy of matching the theories with real-world machine learning techniques and their compatibility with the simulation design was considered. It was crucial to include the theories with relatively established techniques ensuring pragmatic application in the AI alignment process.

The process concluded with a peer review check, where subject matter experts reviewed our selection process and results to minimize biases and strengthen our chosen theoretical framework. This process of thoroughly reviewing each premise fed into our systematic selection

process of evaluating the relevance and applicability of the theories in the context of AI.

The selection was then consolidated and constructed into a comprehensive framework for the digital city simulation. The entire process, from reviewing literature to choosing relevant theories, was iterative in nature, following the cyclic research model presented by Taylor et al. [112]. It involved constant reevaluation, feedback, and modification, ensuring that the theoretical framework was sound, relevant, and comprehensive.

From over 200 theories and methodologies identified in the early stages of the literature review, we whittled the list down to a focused beam of 10 theories and methodologies from various disciplines, which were deemed most relevant and promising for AI testing and alignment using a digital city simulation. This narrowed focus enabled us to make a deep dive into these select disciplines and theories, ensuring a detailed understanding and relevant application of each in the overarching design and operation of the digital city simulation. Over time, this iterative and detailed approach enabled us to create a unique, first-of-its-kind theoretical framework that seamlessly combines multidisciplinary findings to foster AI development, testing and alignment.

This multi-layered selection process utilized across 120 referenced studies and papers furnished the research with a well-rounded theoretical foundation. It also elevated the richness of the digital environment within our simulation approach while ensuring reliability within the context of AI testing and alignment.

Establishing a framework for interactions in an autonomous digital city

This section presents applicability of theories derived from sociology, biology, physics, social psychology and computer science to be used for the process of testing and aligning LLMs and AI. Complexities of LLMs and AI necessitate a rigorous, theory-based approach to aid in testing and alignment processes and to minimize potential risks [93]. Social science theories are valuable in understanding behavioral patterns and decision-making processes, potentially offering explanatory and predictive abilities for AI behavior [87]. Subsequent subsections will offer an in-depth exploration of selected theories and their relevance within AI's context.

The social simulation and reasoned action theories

Integrating social theories in AI research provides perspectives for understanding AI behavior and alignment. Social theories focus on social relations, structures, and institutions that constitute society. They offer a theoretical lens for understanding social phenomena, behavior, and the intricacies of human interaction, which can be extrapolated for enhancing AI alignment. Particularly, the Social Simulation Theory and the Theory of Reasoned Action could be pivotal in this context.

The Social Simulation Theory stems from the broader spectrum of Computational Social Science [35], emphasizing the power of computational methods to simulate, analyze, and draw insights from complex social phenomena. In the context of AI, this theory would be a key approach for testing LLMs. By simulating complex social dynamics computationally, researchers can create a more dynamic and realistic test environment, shedding light on how AI models might interact in various societal scenarios, and thus how to better align AI behavior to human values and expectations [79]. This theory can contribute to the understanding of AI systems' behavior from multilayered perspectives, which is crucial given the complex and often unforeseen consequences that AI systems can have in society [59]. However, the Social Simulation Theory also faces numerous challenges. One significant hurdle is the inherent unpredictability of real-world social systems. Simulating complex social dynamics in abstracted computational models inevitably involves simplifications, which can limit the accuracy and applicability of the results [44].

Originating from social psychology, Theory of Reasoned Action states that intentions drive individual behavior, formed by attitudes towards the behavior, subjective norms, and perceived behavioral control [50]. While originally designed to understand and predict human behavior, the Theory of Reasoned Action may also be extended to autonomous agents in AI. It can guide the modeling of AI behavior in a virtual environment, thus influencing AI's intentions through programming norms and attitudes. By predicting and understanding the possible actions of AI, this theory could assist in aligning AI's actions with the regulatory norms and societal values [100]. A major challenge posed by this theory is derived from the complex nature of emotions and irrational behavior, which greatly influence human decision-making but might be challenging to replicate in an AI. This complexity highlights the importance of multi-faceted approach to AI alignment.

The situated action theory

The Situated Action Theory is an integral aspect of cognitive science that proposes a shift in viewpoint from the classic prescriptive comprehension of behavior to a more adaptive and situation-dependent one [107]. Applying this theory to AI development offers enhanced capabilities for AI behavior within their digital environments, thus making them more in sync with the dynamism and unpredictability of the real world.

Situated Action Theory contends that behavior is not just an outcome of premade plans but is spontaneously formed through continuous and dynamic interaction with the environment. In light of AI, this translates into AI models capable of proactive response modification as a result of changes in their environment rather than being solely driven by an extinctive set of actions [118]. By doing so, we facilitate advanced AI systems that can independently make decisions based on the situation at hand within the parameters of a digital city, thereby being better equipped to navigate the inherent unpredictability present in large-scale neural networks [41].

A situated cognition model offers a more dynamic outline for AI behavior by coupling the capacity to process information with appropriate autonomy to act in context. It emphasizes the need for cognition to be embedded in an understanding of, and interaction with, the environment and not just merely contemplative or inert [69]. Drawing from cognitive sciences, computational approaches to situated cognition help analyze the interaction of AI with the environment, its dynamics, and adaptability to the perceived settings. Acknowledgment of perceptual-action loops can thus be considered as critical in implementing the Situated Action Theory in AI [45].

Translating these concepts into practical AI programming is a demanding task [69]. Real-world factors are multifaceted and ambiguous, which may be difficult to replicate entirely within a digital environment. The dynamism involved in such a set-up would require models to be accurate enough to induce learning while being resource-efficient [114]. This dichotomy presents a major trade-off challenge needing careful consideration and smart solutions.

Though a formidable task, designing AI's adaptive behavior based on Situated Action Theory provides an avenue to unravel cognitive mechanisms in simulated environments [62]. This paves the way for a sophisticated AI model that is not only capable of extracting information from its environment but can also adjust its behavior based on complex contextual information, strengthening alignment with human values, and contributing to secure, reliable AI systems.

Following the exploration of theoretical perspectives in Part I, the paper will expound on how we can apply the theories and insights from social science, robotics, and artificial intelligence into practical testing and alignment of AI. We will look at the creation of a 'digital city' and 'digital citizens' as a central aspect of our innovative simulation-based approach. Their interactions and decision-making processes within this simulated framework will form a crucial part of our study into autonomous behavior. Throughout, we will also discuss the valuable insights and practical applications of this approach, contributing to the refinement and alignment of AI models.

Complex systems theory

In the quest to design realistic and effective simulations, the utility of Complex Systems Theory cannot be overlooked. As a computational and theoretical method, this approach helps to understand the behavior of systems characterized by intricate webs of interdependencies [83].

Complex Systems Theory is founded on the primal assumption that emergent system behaviors can arise from simple local-level rules and interactions [13]. Moreover, this theory particularly focuses on how small changes in the system can potentially herald large-scale effects—a characteristic often referred to as "sensitivity to initial conditions" or, more colloquially, the "butterfly effect".

The population of a city, be it real or simulated, shares numerous characteristics with complex systems. Both environments encapsulate localized systems or agents that independently follow simple rules but collectively generate emergent behavior at the city level. The inherent structures among these agents form a complex network, very similar to a real city that comprises various social, political, and economic networks interacting simultaneously [14].

In the context of creating digital citizens in a simulated city, Complex Systems Theory is a pertinent guide. It can be harnessed in the development phase to design rules and behaviors for digital citizens. By underscoring the relations and dependencies between the different components or inhabitants of the simulated city, this theory can generate unique richness and complexities [12]. Incorporating it benefits our understanding of societal phenomena, forms the foundation for testing the robustness of AI, and assists in aligning LLMs' behavior in digital citizens [7].

However, applications of this theory into AI development are not devoid of challenges. The notorious difficulty in predicting the complex systemic behavior and the associated implications demand a careful approach that combines continual monitoring, learning, and adjusting of the developed AI systems [61]. This complexity accelerates the need to harness various theories, from social and robotics to psychology and game theory, creating a holistic approach to AI development.

Swarm intelligence

In running complex simulations involving hordes of autonomous agents, such as digital citizens of a simulated city, incorporating Swarm Intelligence can present distinct advantages. Swarm Intelligence, a subset of

Artificial Intelligence, refers to the collective behavior of decentralized and self-organized swarming entities [19]. This behavior, inspired by natural phenomena like fish schooling, bird flocking, and ant colonies, is characterized by collective intelligence that emerges from the interactions between simpler, individual agents [68].

In the context of creating digital citizens, Swarm Intelligence can assist in modelling and facilitating intelligent behaviour from a multitude of interacting entities. Agents can share local information and adjust their behaviors based on this shared knowledge, resulting in emergent global strategies that optimize simulated tasks [66]. Examples abound in algorithms inspired by the behaviors of ants (Ant Colony Optimization algorithms), birds (Particle Swarm Optimization algorithms), and bees (Artificial Bee Colony Algorithms) that can be tailored to run complex simulations in a distributed, self-organized system [16].

Furthermore, Swarm Intelligence offers crucial agents' patrolling abilities [33], a potential necessity in our digital city. For instance, agents can be designed to monitor and maintain the city's different sectors' security, following algorithms based on Swarm Intelligence dynamics.

But careful caution must be exercised in incorporating Swarm Intelligence, as transferring biological concepts of swarming behaviour to AI agents may not always manifest desirable results. Unforeseen system behaviours can emerge from Swarm Intelligence algorithms, yielding unexpected and undesirable outcomes [68]. Further, the parameter selection for Swarm Intelligence algorithms can be a complex task due to the interconnected nature of the parameters [36].

In spite of these challenges, Swarm Intelligence provides additional layers of complex behavior for the digital citizens and bolsters the interdisciplinary approach amalgamating social science theories, robotics, game, visual and complex systems theories. This comprehensive perspective creates a robust foundation for AI development, alignment, and research.

The multi-agent system theory

The Multi-Agent System Theory embodies an important cornerstone in the conception and development of autonomous systems [125]. As we venture further into the domain of AI technologies, especially within the context of testing LLMs, understanding how multiple AI agents can work in conjunction or competition becomes increasingly critical.

Multi-agent systems are collections of several autonomous agents that interact with each other within a specific environment. These agents can be both cooperative and competitive [96]. This scenario is highly congruent with the simulated digital city environment envisaged for testing LLMs. In such an environment, each AI agent can be explored as an individual entity, having its unique attributes, characteristics, and decision-making abilities, according to the goals it has been programmed to achieve.

The theory provides us with insights into the potential interaction landscape of AI systems. By creating a multi-agent system, we allow AI models to interact among themselves as well as with the simulated environment, which can expedite the uncovering of emergent behaviors and systemic weaknesses or strengths [80]. With multiple AI agents, we can generate a range of multi-layered, dynamic scenarios that test the robustness and the adaptiveness of the AI models. This could lead to interesting observations about how AI agents learn to cooperate, compete, and negotiate, mimicking the dynamics of real-world complex systems [101]. These insights can lend crucial guidance in aligning AI decision-making processes to desired outcomes.

However, multi-agent systems are also not without challenges. The issues lying in the implementation of this theory could range from achieving synchronization among agents to dealing with conflicts and competition while reaching shared goals [122]. Furthermore, the architecture of AI systems in a multi-agent setup could swiftly evolve from complex to chaotic with the escalating number of agents. This complexity could make troubleshooting a significant challenge [51]. There are also considerable technical hurdles in ensuring smooth, effective communication between agents in a multi-agent system [76].

Despite these challenges, Multi-Agent System Theory holds substantial potential in shaping the future of simulation-based AI testing [86]. Deeper understanding of this theory's implementation can enhance the precision and efficiency of AI models by providing a broader panorama of their potential interactions and making us better equipped to optimize their functionality.

Creating elements of an autonomous digital city

Automated simulations, when performed in a digital domain, offer the potential for high reproducibility and scalability, mimicking complex, interactive scenarios within a digital city structure. They enable a systematic analysis of "real-life" scenarios. Massively complex systems can be simplified into essential actions and interactions for robust analysis and evaluation [9].

Our digital city employs a multi-agent-based simulation framework, which enables the modeling of a population of digital citizens [15]. Each digital citizen is an autonomous AI agent, modeled with a focus on social characteristics to increase the realism and complexity of interactions. These agent-based models offer an effective

method for understanding complex environments—built on vast interactions among individual agents and their interactions with the environment. Indeed, the use of agent-based models has become commonplace across disciplines, from ecology to economics [58].

Creating a complex, interactive environment within a digital city requires a systematic and iterative process. Initial phases require the creation of autonomous agents, the digital citizens, that can operate within a defined parameter space [104]. Silver et al. highlight that the success of these agents in the digital city directly depends on their autonomy level and the parameter spaces within which they operate. In our context, the agents are modeled on LLMs, giving them the capability to engage in sophisticated interaction, including natural language conversation.

The capacity to act on their own, in other words, their autonomy, defines the digital citizens' dynamics within the city [73]. Lehman et al. point out how autonomy has become a focal point in computing and robotics. It is crucial for the digital citizens to make decisions, conduct actions, and participate in the simulations dynamically.

One crucial aspect of developing the digital city is the environment's meticulous design, where the AI agents operate [85]. The city should not be a mere backdrop, instead, it should function as a grounding influence, shaping the decisions and interactions of our digital citizens. Creating such an environment necessitates a detailed focus on various aspects, such as defining the interaction rules, constraints, and the potential choices for the AI agents [85].

In order to ensure continuity in the learned behavior and refinement of AI agents, Leike et al. [75] propose a reinforcement learning approach. Reinforcement learning embedded in simulation allows for the automation of performance improvement, rewarding actions that lead to desired outcomes and penalizing those that do not. Given the high complexity of interactions and decisions occurring in our digital city, such reinforcement learning becomes crucial for developing AI's secure, value-aligned behavior.

The creation and successful implementation of automated simulations involve numerous critical aspects such as those explored in the subsequent sections: infrastructure, citizens, perception and cognition.

Infrastructure through simulation engines

A simulation-based approach to testing and aligning LLMs represents a critical leap in the quest of enhancing the reliability, performance, and security of AI. Kelly et al. [67] propose an argument on the untapped potential of simulation-based approaches for AI-driven technologies. The authors assert that simulations offer a controlled environment where system behaviors can be analyzed under numerous scenarios, a concept that is crucial in the current study.

The proposed digital city mimics a 'real-world' environment but with controlled variables to assess the performance of AI agents and LLMs [82]. McEwan et al. emphasize the effectiveness of such environments in mimicking complex, interactive scenarios. They invite us to draw on the growing evidence that these simulation-based approaches are indispensable in testing AI systems.

This research uses a simulation-based scenario in a virtual reality framework, a paradigm that holds a growing influence in simulation studies [10]. Deploying a virtual reality framework rather than a basic 2D simulation offers an immersive and interactive platform. Coupled with artificial intelligence, the framework has an enormous potential for transforming the testing, design, and alignment of AI constructs [120]. It offers an opportunity to critically investigate how AI agents interact in life-like scenarios, therefore providing a tool for understanding autonomous behavior.

The incorporation of a virtual reality framework into the simulation approach encompasses a broader context for the use and amplification of AI. As Rouse et al. [95] highlight, advancements in virtual reality have a significant influence on artificial intelligence. The use of virtual reality facilitates the concept of 'digital twin' or a 'mirror world' – an almost realistic, virtual representation of the real world that allows for dynamic interaction and learning [17, 18]. The creation of a virtual environment that simulates a realistic city is not only beneficial for the research on AI behavior, but also opens up possibilities for future anthropological, sociological, and psychological studies [108]. A virtual reality-based city could also be used for understanding social behavioral dynamics. As a result, the efficacy of this method in yielding comprehensive and cross-domain insights is undeniable.

A simulation-based approach is a valuable tool for observing, evaluating, and aligning the behavior of AI agents, particularly LLMs through a more realistic and interactive approach [22]. This approach, underpinned by a virtual reality framework for simulations, could potentially revolutionize our understanding and approach towards advancing AI.

Developing an immersive simulated environment, robust in its complexity and capable of hosting myriad digital citizens, necessitates the integration of powerful simulation engines. These engines, a collection of complex algorithms and computational models, serve as the backbone of the simulation, allowing for the creation, interaction, and evolution of entities within a virtual environment [105].

Simulation engines offer immense flexibility and control, allowing researchers to design varied scenarios, inject contingencies, and monitor the system's evolution in real-time. Several game-based technologies such as the Unity Engine [116], Unreal Engine [46], and Godot Engine [54] possess powerful physics, graphics, and AI capabilities that render them fit for creating interactive, immersive environments with realistic physics.

Unity, for example, allows researchers to create diverse, visually rich environments and control them at both macro and micro levels. It supports scripting in C#, facilitating the implementation of unique AI logic to control digital citizens within these environments. The Unreal Engine, on the other hand, is known for its sophisticated visual rendering capabilities, ideal for creating realistic, high-detail environments. It provides native support for Behavior Trees, a tool used to design complex AI behavior [91].

While game-based engines are more suitable for visual simulations, other specialized simulation environments such as MATSim (Multi-Agent Transport Simulation) and MASON multiagent simulation toolkit are better suited for large-scale urban simulations and social dynamics.

Adoption of these engines comes with its challenges, including the need for specialized knowledge and potential hardware limitations. However, the benefits, such as the ability to observe an AI's behavior in a controlled environment, far outweigh the learning curve and resource demands [126].

Combined with robust theories from social sciences, robotics, game theory, visual and complex systems, and swarm intelligence, simulation engines serve as pivotal tools in developing robust, versatile, and impactful AI.

Citizens—the big five theory

Shaping the digital citizens serves a central role in our simulation of a digital city. It is these digital citizens, the primary agents, exhibiting high degrees of autonomy that bring life to the city and the scenarios unfolding within it. In essence, the behaviors, interactions, decisions, and profiles of these AI agents determine the richness and complexity of the simulated environment [11].

Digital citizens can be viewed as AI actors, playing the part of the city's inhabitants, depicting complex behaviors that could resonate with the population of a real city. By endowing the agents with distinctive characters, norms, and behaviors, we mimic a community's diversity in a real city, thereby enriching the simulation's scenarios and insights.

Creating these digital citizens demands an intense focus on the autonomy level. Autonomy, defined as the capacity to act on one's own, forms the crux of these AI agents' dynamics [106]. Stone and Veloso emphasize that portraying autonomy in robotics or AI systems is critical to determining the value and effectiveness of their performance. High-level autonomy of these LLMs represents both an algorithmic challenge and a conceptual step forward, allowing reactions and proactive engagement with unforeseen events in the digital city to be studied [106].

Understanding and effectively replicating the depth and complexity of human personality and behavior is a non-negligible aspect of creating digital citizens within our simulated environment. To this end, delving into the annals of psychology theories can provide the necessary scaffolding. These theories, such as The Big Five personality traits theory, are remarkably useful in devising the specifics of distinct, unique personalities and behaviors within our digital citizens.

Originally conceived by Goldberg [55] and elaborated upon by Costa and McCrae [39], The Big Five theory suggests five broad dimensions of human personality: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Each dimension represents a spectrum, with individuals falling somewhere along this spectrum, thereby shaping unique personality profiles. This model has gained significant acceptance in psychology due to its comprehensiveness and empirical support [65].

In the context of creating digital citizens, the Big Five theory could potentially govern the nature and degree of variance among the inhabitants' personalities in our simulated city. Digital citizens' behaviors can be programmed based on combinations of these five dimensions, thus achieving diversity in behavioral patterns akin to a real-world city populace. This not only enhances the realism and richness of the simulation but also provides a structured means of attributing coherent, consistent behavior patterns to individual AI agents [127].

Applying such psychological theories to AI also involves the task of embodying a breadth of human emotions and motivations, such as intricacies of emotional intelligence and social cognition. Models like the Theory of Mind, which reflects the ability to attribute mental states to oneself and others [90], can enhance digital citizens' interactivity, leading to a more dynamic and authentic simulation. Humans use the Theory of Mind in everyday interactions and empathic understanding, and incorporating this element into AI systems could allow them to predict and respond more flexibly and naturally.

To accurately implement psychology theories in digital citizens, research from computational personality recognition could be utilized, wherein machine learning and natural language processing techniques are used to detect and assign appropriate personality characteristics [117].

Several different learning models can be implemented in creating digital citizens. For instance, reinforcement learning models the dimension in which the AI evolves as each digital citizen learns from interactions with the environment and other agents [109]. It's a powerful, flexible paradigm for defining the learning of digital citizens, thereby encouraging consistently increasingly value-aligned actions [109].

The actions of the LLMs in the virtual city comprise one aspect of our scenario simulation. Equally important is their ability to interact, to respond and engage with their environment, other AI agents, and external inputs [128]. Zhang et al. point out how advanced machine learning techniques can be employed to develop conversational capabilities, further improving the simulation's reality.

These AI agents, significantly enhance the potential of running multi-agent simulations, including collective decision-making scenarios, cooperation, competition, and conflict [74]. Consequently, this results in a higher understanding of the complex social and interactive scenarios that we might encounter in the real world.

Digital citizens' personification allows for a gravity of interaction and personalization that strengthens the depth and breadth of the simulated city. Equipped with a high degree of autonomy, natural language-processing capabilities, character traits, and unique behaviors, these AI agents form the base of our simulated environment and are necessary to fulfill the objectives delineated for AGI development [21].

Perception – computer vision approach

The perceptual abilities of digital citizens feature prominently in the realization of an immersive and engaging simulation environment. To this effect, the incorporation of computer vision is paramount. Computer vision, a critical field within Robotics and AI, involves the automatic extraction, analysis and understanding of useful information from images or video sequences [111]. Consequently, computer vision can provide digital citizens the ability to perceive and interpret their environment.

Several theories and techniques within computer vision can be leveraged to allow AI agents to distinguish between objects and individuals, identify patterns, interact more naturally or even anticipate possible future states of the environment. One such technique is Deep Learning, a class of machine learning algorithms that has been extremely effective in the realm of Computer Vision [72]. Convolutional Neural Networks, a variant of deep learning, can aid digital citizens in identifying and categorizing the wide range of visual stimuli typically present in a cityscape [71].

Semantic segmentation, an advanced computer vision technique, may be used to enable AI agents to understand the varying components of their surroundings better, by classifying different parts of images into distinct categories [77]. This can make for more sophisticated navigation and interaction within the digital city, setting the stage for more realistic scenarios.

Implementing concepts from Visual Scene Understanding, an area of research studying how observer characteristics, such as familiarity, guide anticipation and action planning [60], could also grant our AI agents the ability to make complex decisions based on the visual information they process.

While integrating computer vision into digital citizens, it is essential to bear in mind the potential pitfalls, such as the generalizability of training to different settings [113] and the problem of 'adversarial examples' where small input modifications can make AI outputs incorrect [110]. But despite these challenges, the blend of AI's perceivable ability with behavioral autonomy afforded by psychology, game and social theories, could significantly enhance the richness, realism, and utility of the simulation, pushing the boundaries of AI development.

Cognition – game theory

The richness of the digital environment within our simulation approach allows one to manipulate numerous variables, thereby observing the corresponding influence over the AI agent's behavior. The interactions and decision-making capacity of our digital citizens arise as focal points of this study [115]. Turing suggests that we can attribute intelligence to an entity by observing its ability to make "reasoned" decisions under changing circumstances and its capacity to interact convincingly.

One of the unique traits exhibited by our digital citizens is their capacity to engage in interactions. These interactions could arise between multiple AI agents (multi-agent interaction) or between AI agents and human users [124]. Wilks explains that interactions with multiple agents can lead us to comprehend how complex social dynamics play out. In the context of our AI-driven virtual city, these interactions provide the primary source of data, providing insights into the behavior of AI agents in various scenarios [97].

Interactions in a virtual environment can vary in numerous ways, from simple exchanges – such as greetings – to more complicated ones involving conflict resolution and cooperative tasks [64]. Jennings et al. emphasize the need for a well-defined protocol guiding such interactions, critical in ensuring the scalability and success of multi-agent systems.

As for decision-making, they are crucial in an autonomous agent's capacity to take independent actions. These decisions could range from simple binary choices to complex resolutions involving trade-offs between conflicting interests or values [109]. Sutton and Barto highlight the role of reinforcement learning in shaping an agent's decision-making process—allowing it to 'learn' from its past actions and outcomes.

The application of game theory stands out as a significant tool that provides valuable perspectives for understanding and ultimately influencing the behavior of LLMs. Game theory, originally derived from economic and mathematical contexts, is a well-established theoretical framework that describes and analyzes decision-making scenarios where the outcomes for each participant depend upon the actions of all [52]. In the AI landscape, game theory's main draw is its power to model strategic interactions between multiple agents [102]. In an AI ecosystem constituting multiple self-learning agents, each AI's decision is inherently interdependent, affected by the decisions of others in the same ecosystem. This presents an inherent multi-agent coordination problem. Game-theoretic strategies can provide solutions to this problem by providing a mathematical formalization and an analytic platform for these dynamics, thereby fostering cooperative behaviors [24]. For example, in networked AIs, game-theoretic tactics like the Nash Equilibrium can help align the AIs towards common objectives, by attaining a stable state wherein no AI can gain by unilaterally deviating from its chosen strategy [30].

Game theory contributes to the field of machine ethics, providing a structured methodology for AI to make ethical decisions when pitted against dilemmas. The incorporation of game theory in the design focuses on actual decision-making processes rather than imbuing AIs with abstract, universal principles. By assigning utilities to each possible outcome, ethical AIs can navigate through complex decision trails, evaluating and selecting the most ethical course of action [48].

The interactions and decisions of our digital citizens within this digital space create a rich source of data for multiple applications, from refining future iterations of these AI models to informing policy decisions in digital technology use [4]. Amodei et al. reiterate the significance of the decision-making models, especially in unpredicted scenarios, to enhance our AI agent's ability to align with human values.

It is essential to observe that these simulated interactions and decision-making scenarios remain subject to ongoing refinement and calibration, ensuring they accurately reflect potential real-world situations [4]. Amodei et al. highlight the need for adopting an iterative approach in refining and perfecting the simulations. This perspective resonates with the developing nature of AI.

Interactions and decision-making within our digital city offer a detailed, real-time understanding of how autonomous agents respond to a variety of scenarios. The evolution of strategies and choices reflects the trends and challenges that might be encountered, thereby advancing our ability to guide AI development towards safety and alignment.

Discussion

Valuable insights and practical applications

The application of an innovative simulation-based approach to study LLMs in a digital city provides critical insights that hold valuable implications for AI and its expansive usage. Such an approach can contribute to the development and refinement of AI by providing the opportunity to observe and correct anomalies or undesired behaviors in a controlled, reusable, and adaptable environment [103].

According to the above literature, the project of an autonomous digital city for AI testing may consist of multiple modules, such as infrastructure (simulation engines) and citizens (personality, perception, and cognition). The infrastructure module provides a controlled environment where AI behaviors can be analyzed and tested in numerous scenarios. A simulation engine utilizes advanced algorithms and computational models to create, interact with, and evolve entities within a realistic virtual environment. The Citizens (AI Agents) module simulates city inhabitants, exhibiting complex behaviors and a high degree of autonomy. The Computer Vision and Perception module grants AI agents the ability to perceive and interpret their environment, improving realistic scenarios. The Cognition and Decision-making module emphasizes the autonomous decision-making capabilities of AI agents, allowing them to interact and adapt to their environment.

By deploying LLMs in a virtual reality environment, we emulate life via an imagined space equipped with digital citizens. Here, potential interactions and scenarios are virtually limitless, encompassing casual daily interactions, complex social situations, and even unexpected, novel circumstances [26]. This immersive, controlled environment allows for simulations that hold significant anthropological, sociological, and psychological value in addition to their technological implications [26].

The generated knowledge extends beyond the informational dimension of AI performance and alignment. These insights can be applied to forecast, understand, and even shape the potential effects of AI implementation in different societal sectors [84]. The potential effects of AI integration into areas such as healthcare, education,

governance, and commerce can be tested, adjusted, refined, and optimized within the digital city environment before actual implementation [84].

The diverse data generated within the digital city provides valuable ground for addressing both overt and subtle biases that may become embedded within AI systems [25]. The analysis of AI agent interactions can aid in identifying and correcting these biases to ensure AI deployments remain equitable, efficient, and reflective of the diverse values of humanity at large [25].

The specific domains where AI can bring transformative change – such as autonomous vehicles, robotics, customer service, and language translation – can also benefit from the data and knowledge generated from our approach. The observed patterns and anomalies in the digital city can inform us about potential obstacles, improvements, misalignments, and benefits that AI might face in these specific domains [47].

The digital city as a research environment also serves as a crucible for AI's ethical and moral considerations. As the digital citizens interact, make decisions, and evolve, researchers can gain rich insights into the ethical boundaries and value alignment challenges associated with AI [21]. The novel application of a simulation-based approach within a digital city offers a vital pathway for the sustainable and beneficial development of AI.

The key point would be the autonomy of different agents within the digital city. That means agents would be able to move through the city as they wish, engage in activities of their own choosing, and interact with other agents with whom they share common goals or interests. Conversation between agents would be just one possible activity. The goal would be to increase the complexity of the autonomous city, thus making it more realistic as it evolves into a playground for AIs. Having multiple AI agents participate in the digital city through their agents could be useful to establish testing and correction mechanisms. While more AIs would interact, trustworthy ones could serve as teachers and controllers of those that need to be tested and further aligned to human uses, needs, and values. An illustration of this could be seen in Fig. 4. By fostering a deeper understanding of AI dynamics in a variety of scenarios, society can better seize the opportunities and manage the risks associated with AI innovations.

Fig. 4 In the dynamic digital city, numerous autonomous agents, represented as distinctive yellow and red icons, serve as simulations of AI. This digital city is constructed with diverse structural elements such as buildings, parks, roadways, and transportation vehicles, reflecting the diverse facets of a real metropolis. The autonomous agents move around and interact within this landscape, demonstrating active engagement and reflecting the breadth of potential AI actions and interactions. These agent movements and interactions are indicative of continuous learning, decision-making, and evolution, which are inherent aspects of AI. This complex, multi-agent system within a digital city serves as a critical testing and alignment ground for AI development, capturing the numerous opportunities and complexities seen in AI testing and alignment

Refinement and alignment of AI Models

Fine-tuning and aligning AI models to match human values is one of the most pressing needs of the high-tech world today [21]. This is particularly important when reshaping AI, which could potentially replicate or even surpass human-like reasoning and cognition, including ethical and moral decision-making [21].

The innovative simulation-based approach facilitates the observation and correction of anomalies and misalignments of AI models. Observations from their interactions within the digital city can guide AI developers to identify areas of improvement, thereby achieving a higher degree of alignment with human values [97].

One important factor in AI alignment is understanding how the models generatively encode knowledge and concepts from the data they are trained on [94]. Simulation-based testing can offer insight into whether the AI model cognitively 'understands' what it has learnt, allowing us to understand its decision-making processes better [94].

Decision making in AI is guided by reinforcement learning, which is naturally built into our simulation approach [109]. While reinforcement learning remains a powerful tool, it is necessary to carefully manage this learning process to avoid reinforcing undesired values or behaviors inadvertently [4]. As Amodei et al. observed, any approach intending to align AI with human values must consider the complexities of human value systems and the potential for unintended consequences.

In our simulation approach, the behavior of digital citizens within the digital city provides a rich data set for analysis. This data can be utilized to train AI models to act in ways that adhere to acceptable norms and societal values [34]. The unique, scenario-based insights

generated within the simulations can further inform the development of robust safety measures and the alignment of AI goals with broadly accepted human values [34].

The wealth of behavioral data and interaction analysis gathered through the simulation approach allows for unprecedented refinement of AI decision-making, leading to safer, controlled, and value-aligned AI systems. These insights can aid AI researchers and policymakers in mitigating the risks and maximizing the benefits of AI integration into society. Also, some of the theories discussed in this study could be used to address specific aspects of alignment, as indicated in Table 1.

Repercussions for the simulation argument

The successful production of an artificial world containing independent, intelligent, and interacting digital entities naturally evokes contemplation on the simulation argument. This hypothesis, originally proposed by philosopher Nick Bostrom, suggests that advanced civilizations could possess the technology to produce realistic, convincing simulations of past eras peopled by conscious digital entities [20]. The development of a sophisticated simulated environment could bear significant implications for such a hypothesis and pose intriguing philosophical questions.

The creation of a rich simulated city resided by autonomous digital citizens provides a tangible example of the feasibility of developing such a system, supporting the theoretical possibilities postulated by the simulation argument. Although our technology is primitive compared to the advanced civilizations in Bostrom's hypothesis, our progress underscores the potential reality that future technological advancements might enable the realization of full-scale, hyper-realistic simulations [99].

In this context, the possibility that we might unknowingly be part of such a simulation ourselves becomes a more conceivable prospect. Our ability to create simulated environments that imbue digital citizens with some form of perceived consciousness could mirror an advanced civilization's ability to do the same on a magnitudinally larger scale. As such, there arises the existential question of whether our reality is, indeed, 'real' or merely a highly sophisticated simulation [31].

Validating or debunking the simulation hypothesis remains a vexing challenge. Current scientific methodologies fail to offer any verifiable means to do so. In fact, one argument suggests that if we are living in a perfect simulation, we may never be able to discern our reality's true nature [20].

Expanding our understanding and capability of creating simulated realities should be accompanied by ethical considerations, philosophical reflections, and a commitment to uncovering verifiable truths about our world, simulated or otherwise. Ongoing technological advancements reinforce the plausibility of the simulation hypothesis, impelling continuous exploration of our existence's nature and purpose in increasingly uncertain landscapes.

Conclusion

The accelerated growth of Generative AI, and notably, LLMs, necessitates innovative approaches to ensure security and alignment of these models with human values. An effective way to achieve this is through the development and implementation of simulation-based methodologies. Through the application of various social sciences and robotics theories, considerations of computational social dynamics, human behaviors, ethics, and perception can be thoroughly examined, thereby providing crucial insights into AI behavior in the context of diverse societal scenarios.

To answer the research question posed in this inquiry (RQ1), the application of theories and approaches derived from multiple disciplines to create a comprehensive, multi-faceted simulation-based testing approach for AI encompasses several areas:

From Sociology, Social Simulation Theory and Situated Action Theory can aid in creating AI simulations

Table 1 Ethical considerations and theories informing autonomous AI testing and alignment

Ethical Considerations in Autonomous AI	Corresponding Theoretical Framework
Ethical Alignment	Social Simulation Theory
Controllability	Theory of Reasoned Action
Predictability	Situated Action Theory
Unpredictability in real-world dynamics	Complex Systems Theory
Autonomy in decision-making	Swarm Intelligence
Multi-layered AI behaviours	Multi-Agent System Theory
Autonomy and Interaction abilities in AI	Perception and Cognition in AI
Security and Safety in AI	Game Theory

reflecting human social behaviors and conditions under which actions occur. Theory of Reasoned Action derived from Social Psychology could be used to simulate human decision-making processes within AI, reflecting how individuals process information and make decisions. From Physics, Complex Systems theory can be used as a backdrop for creating AI testing simulations that bear the components of complexity, dynamism, and interconnectivity. Swarm Intelligence rooted in Computer Science and Biology, could be used to replicate collective behavior of decentralized, self-organized systems within AI. Multi-agent System Theory from Computer Science could be essential in building AI simulations where multiple agents interact with each other, mimicking real-world multi-actor scenarios. Utilizing Big 5 Personality theory from Psychology could assist in creating AI that simulates human personality characteristics, enhancing its level of realism and relatability. Computer Vision, a branch from Computer Science, could form the basis for visual processing and recognition capabilities of AI. Game Theory from Economics can construct simulations where AI tests strategic interactions between rational decision-makers. Lastly, Simulation Engines, another facet of Computer Science, are fundamental in creating the simulations upon which AI operates. All noted theories and approaches are listed in Table 2 with adequate references.

These multidisciplinary theories and approaches can collectively construct a robust, versatile simulation-based testing environment for AI, augmenting its adaptability and authenticity in various contexts.

The creation of a controlled and reproducible environment, through the design of a digital city populated with digital citizens, allows researchers to observe and evaluate their behavior under various conditions. Social simulation and theory of reasoned action provide the foundation of our approach, framing the study of digital citizen behavior within the context of social relations.

In an effort to effectively manage the often complex and unpredictable behaviors of AI technologies, the application of various theories and approaches is crucial. The former aids our understanding of the interactions and dynamics between AI agents, while the latter allows AI to adapt its behavior in reaction to changes in its environment. Both theories contribute significantly to the alignment of AI decision-making with desired outcomes, despite the challenges that may arise in the process.

Our innovative approach also emphasizes the development of automated simulations, allowing the behavior of LLMs to be studied exhaustively within the digital city. The use of autonomous digital citizens and the exploration of their interactions and decision-making on a large scale provide a robust framework for understanding autonomous behavior. Not only do these insights inform the refinement of AI, but they also hold significant potential for broader applications across various sectors of society.

Equally critical in this process is the refinement and alignment of AI models to propagate secure, controlled, and value-aligned AI systems. The digital city simulation offers an environment where AI can evolve and be increasingly value-aligned with each iteration. However, the complexities inherent in translating the theories into practical AI programming and replicating real-world effects within an artificially confined environment present new challenges—challenges to be addressed with iterative refinement and disciplined learning and development.

This study demonstrates the potential of a simulation-based approach in testing and aligning AI with human values. Nevertheless, it also reveals the complexities and challenges intrinsic to this process, emphasizing the importance of ongoing refinements in theory, design, and

Table 2 Theories and approaches identified as the most adequate for simulation-based AI testing and alignment framework

Aspect	Approach	Field	References
Interactions	Social Simulation Theory	Sociology	[35, 44, 79]
	Situated Action Theory	Sociology	[69, 107, 118]|
	Theory of Reasoned Action	Social Psychology	[50, 100]|
	Complex Systems	Physics	[7, 12, 13, 83]
	Swarm Intelligence	Computer science and biology	[19, 66, 68]
	Multi-agent system theory	Computer science	[76, 96, 125]
Agents	Big 5 personality	Psychology	[39, 55, 127]%
Perception	Computer Vision	Computer science	[71, 72, 111]
Cognition	Game Theory	Economy	[52, 102]%
Infrastructure	Simulation Engines	Computer science	[46, 54, 116]

application. As we delve into the era of AI and beyond, the steps we take must be exploratory in nature, continually acknowledging the dynamism of the terrain ahead. Through such constant engagement and innovation, we can aspire to design AI technologies that are not only effective and secure, but also respect and uphold the values of the human society in which they operate. The insights offered in this research illuminate a potential path – though perhaps initially challenging, it holds the promise of a secure and value-aligned future for AI.

Similarities with CERN

The theoretical framework presented in this paper to create a simulation of a digital city for AI testing and alignment shares similarities with CERN [29], the European Organization for Nuclear Research, in terms of its large-scale, multidisciplinary nature. Like CERN, it's a significant scientific effort that requires collaboration from various fields. However, there are also notable differences. While CERN is focused on advancing our understanding of the fundamental laws of nature through experiments in high-energy particle physics, the digital city simulation aims to advance AI by providing a controlled environment where AI agents can interact, learn, and evolve. The purpose of this simulation is to provide insights into the behavior and decision-making processes of AI agents, thereby helping to refine and align AI technologies.

CERN relies on physical infrastructure and real-world experiments, whereas the digital city simulation is entirely virtual, relying on AI models, algorithms, and computational resources. Although both CERN and the digital city simulation involve complex, dynamic systems, the nature of these systems differs. CERN investigates the behavior of subatomic particles, while the digital city simulation investigates the behavior of AI agents within a complex, simulated societal environment. Both CERN and the digital city simulation represent grand scientific efforts aiming to improve our understanding of complex systems, whether they are natural or artificially designed. They utilize state-of-the-art technology and draw on a broad range of scientific disciplines, showcasing the power and value of interdisciplinary collaboration in scientific research. They both represent ambitious, forward-looking projects that have the potential to make significant contributions to their respective fields of study.

Limitations of the study

Despite the innovative approach adopted in this study, its limitations must be acknowledged. Firstly, the utility of a simulation-based approach relies heavily on the accuracy and complexity of the simulation or the modeled digital city itself. There is an inherent challenge in replicating the multifaceted features and unpredictability of the real world within a virtual framework. While efforts have been made to ensure a broad range of scenarios and interactions within this study's digital city, the certainty of covering all potential variables and eventualities remains elusive.

The adoption of social science, robotics, and artificial intelligence theories into the LLMs test framework is a task riddled with complexities. This integration is a novel approach, and it brings forth the challenge of developing AI models in a way that accurately reflects these theories' concepts. The translation of abstract theoretical concepts into practical AI programming can be riddled with difficulties, which only increase with the complexity and unpredictability of real-world factors.

Like all AI models, digital citizens introduced in this study also run the risk of incorporating and even amplifying biases present in their training environments or data. While steps are taken to identify and correct these biases, the task's difficulty and the potential impact of these biases should not be underestimated.

Determination of successful AI alignment is a complex feat. It hinges on defining what constitutes "desirable" behavior and ensuring that the AI models exhibit such behavior consistently in a range of scenarios. Currently, our understanding and definition of successful AI alignment is constrained by known human values and societal norms, introducing a probable limitation of oversight or omission of novel or altered values and norms that can form in future societies.

Future research

Future research can build upon the findings of this study in several ways. Advancements in virtual and augmented reality technologies could substantially enhance the digital city's realism in this study, thereby improving the accuracy and relevance of the simulations.

Exploratory endeavors can focus on refining the process of integrating these varied theories into AI programming. New techniques or algorithms might be developed to facilitate this integration, based on learnings gained from the successes and challenges of the current approach.

Systematic efforts could be carried out to identify, understand, and correct biases in AI models with the help of insights gained from the digital city simulations. These techniques could then be incorporated into an automated bias-detection and correction framework for AI development.

There is a pressing need for a comprehensive, universally accepted definition and understanding of successful AI alignment, which respects the diversity and dynamism of human values across cultures and time. Future research

should also focus on developing continuous monitoring and recalibration tools for AI models, ensuring their behavior remains aligned with these values over time, despite the changes and refinements in societal norms and regulations.

This research provides a fertile ground for advancements in the field of AI development and testing. Its pioneering blend of theories and simulation-based approach offers opportunities for further research and exploration towards the goal of secure, controlled, and value-aligned AI.

Acknowledgements
Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them. Grant Agreement no. 101120763—TANGO.
This paper was realized with the support of the Ministry of Science, Technological Development, and Innovation of the Republic of Serbia, according to the Agreement on the Realization and Financing of Scientific Research.
This paper has been supported by the TWON (project number 101095095), a research project funded by the European Union, under the Horizon Europe framework (HORIZON-CL2-2022-DEMOCRACY-01, topic 07). More details about the project can be found on its official website: https://www.twon-project.eu/.

Authors' contributions
Ljubiša Bojić, Ph.D., conceived the original idea, provided overall guidance and wrote significant portions of the text. Matteo Cinelli, Ph.D., contributed to the design of the theoretical framework and wrote relevant parts of the text. Dubravko Ćulibrk, Ph.D., provided essential insights into the potential applications of AI and contributed to writing the paper. Boris Delibašić, Ph.D., helped refine the concepts and contributed to drafting the manuscript. All authors contributed to finalizing the paper, ensuring rigor in the methodology, and reviewing the final draft before submission.

Funding
This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. It is an independent study conducted by the authors listed, contributing to their respective institutional and academic portfolios.

Availability of data and materials
The data that supports the findings of this study are available upon reasonable request from the corresponding authors. It includes a variety of theoretical and practical resources rooted in multiple disciplines, including sociology, social psychology, computer science, physics, biology, and economics.

Declarations

Competing interests
The authors, Ljubiša Bojić, Matteo Cinelli, Dubravko Ćulibrk, and Boris Delibašić, declare that they have no competing interests. They affirm that the submitted work is original, and it has not been influenced by any personal, financial, or commercial relationships or interests.

Received: 20 April 2024 Accepted: 8 August 2024
Published online: 24 August 2024

References

1. Aher G, Arriaga RI and Kalai AT (2023) Using large language models to simulate multiple humans and replicate human subject studies. arXiv. http://arxiv.org/abs/2208.10264
2. AkshitIreddy (2023) Interactive LLM Powered NPCs. GitHub. https://github.com/AkshitIreddy/Interactive-LLM-Powered-NPCs
3. Altman S (2023) Planning for AGI and beyond. OpenAI Blog. https://openai.com/blog/planning-for-agi-and-beyond
4. Amodei D, Olah C. Steinhardt J, Christiano P, Schulman J, Mané D (2016). Concrete problems in AI safety. *arXiv*. https://doi.org/10.48550/arXiv.1606.06565
5. Armstrong S, Sotala K, Óhéigeartaigh SS (2012) The errors, insights and lessons of famous AI predictions – and what they mean for the future. J Exper Theor Artif Intell 26(3):317–342
6. AutoGPT (2023) A simple digital vector art of an octopus like creature, used as the logo of Auto GPT [Illustration], Retrieved October 30, 2023, from: https://en.wikipedia.org/wiki/Auto-GPT#/media/File:Auto_GPT_Logo.png
7. Axelrod R. The dissemination of culture: a model with local convergence and global polarization. J Confl Resolution. 1997;41(2):203-226.
8. Bail CA (2023) Can Generative AI Improve Social Science?. https://doi.org/10.31235/osf.io/rwtzs
9. Banks J, Carson J, Nelson B, Nicol D (2000) Discrete-Event System Simulation. Prentice Hall, New Jersey
10. Barrett RCA, Poe R, O'Camb JW, Woodruff C, Harrison SM, Dolguikh K, Chuong C, Klassen AD, Zhang R, Joseph RB, Blair MR (2022) Comparing virtual reality, desktop-based 3D, and 2D versions of a category learning experiment. PLoS ONE 17(10):e0275119. https://doi.org/10.1371/journal.pone.0275119
11. Bartneck C, Kulić D, Croft E, Zoghbi S (2015) Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. Int J Soc Robot 1(1):71–81
12. Bar-Yam Y (1997) Dynamics of complex systems. Addison-Wesley, Reading
13. Bar-Yam Y (2003) Complexity of military conflict: multiscale complex systems analysis of littoral warfare. New England Complex Systems Institute. https://necsi.edu/complexity-of-military-conflict
14. Batty M, Torrens P (2001) Modeling complexity: the limits to prediction. (CASA Working Papers 36). Centre for Advanced Spatial Analysis: London, UK.
15. Bertacchini E, Grazzini J, Vallino E (2013) Emergence and Evolution of Property Rights: an Agent Based Perspective. Working Papers 201340, Department of Economics and Statistics Cognetti de Martiis, University of Turin.
16. Blum C, Li X (2008) Swarm intelligence in optimization. In: Blum C, Merkle D (eds) Swarm intelligence. Natural Computing Series. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-74089-6_2
17. Bojic L (2022) Metaverse through the prism of power and addiction: what will happen when the virtual world becomes more attractive than reality? Eur J Futur Res. 10(22). https://doi.org/10.1186/s40309-022-00208-4
18. Bolton, R.N., McColl-Kennedy, J.R., Cheung, L., Gallan, A., Orsingher, C., Witell, L. and Zaki, M. (2018). Customer experience challenges: bringing together digital, physical and social realms, *Journal of Service Management*, *29*(5), 776-808. https://doi.org/10.1108/JOSM-04-2018-0113
19. Bonabeau E, Dorigo M, Theraulaz G (1999) Swarm intelligence: from natural to artificial systems. Oxford University Press, New York
20. Bostrom N (2003) Are you living in a computer simulation? Philos Quart 53(211):243–255
21. Bostrom, N. (2014). Superintelligence: paths, dangers, strategies. Oxford University Press.
22. Bostrom N, Yudkowsky E (2014) The ethics of artificial intelligence. In: Frankish K, Ramsey WM (eds). The Cambridge Handbook of Artificial Intelligence. Cambridge University Press. Pp. 316–334. https://doi.org/10.1017/CBO9781139046855.020
23. Brundage M, Avin S, Wang J, et al (2018) The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv. https://doi.org/10.48550/arXiv.1802.07228
24. Busoniu L, Babuska R, De Schutter B (2008) A comprehensive survey of multi-agent reinforcement learning. IEEE Transact Syst Man Cybernet Part C (Applications and Reviews) 38(2):156–172
25. Caliskan A, Bryson JJ, Narayanan A (2017) Semantics derived automatically from language corpora contain human-like biases. Science 356(6334):183–186
26. Castelfranchi C (2000) Artificial liars: why computers will (necessarily) deceive us and each other. Ethics Inf Technol 2(2):113–119

27. Cave S, Craig C, Dihal K, Dillon S, Montgomery J, Singler B, Taylor L. (2018) Portrayals and perceptions of AI and why they matter. In: Artificial Intelligence Safety and Security. CRC Press. pp. 283–296
28. Cave S, ÓhÉigeartaigh SS, Weller A (2019) Bridging near- and long-term concerns about AI. Nat Mach Intell 1:5–6
29. CERN (2022) The Standard Model. https://home.cern/science/physics/standard-model
30. Chalkiadakis G, Elkind E, Wooldridge M (2011) Computational aspects of cooperative game theory. Morgan & Claypool Publishers, Princeton
31. Chalmers DJ (2010) The character of consciousness. Oxford University Press
32. Chen MX, Firat O, Bapna A, Johnson M, Macherey W, Foster G, … Wu Y (2020) The best of both worlds: Combining recent advances in neural machine translation. *arXiv*. https://doi.org/10.48550/arXiv.1804.09849
33. Chevaleyre Y (2004) Theoretical analysis of the multi-agent patrolling problem. Proceedings of International Conference on Intelligent Agent Technology (IAT 2004), 302–308. https://doi.org/10.1109/IAT.2004.1342959
34. Christano P, Leike J, Brown T, Martic M, Legg S, Amodei D (2017) Deep reinforcement learning from human preferences. In: Advances in Neural Information Processing Systems. pp. 4299–4307.
35. Cioffi-Revilla C (2014) Introduction to Computational Social Science: Principles and Applications. Springer.
36. Clerc M, Kennedy J (2002) The particle swarm-explosion, stability, and convergence in a multidimensional complex space. IEEE Trans Evol Comput 6(1):58–73. https://doi.org/10.1109/4235.985692
37. Convai (2023). Convai. https://www.convai.com/
38. Convex (2023) AI Town: A virtual town where AI characters live, chat and socialize. *Convex.dev*. https://www.convex.dev/ai-town
39. Costa PT, McCrae RR (1992) Normal personality assessment in clinical practice: The NEO Personality Inventory. Psychol Assess 4(1):5
40. Creswell JW (2009) Research design: qualitative, quantitative, and mixed methods approaches. 3rd ed. Sage Publications, Inc.
41. Cruz F, Solis MA, Navarro-Guerrero N (2023) Editorial: Cognitive inspired aspects of robot learning. Front Neurorobot 17:1256788. https://doi.org/10.3389/fnbot.2023.1256788
42. Dafoe A (2018)AI governance: a research agenda. Governance of AI Program, Future of Humanity Institute, University of Oxford
43. Dillion D, Tandon N, Gu Y, Gray K (2023) Can AI language models replace human participants? Trends Cogn Sci 27(7):597–600. https://doi.org/10.1016/j.tics.2023.04.008
44. Edmonds B, Moss S (2005) From KISS to KIDS – an 'anti-simplistic' modelling approach. In: Davidsson P, Logan B, Takadama K (eds) Multi-Agent and Multi-Agent-Based Simulation. Springer, pp 130–144
45. Engel AK, Maye A, Kurthen M, König P (2013) Where's the action? The pragmatic turn in cognitive science. Trends Cogn Sci 17(5):202–209
46. Epic Games (2020) Unreal Engine. Retrieved from https://www.unrealengine.com/
47. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K,…& Dean J (2019) A guide to deep learning in healthcare. Nature Med25(1):24–29.
48. Etzioni, A (1990) The moral dimension: Toward a new economics. Free Press.
49. Everitt BS, Landau S, Leese M, Stahl D. (2011). Cluster analysis. Wiley. https://doi.org/10.1002/9780470977811
50. Fishbein M, Ajzen I (2010) Predicting and changing behavior: the reasoned action approach. Psychology Press.
51. Franklin S, Graesser A (1997) Is it an Agent, or just a Program?: a taxonomy for autonomous agents. In: Proceedings of the 3rd International Workshop on Agent Theories, Architectures, and Languages. Springer-Verlag. pp. 21–35
52. Fudenberg D, Tirole J (1991) Game theory. MIT Press, Cambridge
53. Gartner (2023) Definition of Artificial general intelligence (AGI). https://www.gartner.com/en/information-technology/glossary/artificial-general-intelligence-agi
54. Godot Engine contributors. (2020). Godot Engine. Retrieved from https://godotengine.org/
55. Goldberg LR (1990) An alternative "description of personality": the big-five factor structure. J Pers Soc Psychol 59(6):1216
56. Guo F (2023) GPT agents in game theory experiments. arXiv. http://arxiv.org/abs/2305.05516
57. Hart C (1998) Doing a literature review: releasing the research imagination. SAGE Publications.
58. Heath B, Hill R, Ciarallo F (2019) A survey of agent-based modeling practices (January 1998 to July 2008). J Artif Soc Soc Simul 12(4):9
59. Helbing D (2015) Thinking ahead-essays on big data, digital revolution, and participatory market society. Springer
60. Henderson JM, Hayes TR (2017) Meaning-based guidance of attention in scenes as revealed by meaning maps. Nat Hum Behav 1(10):743–747
61. Holland JH (2006) Studying complex adaptive systems. J Syst Sci Complexity 19(1):1–8. https://doi.org/10.1007/s11424-006-0001-z
62. Hutchins E (1995) Cognition in the Wild. MIT press.
63. Irving G, Askell A (2019) Ai safety needs social scientists. Distill 4(2). https://doi.org/10.23915/distill.00014
64. Jennings NR, Sycara K, Wooldridge M (1998) A roadmap of agent research and development. Auton Agent Multi-Agent Syst 1(1):7–38
65. John OP, Naumann LP, Soto CJ (2008) Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In: John OP, Robins RW, Pervin LA (eds) Handbook of personality: Theory and research. The Guilford Press, pp 114–158
66. Karaboga D, Akay B (2009) A survey: algorithms simulating bee swarm intelligence. Artif Intell Rev 31(1–4):61–85. https://doi.org/10.1007/s10462-009-9127-4
67. Kelly SDT, Suryadevara NK, Mukhopadhyay SC (2013) Towards the implementation of IoT for environmental condition monitoring in homes. IEEE Sens J 13(10):3846–3853
68. Kennedy J, Eberhart R, Shi Y (2001) Swarm intelligence. Morgan Kaufmann
69. Kirsh D (2009) Problem solving and situated cognition. The Cambridge Handbook of Situated Cognition. Cambridge University Press, Cambridge, pp 321–339
70. Kopecky F (2022) Arguments as drivers of issue polarisation in debates among artificial agents. J Artif Soc Soc Simul 25(1):4
71. Krizhevsky A, Sutskever I, Hinton G E (2012) Imagenet classification with deep convolutional neural networks. In: *Proceedings of the 25th International Conference on Neural Information Processing Systems* (pp. 1097–1105).
72. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444
73. Lehman J, Clune J, Misevic D, Adami C, Altenberg L, Beaulieu J, … & Hod LLE (2018) The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities. Artif Life 26(2):274–306.
74. Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J., & Graepel, T. (2017). Multi-agent reinforcement learning in sequential social dilemmas. *In Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems* (pp. 464–473).
75. Leike, J., Martic, M., Krakovna, V., Ortega, P. A., Everitt, T., Lefrancq, A., ... & Legg, S. (2017). AI safety gridworlds. *arXiv*. https://doi.org/10.48550/arXiv.1711.09883
76. Lesser V (1999) Cooperative multiagent systems: A personal view of the state of the art. IEEE Trans Knowl Data Eng 11(1):133–142
77. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431–3440.
78. Lutkevich B (2023). Auto-GPT. TechTarget. https://www.techtarget.com/whatis/definition/Auto-GPT
79. Macy MW, Willer R (2002) From factors to actors: computational sociology and agent-based modeling. Ann Rev Sociol 28:143–166
80. Mataric MJ (1998) Behavior-based robotics as a tool for synthesis of artificial behavior and analysis of natural behavior. Trends Cogn Sci 2(3):82–86
81. Mauhe N, Izquierdo LR, Izquierdo SS (2023) Social simulation models as refuting machines. J Artif Soc Soc Simul 26(2):8. https://doi.org/10.18564/jasss.5076
82. McEwan GF, Groner ML, Fast MD, Gettinby G, Revie CW (2015) Using agent-based modelling to predict the role of wild refugia in the evolution of resistance of sea lice to chemotherapeutants. PLoS ONE 10(10):e0139128. https://doi.org/10.1371/journal.pone.0139128
83. Miller JH, Page SE (2007) Complex adaptive systems: an introduction to computational models of social life. Princeton University Press.

84. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, … Petersen S (2015) Human-level control through deep reinforcement learning. Nature 518(7540):529–533.
85. Olson RS, Hintze A, Dyer FC, Knoester DB, Adami C (2013) Predator confusion is sufficient to evolve swarming behaviour. J Royal Soc Interface 10(85):20130305. https://doi.org/10.1098/rsif.2013.0305
86. Ossowski S (2013) Agreement technologies. Springer Science & Business Media.
87. Ostrom E (2014) A general framework for analyzing sustainability of social-ecological systems. Science 325:419–422
88. Park JS, O'Brien JC, Cai CJ, Morris MR, Liang P, Bernstein MS (2023) Generative Agents: Interactive Simulacra of Human Behavior. arXiv. http://arxiv.org/abs/2304.03442
89. Park JS, Popowski L, Cai CJ, Morris MR, Liang P, Bernstein MS (2022) Social Simulacra: Creating Populated Prototypes for Social Computing Systems. arXiv. http://arxiv.org/abs/2208.04024
90. Premack D, Woodruff G (1978) Does the chimpanzee have a theory of mind? Behavior Brain Sci 1(4):515–526
91. Rabin S (2014) Introduction to Game Development, Second Edition. Charles River Media.
92. Radford, A., Brown, T. B., Sutskever, I., et al. (2019). *Language models are unsupervised multitask learners*. Openai. https://openai.com/blog/better-language-models/.
93. Rafols I (2014) Knowledge integration and diffusion: Measures and mapping of diversity and coherence. *arXiv*. http://arxiv.org/abs/1412.6683
94. Ring M, Orseau L (2011) Delusion, survival, and intelligent agents. In: Schmidhuber J, Thórisson KR, Looks M eds. Artificial General Intelligence. AGI 2011. Lecture Notes in Computer Science, vol 6830. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-22887-2_2
95. Rouse WB, Cannon-Bowers JA, Salas E (1992) The role of mental models in team performance in complex systems. IEEE Trans Syst Man Cybern 22(6):1296–1308
96. Russell S & Norvig P (1995) Artificial Intelligence: A Modern Approach. Pearson.
97. Russell S, Dewey D, Tegmark M (2015) Research priorities for robust and beneficial artificial intelligence. AI Mag 36(4):105–114
98. Sartori G, Orrù G (2023) Language models and psychological sciences. Front Psychol 14:1279317. https://doi.org/10.3389/fpsyg.2023.1279317
99. Schneider S (2008) Future minds: transhumanism, cognitive enhancement and the nature of Persons. Neuroethics Publications. https://repository.upenn.edu/cgi/viewcontent.cgi?article=1037&context=neuroethics_pubs
100. Sheeran P, Webb TL (2016) The intention–behavior gap. Soc Pers Psychol Compass 10(9):503–518
101. Shoham Y (1993) Agent-oriented programming. Artif Intell 60(1):51–92
102. Shoham Y, Leyton-Brown K (2008) Multiagent systems: Algorithmic, game-theoretic, and logical foundations. Cambridge University Press.
103. Shoham Y, Perrault R, Brynjolfsson E, Clark J, Manyika J, Niebles JC, … Etchemendy J (2018) The AI Index 2018 Annual Report. AI Index Steering Committee, Human-Centered AI Initiative, Stanford University, Stanford.
104. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, Van Den Driessche G, … & Hassabis D (2016) Mastering the game of Go with deep neural networks and tree search. Nature 529(7587):484–489.
105. Smith R (2010) The long history of gaming in military training. Simul Gaming 41(1):6–19. https://doi.org/10.1177/1046878109334330
106. Stone P, Veloso M (2000) Multiagent systems: a survey from a machine learning perspective. Auton Robot 8(3):345–383
107. Suchman LA (1987) Plans and situated actions: the problem of human-machine communication. Cambridge University Press.
108. Sun R (2005) Cognition and multi-agent interaction: From cognitive modeling to social simulation. Cambridge University Press.
109. Sutton RS, Barto AG (2018) Reinforcement learning: An introduction. MIT press.
110. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R (2014) Intriguing properties of neural networks. arXiv. http://arxiv.org/abs/1312.6199
111. Szeliski R (2010) Computer Vision: Algorithms and Applications. Springer Science & Business Media.
112. Taylor SJ, Bogdan R, DeVault M (2015) Introduction to Qualitative Research Methods: A Guidebook and Resource, 4th edn. John Wiley & Sons, London
113. Torralba A, Efros AA (2011) Unbiased look at dataset bias. In CVPR 2011. https://people.csail.mit.edu/torralba/publications/datasets_cvpr11.pdf
114. Troitzsch KG, Mueller U, Gilbert GN, Doran JE (eds) (1996) Social science microsimulation. Springer-Verlag, Berlin
115. Turing AM (1950) Computing machinery and intelligence. Mind 59(236):433–460
116. Unity Technologies. (2020). Unity. Retrieved from https://unity.com/
117. Van Pinxteren MME, Pluymaekers M, Lemmink JGAM (2020) Human-like communication in conversational agents: A literature review and research agenda. J Serv Manag 31(2):203–225. https://doi.org/10.1108/JOSM-06-2019-0175
118. Varela FJ, Rosch E, Thompson E (1992) The Embodied Mind: Cognitive Science and Human Experience. MIT press, Massachusetts
119. Véliz C (2020) Privacy is power: why and how you should take back control of your data. Transworld.
120. Vora J, Nair S, Gramopadhye AK, Duchowski AT, Melloy BJ, Kanki B (2002) Using virtual reality technology for aircraft visual inspection training: Presence and comparison studies. Appl Ergon 33(6):559–570. https://doi.org/10.1016/S0003-6870(02)00039-X
121. Wang L, Ma C, Feng X, Zhang Z, Yang H, Zhang J, Chen Z, Tang J, Chen X, Lin Y, Zhao WX, Wei Z, Wen J-R (2023) A survey on large language model based autonomous agents. arXiv. http://arxiv.org/abs/2308.11432
122. Weiss G (2000) Multiagent systems: a modern approach to distributed artificial intelligence. MIT press.
123. Whittlestone J, Nyrup R, Alexandrova A, et al (2019) Ethical and societal implications of algorithms, data, and artificial intelligence: A roadmap for research. Nuffield Foundation.
124. Wilks Y (2010) Close engagements with artificial companions: Key social, psychological, ethical and design issues. John Benjamins Publishing.
125. Wooldridge M (2009) An introduction to multiagent systems. Wiley
126. Yan X, Zeng Z, He K, Hong H (2023) Multi-robot cooperative autonomous exploration via task allocation in terrestrial environments. Front Neurorobot 17:1179033. https://doi.org/10.3389/fnbot.2023.1179033
127. Yarkoni T (2010) The abbreviation of personality, or how to measure 200 personality scales with 200 items. J Res Pers 44(3):180–192
128. Zhang S, Dinan E, Urbanek J, Szlam A, Kiela D, Weston J (2018) Personalizing Dialogue Agents: I have a dog, do you have pets too?. In: *Proceedings of the 57th annual meeting of the association for computational linguistics*, 2200–2210.

Publisher's Note