The Interplay of Social and Robotics Theories in AGI Alignment: Navigating the Digital City Through Simulation-based Multi-Agent Systems

Ljubiša Bojić[1,2,*], Vladimir Đapić[1]

[1]The Institute for Artificial Intelligence Research and Development of Serbia, Fruskogorska 1, 21000 Novi Sad, Serbia

[2]Digital Society Lab, Institute for Philosophy and Social Theory, University of Belgrade, Kraljice Natalije 45, 11000 Belgrade, Serbia

Abstract

This study delves into the task of aligning Artificial General Intelligence (AGI) and Large Language Models (LLMs) to societal and ethical norms by using theoretical frameworks derived from social science and robotics. The expansive adoption of AGI technologies magnifies the importance of aligning AGI with human values and ethical boundaries. This paper presents an innovative simulation-based approach, engaging autonomous 'digital citizens' within a multi-agent system simulation in a virtual city environment. The virtual city serves as a platform to examine systematic interactions and decision-making, leveraging various theories, notably, Social Simulation Theory, Theory of Reasoned Action, Multi-Agent System Theory, and Situated Action Theory. The aim of establishing this digital landscape is to create a fluid platform that enables our AI agents to engage in interactions and enact independent decisions, thereby recreating life-like situations. The LLMs, embodying the personas in this digital city, operate as the leading agents demonstrating substantial levels of autonomy. Despite the promising advantages of this approach, limitations primarily lie in the unpredictability of real-world social structures. This work aims to promote a deeper understanding of AGI dynamics and contribute to its future development, prioritizing the integration of diverse societal perspectives in the process.

Keywords

Artificial General Intelligence, Large Language Models, Social Theories, Robotics Theories, Simulation-Based Approach

1. Introduction

The increasingly pervasive role of AI, especially natural language processing (NLP), signifies a new frontier of technological development. AI-driven applications like Generative Pretrained Transformers (GPT) pioneer transformations across society [1]. As reliance on such AI systems rises, so does the challenge of adapting these models to human values, prompting deeper research and development.

Despite rapid advancements, achieving full controllability and value alignment with AI is a notable hurdle, especially with large-scale neural networks [2]. The rise of powerful AI models like GPT further amplifies concerns about their ethical alignment, controllability, and unpredictability [3]. This pressure intensifies the exploration of better testing and mitigation strategies [1].

Large Language Models (LLMs) are artificial intelligence (AI) programs capable of language generation, translation, question answering, summarization, and code generation [4]. Unlike traditional AI models, which are trained on specific datasets and for particular tasks, LLMs are trained on diverse internet text content. They have demonstrated performance in a wide range of tasks and languages without any task-specific training [4], a capability that resonates with the concept of artificial general intelligence (AGI).

AGI refers to a type of AI with cognitive capabilities that can successfully understand, learn, and implement intellectual tasks equivalent to those of a human being [5]. Contrary to traditional AI that is limited to expert-level competence in specific tasks, AGI can understand, learn, and adapt to any intellectual task that can be performed by humans [5] The universality of this ability in AGI is often considered as both beneficial and dangerous. While it promises extensive progress and efficiency in virtually all fields of life, it also imposes significant risks related to misuse and unintended consequences.

Aside from text generation, sophisticated Large Language Models (LLMs) also exhibit the capacity to simulate understanding of inquiries and perform complex cognitive tasks [6]. Among numerous 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 continues the development of Artificial General Intelligence publicly while devising strategies that ensure AGI's safety and alignment with human values [7]. On the other hand, LLMs can be given various degrees of autonomy while creating multiple agents with different

BISEC'23: 14th International Conference on Business Information Security, November 24, 2023, Niš, Serbia

[*] Corresponding author.

ljubisa.bojic@ivi.ac.rs (L. Bojić); vladimir.djapic@ivi.ac.rs (V. Đapić)

0000-0002-5371-7975 (L. Bojić); 0000-0002-8661-0269 (V. Đapić)

prompts capable of interacting with each other [8].

AI Alignment represents the proposition of ensuring that the behavior of AGI system is congruent with human intentions and values. As Bostrom [9] argues in his book "Superintelligence," it is incredibly challenging to specify what is meant by human values in a way that an AI can understand. The alignment of AGI is considered crucial due to multiple reasons. The development of AGI might lead to an intelligence explosion where AGI surpasses human intelligence. If such a situation arises, it is important to ensure that AGI is beneficially aligned and promotes the interests of humanity [9]. Moreover, poorly aligned AI could result in negative ramifications if it can impact significant resources or make autonomous decisions. Hence, dedicated research is needed to ensure that AGI development is carried out responsibly and with necessary precautions.

AI and AGI advancements come with benefits, complexities, risks, and ethical challenges. With traditional risk management methodologies proving inadequate, there's a shift towards exploring more multi-layered methodologies [10]. The unpredictability of AI and AGI systems poses risks, underpinning the necessity of embedding human values and ethics into AI systems [3]. Transparent, accountable AI systems developed with public involvement are advocated by scholars like Véliz [11] and Whittlestone et al. [12], leading to the democratization of technology. The unification of social science theories and technology offers a promising path for developing socially-responsible AI and AGI [13, 14].

This paper delves into the potentials and challenges of AI and social robotics theory convergence for aligning AGI and LLMs. It explores theories and their application in AI alignment, demonstrating their relevance in simulation-based approaches within a digital city environment. The paper concludes with reflections on limitations and directions for future research, essential for ensuring AGI technologies are effective, secure, and uphold societal values

2. Theoretical framework

Exploring social science and robotics theories can provide critical insights for testing and aligning Large Language Models (LLMs) and artificial general intelligence (AGI). The complexity of LLMs and AGIs demand a stringent, theory-based approach [13]. Social science theories aid in understanding and predicting AI behavior [15], while robotics theories provide essential insights on machine ethics and multi-agent system operation for AGI design and refining [16].

Incorporating social science theories in AI research grants a lens for understanding AI alignment and behavior. The relevance of Social Simulation Theory and Theory of Reasoned Action is considerable. Stemming from the Computational Social Science spectrum, Social Simulation Theory leverages computational methods for simulating and analyzing social dynamics, thus driving tests for large language models and better aligning AI behavior to social norms [17, 18]. However, representing the unpredictable nature of real-world social systems in abstract computational models is a significant challenge, limiting the theory's accuracy and applicability [19].

The Theory of Reasoned Action, from social psychology, asserts that intentions drive behavior, influenced by attitudes towards the behavior, norms, and perceived control [20]. While originally for understanding human behavior, it can guide AI behavior modeling, influencing AI intentions via programmed norms and attitudes, and helping align AI actions with societal values [21]. However, the challenge lies in replicating the complex nature of human emotions and irrational behavior in AI, emphasizing the need for a multifaceted AI alignment approach.

Asimov's Laws of Robotics and The Uncanny Valley Hypothesis offer insights for AI security, concerning human-AI interactions [22]. Asimov's Laws provide ethical guidelines enhancing AI system's controllability and ethical behavior. Yet, ambiguity in AI behavior complicates adherence to these laws [23].

The Uncanny Valley Hypothesis highlights the comfort of users with human-like AI, stressing careful design to ensure secure AI usage [24]. Despite the theory's cultural subjectivity, considering such perceptions augments holistic AI system design, balancing advancement with ethical responsibility and security. Multi-Agent System Theory offers valuable insights for developing autonomous systems and testing LLMs and AGI. Multi-agent systems of AI agents, each with unique attributes and decisions in a simulated digital city, can reveal emergent behavior and systemic strengths or weak points. Challenges, though, include agent synchronization, conflict resolution, and handling competition [25]. Despite these, the theory provides crucial support for AI testing in simulated environments. Situated Action Theory encourages adaptive, situation-driven behavior, enhancing AI responses to digital environments. This theory implies AI models should adapt dynamically to changes rather than sticking to prescribed actions. This approach equips AI to navigate unpredictability inherent in large networks.

However, translating these concepts into AI programming proves challenging due to reality's multidimensional and ambiguous nature. Designing adaptive behavior based on Situated Action Theory helps decipher cognitive functions in simulated environments, paving the way for advanced, reliable AI systems.

Next, we examine the practical implementation of these theories for AGI, focusing on developing a digital city. Subsequent section will reflect on the simulation's results, offering insights for alignment of AI models.

3. Towards simulation of a digital city

A simulation-based methodology enhances the reliability, efficacy, and safety of Large Language Models (LLMs) in AGI development [26]. The authors note that simulations provide controlled settings for testing AI behaviors under various scenarios. This digital city simulation, inspired by McEwan et al. [27], effectively mimics real-life complex interactions in a controlled setting. As such, these tested procedures have become instrumental in AGI development.

In this research, a virtual reality framework adds a potent and immersive dimension to simulation studies, a paradigm gaining wider acceptance [28]. Enhanced with AI, this approach offers opportunities for in-depth analysis of AI interactions in realistic scenarios [29].

By incorporating virtual reality, we tap into a broader context for AI implementation. Lending support to Bolton et al. [30], the creation of a 'digital twin' or 'mirror world' facilitates dynamic AI learning. It triples as a platform for appreciating AI behaviors, an arena for future social sciences research, and a toolkit for understanding social dynamics [31].

A simulation-based approach as noted by Bostrom & Yudkowsky [32], enhances the evaluation of AI, especially LLMs behavior. This methodology, bolstered by a virtual reality dimension, holds potential to remarkable breakthroughs in AGI understanding and enhancement.

Automated simulations for LLMs form the cornerstone of our approach, offering reproducible, scalable, and complex interactive environments [33]. Our digital city employs a multi-agent-based simulation framework, modeling a population of autonomous AI agents or 'digital citizens' [34]. Heath et al. [35] affirm the effectiveness of such agent-based models in understanding complex environments.

The development of this digital realm involves iterative creation of autonomous agents operating within defined parameter spaces [36]. Their autonomy determines their dynamics within the city [37]. A meticulously designed environment, where the AI agents function, necessitates a thorough attention to interactions, constraints, and choices [38]. Continuity in learning behavior and refinement of AI agents are ensured by a reinforcement learning approach, as proposed by Leike et al. [2]. The creation of these simulations significantly influences the lockdown approach's effectiveness in providing real-life scenario-based insights for AGI.

Describing the digital citizens, Bartneck et al. [39] underscore their importance in our simulation strategy. Acting as AI actors, these agents vary in personality, norms, and behaviors, enriching the simulation's scenarios and insights. Autonomy, or the capacity to act independently, is critical for AI agents' value and effectiveness [40].

Various learning models, such as reinforcement learning, are utilized for shaping digital citizens [41]. Interaction and responsiveness to their environment, other AI agents, and external inputs is paramount [42]. Personified digital citizens, complete with autonomy, natural language-processing capabilities, character traits, and unique behaviors, significantly enhance multi-agent simulations [43]. Such enhancement underpins our objectives for AGI development [9].

Our digital environment's richness allows observation and manipulation of variables influencing AI behavior, with significant emphasis on interactions and decision-making of digital citizens [44]. Interactions and decisions form the crux of our simulation, driving insights into AI behavior under various scenarios.

Interactions can range from simple exchanges to conflict resolutions and cooperative tasks [45]. Decision-making forms a crucial part of an autonomous agent's function, stretching from simple choices to complex trade-offs [41]. These interactions and decisions provide data useful in refining AI models and informing digital technology policies [46]. Our simulation-based approach provides invaluable insights for AGI and influences its use [47]. The immersive environment offers simulations of significant clinical, social, and psychological interest [48]. These understandings, extending beyond AGI performance, help anticipate and shape AGI's potential societal impact [49].

Data from the digital city facilitates bias addressing in AGI systems [50]. Areas like autonomous vehicles, robotics, customer service, and translation would gain from information acquired in the digital city environment [51]. The virtual city also underlines the ethical considerations and value alignment issues concerning AGI [9]. The use of a simulation approach in a digital city enriches understanding of AGI dynamics, helping society harness AGI innovations responsibly.

Aligning AI models with human values is critical, especially in AGI, which has the potential to mimic human-like reasoning, including ethical decision-making [9]. Observations from interactions within our simulated approach assist in identifying and rectifying AGI's anomalies and misalignments.

Understanding how models encode knowledge is crucial for AI alignment [52]. Our simulation-based testing offers insights into AI's cognitive understanding, giving a better overview of its decision-making processes. Decision-making in AGI leverages reinforcement learning, but it requires careful management to avoid endorsing undesired behaviors [46].

The behaviors and interactions of digital citizens within our simulation offer rich data for AGI refinement [53]. This scenario-based data aids in developing safety measures, aligning AGI with human values, and mitigating the risks of AI integration into society. Consequently, this enables the creation of safer, controlled, and value-aligned AI systems.

4. Conclusion

AI growth necessitates innovative security solutions and alignment with human values. Through contriving a digital city with digital citizens, various societal interactions can be explored to gain insights into AI behavior. Key theories guiding our approach include social simulation and theory of reasoned action for studying AI behavior in social contexts. Robotics theories illuminate ethical considerations, informed by Asimov's Laws of Robotics and the Uncanny Valley Hypothesis.

The application of Multi-Agent System Theory and Situated Action Theory helps manage AI behaviors, guiding interactions, and environment-response adaptations. This accentuates AI alignment with desired outcomes despite potential challenges. Our approach highlights automated simulations for exhaustive study of AI behavior. Autonomous citizens' interactions provide rich data for understanding autonomy, crucial for AGI refinement and broader societal applications. Simulations also help design value-aligned AGIs. However, challenges exist with theory application to AI programming and replicating real-world effects. Nevertheless, simulation-based approaches show promise for aligning AI with human values, despite complexities.

Our approach also has limitations, primarily the difficulty in replicating complexities of real societies within a digital space. Translating theoretical concepts into AI programming presents additional challenges. Biases in AI models can be perpetuated from training environments, and defining "desirable" behavior for AI alignment proves complex.

Future research can enhance simulation realism using advanced VR and AR technology. Focus should also be on refining theory integration into AI programming and developing automated bias correction frameworks. There's also the need to build definitions of AI alignment that respect the dynamism of values across cultures. This research is a starting point for harnessing theories and simulation-based approaches towards value-aligned AGI.

Acknowledgment

This paper monograph was realised with the support of the Ministry of Science, Technological Development and Innovation of the Republic of Serbia, according to the Agreement on the realisation and financing of scientific research.

References

[1] M. X. Chen, O. Firat, A. Bapna, M. Johnson, W. Macherey, G. Foster, L. Jones, N. Parmar, M. Schuster, Z. Chen, et al., The best of both worlds: Combining recent advances in neural machine translation, arXiv preprint arXiv:1804.09849 (2018).

[2] J. Leike, M. Martic, V. Krakovna, P. A. Ortega, T. Everitt, A. Lefrancq, L. Orseau, S. Legg, Ai safety gridworlds, arXiv preprint arXiv:1711.09883 (2017).

[3] G. Irving, A. Askell, Ai safety needs social scientists, Distill 4 (2019) e14.

[4] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.

[5] Gartner, Definition of Artificial general intelligence (AGI), https://www.gartner.com/en/information-technology/glossary/artificial-general-intelligence-agi, 2023.

[6] G. Sartori, G. Orrù, Language models and psychological sciences, Frontiers in Psychology 14 (2023).

[7] S. Altman, Planning for agi and beyond, OpenAI Blog, February (2023).

[8] B. Lutkevich, Auto-GPT, https://www.techtarget.com/whatis/definition/Auto-GPT, 2023.

[9] T. Mulgan, Superintelligence: Paths, dangers, strategies, 2016.

[10] S. J. Russell, P. Norvig, Artificial intelligence a modern approach, London, 2010.

[11] B. Rumbold, Privacy is power: Why and how you should take back control of your data, written by carissa véliz, Journal of Moral Philosophy 20 (2023) 585–587.

[12] J. Whittlestone, R. Nyrup, A. Alexandrova, K. Dihal, S. Cave, Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research, London: Nuffield Foundation (2019).

[13] I. Rafols, Knowledge integration and diffusion: Measures and mapping of diversity and coherence, in: Measuring scholarly impact: Methods and practice, Springer, 2014, pp. 169–190.

[14] S. Cave, S. S. ÓhÉigeartaigh, Bridging near-and long-term concerns about ai, Nature Machine Intelligence 1 (2019) 5–6.

[15] E. Ostrom, A general framework for analyzing sustainability of social-ecological systems, Science 325 (2009) 419–422.

[16] P. Lin, K. Abney, G. A. Bekey, Robot ethics: the ethical and social implications of robotics, MIT press, 2014.

[17] C. Cioffi-Revilla, Introduction to computational social science, Springer, 2014.

[18] M. W. Macy, R. Willer, From factors to actors: Computational sociology and agent-based modeling, Annual review of sociology 28 (2002) 143–166.

[19] B. Edmonds, S. Moss, From kiss to kids–an 'anti-simplistic'modelling approach, in: International workshop on multi-agent systems and agent-based simulation, Springer, 2004, pp. 130–144.

[20] M. Fishbein, I. Ajzen, Predicting and changing behavior: The reasoned action approach, Taylor & Francis, 2011.

[21] P. Sheeran, T. L. Webb, The intention–behavior gap, Social and personality psychology compass 10 (2016) 503–518.

[22] I. Asimov, I, Robot., New York: Gnome Press., 1950.

[23] J. J. Bryson, Robots should be slaves, Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues 8 (2010) 63–74.

[24] M. Mori, The uncanny valley, Energy 7 (1970) 33–35.

[25] G. Weiss, Multiagent systems: a modern approach to distributed artificial intelligence, MIT press, 1999.

[26] S. D. T. Kelly, N. K. Suryadevara, S. C. Mukhopadhyay, Towards the implementation of iot for environmental condition monitoring in homes, IEEE sensors journal 13 (2013) 3846–3853.

[27] G. F. McEwan, M. L. Groner, M. D. Fast, G. Gettinby, C. W. Revie, Using agent-based modelling to predict the role of wild refugia in the evolution of resistance of sea lice to chemotherapeutants, PLoS One 10 (2015) e0139128.

[28] R. C. A. Barrett, R. Poe, J. W. O'Camb, C. Woodruff, S. M. Harrison, K. Dolguikh, C. Chuong, A. D. Klassen, R. Zhang, R. B. Joseph, et al., Comparing virtual reality, desktop-based 3d, and 2d versions of a category learning experiment, Plos one 17 (2022) e0275119.

[29] J. Vora, S. Nair, A. K. Gramopadhye, A. T. Duchowski, B. J. Melloy, B. Kanki, Using virtual reality technology for aircraft visual inspection training: presence and comparison studies, Applied ergonomics 33 (2002) 559–570.

[30] R. N. Bolton, J. R. McColl-Kennedy, L. Cheung, A. Gallan, C. Orsingher, L. Witell, M. Zaki, Customer experience challenges: bringing together digital, physical and social realms, Journal of service management 29 (2018) 776–808.

[31] L. Bojic, Metaverse through the prism of power and addiction: what will happen when the virtual world becomes more attractive than reality?, European Journal of Futures Research 10 (2022) 1–24.

[32] N. Bostrom, E. Yudkowsky, The ethics of artificial intelligence, in: Artificial intelligence safety and security, Chapman and Hall/CRC, 2018, pp. 57–69.

[33] J. Banks, Discrete event system simulation, Pearson Education India, 2005.

[34] E. E. Bertacchini, G. Jakob, E. Vallino, et al., Emergence and evolution of property rights. an agent-based perspective, WORKING PAPER SERIES 40 (2013).

[35] B. Heath, R. Hill, F. Ciarallo, A survey of agent-based modeling practices (january 1998 to july 2008), Journal of Artificial Societies and Social Simulation 12 (2009) 9.

[36] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al., Mastering the game of go with deep neural networks and tree search, nature 529 (2016) 484–489.

[37] J. Lehman, J. Clune, D. Misevic, C. Adami, L. Altenberg, J. Beaulieu, P. J. Bentley, S. Bernard, G. Beslon, D. M. Bryson, et al., The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities, Artificial life 26 (2020) 274–306.

[38] R. S. Olson, A. Hintze, F. C. Dyer, D. B. Knoester, C. Adami, Predator confusion is sufficient to evolve swarming behaviour, Journal of The Royal Society Interface 10 (2013) 20130305.

[39] C. Bartneck, D. Kulić, E. Croft, S. Zoghbi, Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots, International journal of social robotics 1 (2009) 71–81.

[40] P. Stone, M. Veloso, Multiagent systems: A survey from a machine learning perspective, Autonomous Robots 8 (2000) 345–383.

[41] R. S. Sutton, A. G. Barto, Reinforcement learning: An introduction, MIT press, 2018.

[42] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, J. Weston, Personalizing dialogue agents: I have a dog, do you have pets too?, arXiv preprint arXiv:1801.07243 (2018).

[43] J. Z. Leibo, V. Zambaldi, M. Lanctot, J. Marecki, T. Graepel, Multi-agent reinforcement learning in sequential social dilemmas, arXiv preprint arXiv:1702.03037 (2017).

[44] A. M. Turing, Computing machinery and intelligence, mind 59 (1950) 433–460.

[45] N. R. Jennings, K. Sycara, M. Wooldridge, A roadmap of agent research and development, Autonomous agents and multi-agent systems 1 (1998) 7–38.

[46] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, D. Mané, Concrete problems in ai safety, arXiv preprint arXiv:1606.06565 (2016).

[47] Y. Shoham, R. Perrault, E. Brynjolfsson, J. Clark,

J. Manyika, J. C. Niebles, J. T. Etchemendy, et al., The ai index 2018 annual report – ai index steering committee, human-centered ai initiative, 2018.

[48] C. Castelfranchi, Artificial liars: Why computers will (necessarily) deceive us and each other, Ethics and Information Technology 2 (2000) 113–119.

[49] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al., Human-level control through deep reinforcement learning, nature 518 (2015) 529–533.

[50] A. Caliskan, J. J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases, Science 356 (2017) 183–186.

[51] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, J. Dean, A guide to deep learning in healthcare, Nature medicine 25 (2019) 24–29.

[52] M. Ring, L. Orseau, Delusion, survival, and intelligent agents, in: Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3-6, 2011. Proceedings 4, Springer, 2011, pp. 11–20.

[53] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, D. Amodei, Deep reinforcement learning from human preferences, Advances in neural information processing systems 30 (2017).