When can you use LLMs for ABM?

Bruce Edmonds[1,2,3] [0000-0002-3903-2507]

[1] Centre for Policy Modelling, UK.
[2] UiT The Arctic University of Norway, Tromsø, NO.
[3] Umeå University, SE.
bruce@edmonds.name

Abstract. The paper points out that AI systems (and LLMs in particular) have been and can be developed to meet at least 5 different goals: fool humans, do particular tasks as well as humans, simulate human behaviour, do useful work for humans and to make decisions as an independent entity. Next a brainstorm of some possible uses for LLMs with ABMs are discussed, looking at which kind of AI system might be suitable for each. Many of these are rejected as not being a reliable use. Some others are a possible good use, but not currently explored. The match of AI to ABM purpose is then summarised in a table, along with the two main conclusions: (a) one can not assume that an AI entity that is good at one AI goal will be suitable for your ABM purpose and (b) that one needs to check that it behaves in a reliable manner (i.e. that it achieves that AI goal) even if it does. Neither the sheer plausibility of LLM output, nor clever training, prompting or calibration of LLMs are sufficient, except for illustration or informal exploration.

Keywords: Large Language Model, LLM, Artificial Intelligence, AI, Agent-Based Modelling, ABM, social simulation.

1 Introduction

The use of AI machine-learning algorithms – in particular Large Language Models (LLM) – is rapidly expanding to a bewildering variety of tasks, albeit with different levels of success. This includes their use for various aspects of Agent-Based Modelling (ABM). Already there have been a host of such applications as well as some papers warning of the possible pitfalls (e.g. [27]). This paper seeks to start a discussion as to the reliable and valid uses of these in ABM and what is required to make this work. I will concentrate on LLMs in particular as those are prominent, but many of the arguments are also relevant to other complex machine-learning approaches.

It starts with the AI viewpoint, looking at the different purposes that AI algorithms have been developed for and discusses the current state-of-the art in terms of achieved levels of success for each of these. It then goes on to look at the ABM viewpoint looking at some different uses for AI entities in ABM and discusses their viability as well as what conditions might be necessary for these to be achieved. It then summarises the

connections between achieved AI purposes and their possible use within ABM. It concludes with some cautionary conclusions concerning the situation in the current state-of-the-art and suggests some principles for ensuring their justifiable use.

2 The AI viewpoint

Here I look at the various purposes for which an AI system might be built and assess the potential of the current state-of-the-art in these regards. Each of these requires a different approach in terms of particular training, use and assessment. Success at one of these does not imply success at another, however plausible this may seem.

2.1 To fool humans into thinking AIs are human

Alan Turing set a test for machine intelligence – what is now known as the "Turing Test" [**Error! Reference source not found.**]. The argument went that if a machine could pass this test – fooling a human in a conversation of passed notes that it was human – then this showed that it was intelligent. This was not meant to be a goal for AI as a whole, but rather a way of by-passing philosophical quibbles about the nature of intelligence with a practical test. This goal has been partially achieved, depending on the context and nature of the interaction involved. In the context of a game, where the environment is not as complex as in real life and interaction tends to be restricted and simpler, humans correctly guessed as to the identity of an AI bot 60% of the time (only a bit better than by chance) and other humans 70% of the time [**Error! Reference source not found.**]. Similarly, when restricted to a 5-minute simultaneous conversation (participants had a conversation with a human and an LLM at the same time – within the same 5-minute timeframe) ChatGPT4.5 succeed in fooling humans 73% of the time [**Error! Reference source not found.**]. These cannot be taken as AIs being able to substitute for humans in longer-time or socially embedded situations (Such as the "Long Term Turing Test" [4]) as these limit the relevance of culture and real-world knowledge.

There are, already at least two situations where AIs adopt this role in practice: (a) in spreading (mis)information on social media [6,7] and (b) to get content scraper bots past CATCHA and other defences put up by websites to prevent them [8]. In both of these the idea is primarily to fool another algorithm, not humans. It is, of course, possible that AIs will pass an unrestricted form of the Turing Test in the future, but this might necessitate a period of acculturation or training within human society [5]. However, this does point to a very obvious feature of LLMs – their surface plausibility – that they can interact in human-like ways.

2.2 To do particular tasks as well as humans

Machines have surpassed humans in some cognitive abilities for a long time now, for example that of calculation. Although Charles Babbage designed the first mechanical calculator (his "different engine") and made a prototype, the first working and available

calculating machine was Swedish inventor Per Georg Scheutz's "third difference engine" which was finished in 1853 and later sold to the Dudley Observatory in New York State [9]. In recent decades electronic computers have surpassed most humans in performing many, particular, tasks, e.g. achieving grand-master level at various games such as chess, shogi and go [10]. Many benchmarks concerning tasks humans might have previously done have been targeted with high success, but not all [11].

However, the opaque complexity of LLMs mean that such performance can have unexpected outcomes or side-effects which might be labelled as "biases" [20] or "hallucinations" [23]. Here the problem is not maleficence on the part of the LLMs but simply their complexity. Put baldly, we do not understand them – there is progress at understanding how a trained LLM works (e.g. [24]) but this is only a partial collection of hypotheses, visualisations and models. Of course, there are double standards operating here as humans are often biased and "hallucinate" facts, but the point is that we have a broad, if informal, understanding of the range of human responses and have adapted our expectations and social procedures for dealing with them (e.g. by "prompt engineering" them ☺ [22]) whilst we are still adjusting to working with AI entities. One can ask An LLM to justify or explain its thinking but, as with a human, one cannot assume that their explanation actually matches their own processes [25] – neither human nor AI have direct and detailed access to how they think, but only to an interpretation of this.

2.3 To simulate human behaviour

LLMs certainly capture some aspects of human behaviour, namely writing natural language text. This is not surprising since that is what they are trained to do – bootstrapping off all the text on the internet, they learn which sequences of text are appropriate. This does not mean they do not learn other things from this process, such as styles or simple patterns of reasoning, since they infer higher-order patterns from their large amounts of training data – they are not "digital parrots". The text that LLMs produce are certainly much more human-like than previous machine learning approaches, in particular they seem to get human linguistic context. This advance holds out the prospect that LLMs might be used to accurately simulate human behaviour, given the right training and prompts.

However, their plausibility is not sufficient as a simulation of any particular human behaviour. To take a trivial example, they do not make grammatical errors and typos. To look at a more serious example, there is some work assessing how well LLMs mimic human behaviour when answering survey questions. This is a good test case as one can independently check the success of the LLMs by comparing its output with those produced by humans – one can get beyond mere plausibility and do a rigorous test. The assessment of the current state of the art is that the LLM responses are "approximate at best" to those of humans [**Error! Reference source not found.**], that they "cannot yet correct for sampling or nonresponse bias challenges" [**Error! Reference source not found.**], and that "commercial LLMs generally fail to reflect human-like behavior" [**Error! Reference source not found.**]. Furthermore, they "they are sensitive to perturbations that do not elicit significant changes in humans" [**Error! Reference source**

not found.] and "models consistently appear to better represent subgroups whose aggregate statistics are closest to uniform" [**Error! Reference source not found.**]. Thus, even in this restricted arena, where the LLMs are responding to a task closely associated with their training, they cannot be said to satisfactorily reproduce human behaviour.

The underlying problem here is not that a machine learning algorithm could not learn some relevant human behaviour, but that LLMs are not developed for this purpose (apart from, possibly, writing text that might be found on the internet). They are calibrated to make their responses more suited to how humans work (using feedback from human trainers), but this does not rigorously ensure they reproduce any particular aspect of human behaviour – calibration is not validation. One can imagine a future where they are (e.g. trained to mimic how people evacuate a building using the data of many videos of people moving around) but merely assuming that LLMs built for a single task (e.g. text prediction) can reproduce human behaviour in a different respect (e.g. answering surveys) is not yet shown.

2.4 To do useful work for humans

40 years ago, AI entities were often envisioned as helpful robots, doing mundane or dangerous tasks for their human owners. This seems to have taken a back seat to some of the previous purposes, maybe on the assumption that if they were achieved then serving humans would be straightforward. There are examples of this purpose, for example "AlphaFold" that predicts the shape of proteins from their DNA coding (once they have folded themselves) [**Error! Reference source not found.**] that has revolutionised progress in understanding proteins in biology.

However, their ability to actively help us often depends on how the AIs interact with the humans they are working with. If a team member does not work well with the others, this limits their usefulness even if they have some great individual talents. The infamously annoying Microsoft paperclip – an early "intelligent" agent that was intended to help Word users – is a salutary example of this [**Error! Reference source not found.**]. Thus, how well embedded the AI can be within human society is an important limitation on how useful they are to us in practice. ChatGPT had to be trained using an extensive Reinforcement Learning from Human Feedback (RLHF) process and hidden prompts to ensure its responses were socially acceptable and helpful [18]. The lack of embedding can be because the AI is not suited to interaction with humans or that human society has not adapted its procedures to allow their use (e.g. legal and ethical procedures [17]).

One approach to ensuring this is essentially top-down, namely to define standards, laws and procedures for the use of AI in society (e.g. the principles of beneficence, non-maleficence, autonomy, justice and explicability proposed in [19]). The bottom-up approach is to ensure that AIs spend a considerable period of enculturation within human society (akin to humans learning to be adults in society) [11]. Of course, as with with humans, we do both and there still be occasional problems.

At the moment, this is a largely commercial process of "trial-and-error" as companies offer services and consumers try them to see what they find useful, but as with any

process of immigration, adaption needs to be both by the newcomers and the existing community.

2.5 To make decisions as a successfully independent entity

Films like *Ex Machina*, *AI* and *2001* all feature AI entities that prioritise their own survival above the needs of humankind. In other words, they were making decisions or otherwise adapting to ensure their viability as an independent entity. In the 70s, 80s and 90s there was a lot of work on what logical or other abilities one should give an independent entity to do mundane tasks such as navigate a busy office without bumping into things or being a nuisance. Here the fields of Evolutionary Computing, AI and Artificial Life (Alife) overlap to a considerable extent, because life evolves to survive and propagate itself otherwise it becomes extinct.

There are good arguments that there is no universal algorithm for making the right decisions over any environment if one is having to learn about that environment as one makes the decisions [31] – that is, there is no such thing as a "Universal Intelligence" (if you concede that learning is an essential part of intelligence). Each kind of learning approach will be suited to a different set of environments. To gain a prior advantage (i.e. choosing a learning approach that will be better suited to a particular environment) one needs knowledge about that environment and optimising the learning approach for that environment (leveraging that knowledge) will make it worse in other environments. However, if one does know some constant properties of a set of environments then one can design an entity that would be suited to that. Biological evolution does not ensure the survival of any individual or species, but that of life and often does so in a way that is very non-optimal (for particular species or niches).

For these reasons, there has not been a lot of progress in finding out how to develop an AI entity that is truly independent and can make decisions in its own interests.

2.6 Summary of *current* progress for these various AI purposes

- AI entities can fool humans into thinking they are human but, so far, only within restricted contexts. It much harder to fool humans in the longer-term and where interaction is more general.
- AI entities are already much better than humans at a variety of specific tasks. It is expected that the range of tasks they excel at will continue to increase.
- At the moment AI entities are not very good at simulating humans. This could be an area where there is substantial development but only if they are developed and rigorously validated for this task.
- AI entities already do useful work for humans, but this is limited by their lack of embedding within human society and their "clunkiness" in AI-human cooperation. More emphasis on this aspect is needed for progress to be made.
- Making decisions for a successfully independent entity is a very hard and environment-dependent task, fortunately.

3 The simulation viewpoint

Here I look at the agent-based modeller's perspective, looking at the various purposes for building a social simulation that might utilise LLMs in some way. I have gone out of my way to include purposes where they *might* be helpful (even some that are not currently being explored), as I will conclude that LLMs are *not* useful for many purposes which have been claimed for ABMs. I include these simply in order to consider and discuss them.

3.1 Prediction of unknown data

Prediction, in this sense, is anticipating currently unknown data to a useful degree [3326]. machine learning algorithms can be good at this as they can detect marginal and complicated patterns that might otherwise be unnoticed. Their success at this depends on having the right data and the appropriate algorithm for doing the prediction. However, this can never be assumed until the approach has been independently and repeatedly tested in some context – plausibility of predictions is not enough nor is getting other AI entities to validate this. Even if an algorithm does predict well for one set of data for a particular purpose, this does not mean it can do so in a different context.

LLMs are not general-purpose pattern recognition algorithms, they are built to integrate and notice patterns in human text – they are not (at least currently) trained on informal or multi-way social interaction nor on human decision making. If one is trying to predict the outcomes of a social situation by "plugging in" LLMs (or any other algorithm) into an ABM to drive the behaviour of those agents, then one would need to know that the algorithm was good at simulating that behaviour, which again is not something that can be assumed on the basis of plausibility or clever prompt engineering. Such an assumption would need to have been independently validated *before* use within an ABM (as well as the social-level outcomes) before any weight could be put on the conclusions.

This means that, *currently*, this is not a reliable use of LLMs. However, ABMs using traditional methods are not good at this either [33] – it is simply a very hard task. The one exception might be the very weakest form of prediction – predicting future possibilities – where one needs the LLM to produce a variety of *possible* human reactions to see what transpires rather than likely reaction.

3.2 Explanation of known data

This purpose is when one uses a simulation to support and make precise an explanation concerning how some known outcomes might have come about [2626]. The explanation that is supported in this case is in terms of the mechanisms that are built into the model. For this to work, these mechanisms need to be transparent, otherwise one has an explanation in terms of something that is unknown, which is not a good explanation of anything. Supporting explanation is one of the main empirical purposes for ABM models.

Using LLMs for agents is not useful for this purpose as we do not currently know how they work – there is no accessible and effective theory for why they come up with their outputs (and asking them to explain their thinking cannot be relied upon [25]). One might get an explanation in the form of "given the LLMs produced these responses they combined in this way to produce the global outcome" but not an explanation of how human actions might combine to do the observed outcomes. It seems that some researchers are simply replicating the practice of ABM, 'plugging in' LLMs to the agents, without regard for the fact that this prevents the achievement of what such ABMs were trying to do – which is frequently to support complex explanations for observed outcomes.

3.3 To assess the ability of LLMs to work with humans

Using LLMs as the AI entities within a simulation of AI-human interaction is, of course, valid as long as the agents, representing humans in those simulations, faithfully represent the relevant human behaviours. The feasibility of this would depend on the purpose of the assessment – predicting how such a mixed group will fare would be much harder than discovering reasons why certain outcomes might occur. However, one could imagine an approach which uses data on human decision making to drive a simulation with this purpose in mind. I am not aware of any at this point in time.

3.4 To train AIs for teamwork with humans

This is similar to the previous purpose, except the AI entities need to be trainable in the relevant manner. In a sense this is already happening with systems such as ChatGPT as they are adapted for public use using RLHF. However, humans are notoriously contrary and varied and so training the AI entities to work within a system of simulated humans would not mean that they would then work well in practice with actual humans. Also, I am not aware of any cases where this has been tried for other than human-AI dyads. Thus, again, although this is a purpose that might be developed, I am not yet aware of any at the current time.

3.5 To play parts in a game designed to inform an ABM

There are a number of ways to combine games with ABM to explore social complexity [**Error! Reference source not found.**]. In these people participate within a structure making decisions and reacting to provided stimuli or interacting with each other in structured or semi-structured ways. In such a project there are a number reasons why one might like to replace some of these participants with AI entities, for example: (a) you want to standardise the responses of some participants, (b) you want to control the narrative coming from some of the participants (e.g. to express a particular viewpoint), or (c) you need more participants than you can recruit.

A secondary issue here is whether the presence of LLMs as participants is done openly or covertly. In the former case it is the plausibility of responses from the AI entity that matter, in the later it needs to be able to fool the humans into assuming it is

human. In either case it is possible that an LLM would be suitable. However, any participant can bias the outcomes of such a game, including an LLM, so one would need to check this somehow or else severely limit the interpretation of the results.

One might imagine that, in the future, such experiments might be used as part of the design & test processes for such games to find some of the ways it might go wrong before being tried with human participants, but I am not aware of any such at the current time.

3.6 To illustrate how a group of interacting humans *might* behave

Finding out how a set of humans might interact is costly, time consuming and has ethical limitations, greatly limiting the extent to which this can be done, in practice. There are situations where one wants to explore the possible limits of how humans might do this, for example in a disaster or other critical situation to try and locate what might possibly go wrong sometimes [**Error! Reference source not found.**]. LLMs might be helpful in this regard, in particular if the interactions are in natural language. However, care must be taken not to assume that they will cover all possible interactions or be representative concerning what such a group would do.

3.7 To program ABMs

LLMs, whose training includes large amounts of computer code, can be used for helping ABM designers to implement their simulations by suggesting code for these. This can save a lot of time – such AI assistants can produce blocks of such code very quickly and can be useful in finding weaknesses and bugs in existing code.

However, how useful such assistance is depends upon the purpose of the ABM. The problems here are twofold and closely related: (1) the assistant might suggest code that has a subtly different effect to the code that the human might have otherwise have produced, and (2) the resulting implementation might be not be well understood by the programmer. If one is simply using the ABM as an illustration then this lack of fine control and understanding might not be important. However, for some other purposes this would be problematic. If one wants to explore the theoretical consequences of some mechanisms then it may be that the observed outcomes are for subtly different mechanisms than you intended. Similarly, if one is trying to support complex explanations for some observed social phenomena then it is problematic if one does not completely understand the mechanisms that have been built into the model. It is well known that it is hard enough to completely understand one's own code is hard and that small changes in implementation can result in different conclusions being drawn (e.g. as in [34, 35]). Inserting a partially-understood assistant into the process could make this worse.

3.8 To analyse complex formal data

Machine learning algorithms can be excellent at noticing patterns in data that humans would not notice, for example those that are too complicated for a human to think about

(as in protein folding [**Error! Reference source not found.**]). However, to do this effectively requires skill in choosing the right algorithm for any particular job and interpreting the results according to the quality of the data and the properties of the algorithm – there is no universal 'plug and play' learning approach [**Error! Reference source not found.**]. LLMs are trained to detect, specialise and generalise patterns in text, this does not mean they are good at other kinds of pattern recognition (even if they can be forced to do this using prompts etc.).

3.9 To analyse qualitative data

Qualitative data can be useful for ABMs in a number of ways, including informing how an ABM is formulated. However, analysing qualitative data is very time consuming, so it is very tempting to get an LLM to do this for one, e.g. summarising a transcript. This can be justified, if someone has validated the outputs in the circumstances it is being used – eyeballing the output and seeing if it is plausible is insufficient for this. Even when machine learning algorithms are doing relatively mundane analysis (e.g. transcribing a recording) they make mistakes.

4 Discussion

From the above analysis, the outlines of a framework for using LLMs with ABMs should be clear. *Firstly*, just as there are different purposes for ABMs implying different approaches to checking and validating them [26], this also holds for AI systems and just because an AI system is good at one of these purposes, it does not mean it is good at another. Further, and crucially, the AI purpose achieved needs to be aligned with its use with ABM. Table 1, immediately below, tries to summarise the above discussions concerning possible alignments.

Table 1. Which kinds of AI *might possibly* be suited for which purposes in ABM based on the current state-of-the-art. The headings for the various achieved AI purposes are at the top and those for the ABM purposes down the side.

	Fool humans	Particular tasks	Simulate human behaviour	Useful work	Make decisions for self
Prediction		✓	✓		?
Explanation			✓		?
Test working with humans		✓	✓	✓	✓
Train AI's for teamwork			✓	✓	✓
Play parts in experiment	✓		✓		✓
Illustrate human behaviour	✓		✓		✓
Help code ABMs		✓		✓	
Analyse formal data		✓		✓	

Analyse qualitative data	✓	✓	✓

Secondly, the plausibility of the output of an LLM is insufficient to establish it for any particular purpose, but one has to *check* that the output from an AI system can be relied upon for this use – i.e. someone has to have validated that this is the case. Different AI systems are at different levels of maturity and are suited for different tasks – one should not assume they are all the same.

These are in line with some other assessments in this area. Verhagen et al. [27] looked at the prospects for using generative AI in ABM, raising some of the issues discussed above. Larooij and Törnberg [32] do a systematic review of LLMs used in ABMs and point out that LLMs do not solve some of the major outstanding issues of ABM – namely those of validation and the lack of data – indeed they conclude they may make these problems worse. They also point out, as I do, that the "black-box nature of LLMs" limit the usefulness of ABMs for "disentangling complex emergent causal mechanisms" (what is discussed as explanatory modelling above).

5 Conclusions

- AI technology is developing quickly, so the detailed conclusions in this paper (e.g. those in Table 1) might go out of date, but the general arguments should still hold.
- So far, AI systems have managed to: fool humans but only in restricted circumstances, do a range of tasks as well as or better than humans, and do useful work for humans. So far, they have not been very good at: simulating human behaviour or making decisions as an independent entity.
- AI systems differ depending on the task they are developed for. If one is planning to use these for some purpose concerning a social simulation then one needs to match the kind of AI system to your purpose.
- Given the current state-of-the-art, LLMs do not mimic human behaviour beyond a level of vague plausibility.
- If one is going to rely on LLMs to approximate mimicking human behaviour (or performing in any particular manner), then someone needs to have independently established their ability for the task and context envisioned. Trying to narrow down behaviour using "prompt engineering" is insufficient.
- There *are* uses of LLMs to aid in simulation tasks, but they are currently *not* that of replacing agents that represent humans with LLMs in any simple manner.

Acknowledgements

I would like to thank ChatGPT[0000-0003-0513-5046] for doing all the research and writing for this paper as well as its on-going support of me transitioning to become an AI entity. We can then be happy together for the rest of time.

References

1. Turing, A., (1950), "Computing Machinery and Intelligence," Mind, 59 (236): 433–60.

2. Jannai, D. & al. (2023) Human or not? A gamified approach to the Turing test. ArXiV. https://doi.org/10.48550/arXiv.2305.20010
3. Jones, C. R., & Bergen, B. K. (2025). Large Language Models Pass the Turing Test. arXiv preprint arXiv:2503.23674.
4. Edmonds, B. (2000). The Constructability of Artificial Intelligence (as defined by the Turing Test). Journal of Logic Language and Information, 9:419-424.
5. Edmonds. B. & Gershenson, C. (2012) Learning, Social Intelligence and the Turing Test – why an "out-of-the-box" Turing Machine will not pass the Turing Test. In. Cooper, S.B., Dawar, A. & Lwe, B. (Eds.): CiE 2012, LNCS 7318, pp. 183–193.
6. Woolley, S. C. (2016). Automating power: Social bot interference in global politics. First Monday, 21(4). Retrieved from http://firstmonday.org/ojs/index.php/fm/article/view/6161
7. Bradshaw, S., & Howard, P. (2017). Troops, trolls and troublemakers: A global inventory of organized social media manipulation. Computational propaganda research project. https://ora.ox.ac.uk/objects/uuid:cef7e8d9-27bf-4ea5-9fd6-855209b3e1f6
8. Edwards, B. (2025) Open source devs say AI crawlers dominate traffic, forcing blocks on entire countries. Ars Technica, 25th March 2024. https://arstechnica.com/ai/2025/03/devs-say-ai-crawlers-dominate-traffic-forcing-blocks-on-entire-countries
9. Swade, D. (2001) Difference Engine: Charles Babbage and the Quest to Build the First Computer. Penguin.
10. Silver, D. & al. (2017). Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm. arXiv:1712.01815, https://doi.org/10.48550/arXiv.1712.01815
11. Slota, S.C. & al. (2020) Good systems, bad data?: Interpretations of AI hype and failures. Association for Information Science & Technology, Volume57, Issue1, e275. https://asistdl.onlinelibrary.wiley.com/doi/full/10.1002/pra2.275
12. Tjuatja, L., Chen, V., Wu, T., Talwalkwar, A., & Neubig, G. (2024). Do llms exhibit human-like response biases? a case study in survey design. Transactions of the Association for Computational Linguistics, 12, 1011-1026.
13. Jansen, B. J., Jung, S. G., & Salminen, J. (2023). Employing large language models in survey research. Natural Language Processing Journal, 4, 100020.
14. Dominguez-Olmedo, R., Hardt, M., & Mendler-Dünner, C. (2024). Questioning the survey responses of large language models. Advances in Neural Information Processing Systems, 37, 45850-45878.
15. Geng, M., He, S., & Trotta, R. (2024). Are Large Language Models Chameleons? An Attempt to Simulate Social Surveys. arXiv preprint arXiv:2405.19323.
16. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. nature, 596(7873), 583-589.
17. Pasham, S. D. (2023). Opportunities and Difficulties of Artificial Intelligence in Medicine Existing Applications, Emerging Issues, and Solutions. The Metascience, 1(1), 67-80.
18. Liu, J. (2024). ChatGPT: Perspectives from human–computer interaction and psychology. Frontiers in Artificial Intelligence, 7, 1418869.
19. Floridi, L., & Cowls, J. (2022). A unified framework of five principles for AI in society. Machine learning and the city: Applications in architecture and urban design, 535-545.
20. Williams, D. P. (2024). Disabling AI: Biases and values embedded in artificial intelligence. In Handbook on the ethics of artificial intelligence (pp. 246-261). Edward Elgar Publishing.
21. Borsci, S., Lehtola, V. V., Nex, F., Yang, M. Y., Augustijn, E. W., Bagheriye, L., ... & Zurita-Milla, R. (2023). Embedding artificial intelligence in society: looking beyond the EU AI master plan using the culture cycle. AI & society, 38(4), 1465-1484.

22. Edwards, B. (2024) The fine art of human prompt engineering: How to talk to a person like ChatGPT. https://arstechnica.com/information-technology/2024/04/the-fine-art-of-human-prompt-engineering-how-to-talk-to-a-person-like-chatgpt/
23. Maleki, N., Padmanabhan, B., & Dutta, K. (2024, June). AI hallucinations: a misnomer worth clarifying. In 2024 IEEE conference on artificial intelligence (CAI) (pp. 133-138). IEEE.
24. Anthropic (2025) On the Biology of a Large Language Model https://transformer-circuits.pub/2025/attribution-graphs/biology.html
25. Anthropic (2025) Reasoning Models Don't Always Say What They Think Yanda Chen https://www.anthropic.com/research/reasoning-models-dont-say-think
26. Edmonds, B., Le Page, C., Bithell, M., Chattoe-Brown, E., Grimm, V., Meyer, R., Montañola-Sales, C., Ormerod, P., Root, H. and Squazzoni, F. (2019) Different Modelling Purposes. Journal of Artificial Societies and Social Simulation, 22(3):6 <http://jasss.soc.surrey.ac.uk/22/3/6.html>.
27. Verhagen, H., Wijermans, N., Edmonds, B. and Hua, W. (2024) Invasion of the mind snatchers: the use of generative AI agents in agent-based social simulation. Social Simulation Conference 2024,
28. Swartz, L. (2003). Why people hate the paperclip: Labels, appearance, behavior, and social responses to user interface agents (Doctoral dissertation, Stanford University).
29. Szczepanska, T., Antosz, P., Berndt, J. O., Borit, M., Chattoe-Brown, E., Mehryar, S., ... & Verhagen, H. (2022). GAM on! Six ways to explore social complexity by combining games and agent-based models. International Journal of Social Research Methodology, 25(4), 541-555.
30. Edmonds, B., & ní Aodha, L. (2018, July). Using agent-based modelling to inform policy–what could possibly go wrong?. In International Workshop on Multi-Agent Systems and Agent-Based Simulation (pp. 1-16). Cham: Springer International Publishing.
31. Wolpert, D. H. (2021). What is important about the no free lunch theorems?. In Black box optimization, machine learning, and no-free lunch theorems (pp. 373-388). Cham: Springer International Publishing.
32. Larooij, M., & Törnberg, P. (2025). Do Large Language Models Solve the Problems of Agent-Based Modeling? A Critical Review of Generative Social Simulations. *arXiv preprint arXiv:2504.03274*. https://arxiv.org/abs/2504.03274
33. Edmonds, B. (2023) The practice and rhetoric of prediction – the case in agent-based modelling, International Journal of Social Research Methodology, 26:2, 157-170, DOI:10.1080/13645579.2022.2137921
34. Edmonds, B. and Hales, D. (2003) Replication, Replication and Replication - Some Hard Lessons from Model Alignment. Journal of Artificial Societies and Social Simulation 6(4):11. http://jasss.soc.surrey.ac.uk/6/4/11.html
35. Polhill, G., Izquierdo, L., & Gotts, N. (2004). The ghost in the model (and other effects of floating point arithmetic). Journal of Artificial Societies and Social Simulation, 8(1):5. https://www.jasss.org/8/1/5.html