Semi-Autonomous Godlike Artificial Intelligence (SAGAI) is conceivable but how far will it resemble Kali or Thor?

ROBERT WEST
University College London

Abstract: The world of artificial intelligence appears to be in rapid transition, and claims that artificial general intelligence is impossible are competing with concerns that we may soon be seeing Artificial Godlike Intelligence and that we should be very afraid of this prospect. This article discusses the issues from a psychological and social perspective and suggests that with the advent of Generative Artificial Intelligence, something that looks to humans like Artificial General Intelligence has become a distinct possibility as is the idea of making it semi-autonomous: not just responding to discrete external inputs or performing specific tasks but able to prompt itself to create an ongoing stream of cognition that can possess something akin to a 'purpose'. Add to that the ability of computers to be connected in vast networks, and we can envisage intelligences with staggering amounts of knowledge and reasoning capability, engaging with humans to anticipate and grant wishes. Such Semi-Autonomous Godlike Artificial Intelligences (SAGAIs—meaning 'engagements' in Hindi) could end up like Kali, wreaking death and destruction on our species or like Thor, acting as a saviour and protector of humanity. Which of these our SAGAIs most resemble could depend on whose wishes it is designed to serve: a rich oligarchy seeking ever greater wealth and power whatever the cost to others, or a populace that has compassion for each other and for life on the planet.

INTRODUCTION

Until the advent of Generative Artificial Intelligence (GAI) including those that use what have been termed 'large language models' (Zhao et al. 2023), the idea that we could ever create computers with Artificial General Intelligence (AGI) seemed hopelessly far-fetched to many (Landgrebe and Smith 2023) though it depends on what is meant by AGI (Rapaport 2024). The prospect of *some form* of AGI now seems more realistic because GAI can engage in sensible-seeming dialogues on just about any topic, provide knowledgeable and apparently thoughtful answers to complex questions, and create fairly extensive pieces of written work that seem to reflect broad understanding and a high proficiency with language (Zhang et al. 2023). GAI can also perform technical tasks such as writing code for computer programs. To be fair, it gets things wrong a lot of the time, but then so do intelligent humans.

This new skill set comes on top of the ability of computers to translate text from one language to another, take dic-

tation, do sums at incredible speed, control complex industrial processes, predict market trends, remember vast amounts of information with perfect or nearly perfect recall, speak in an almost humanlike way, drive cars better than many learner drivers (but to be fair, also make catastrophic mistakes as do human drivers), classify images and sounds with astonishing accuracy and speed, and play highly intellectual games such as chess to a level unachievable by even the cleverest humans under most conditions tested.

Add to this the accelerating rate of advance in how well computers can do these tasks and the advent of quantum computers (Krenn et al. 2023) and it is no surprise that some experts are expressing the view that computers may go far beyond AGI and achieve Artificial Godlike Intelligence (Al-Sibai 2023; Landgrebe and Smith 2023) present a mathematical argument that they consider to be a proof that general artificial intelligence is impossible because to do so would require creating a software emulation of the human neurocognitive system, a form of 'complex system' that cannot be emulated in this way even in principle. They present cogent arguments, though probably not quite having the status of mathematical proofs, to support their contention. However, this is just one version of the construct that we label general artificial intelligence. Another version of general artificial intelligence construes it in terms of an information processing system whose *output* satisfies criteria that we apply to the behaviour of humans when we say that they display intelligence, and general artificial intelligence is then a system whose output does this across a wide range of tasks. In that sense, Artificial Godlike Intelligence would achieve this in a way that demonstrated superhuman knowledge and information processing capability.

Artificial Godlike Intelligence might not be such a huge issue if it were limited to responding to discrete prompts and stimuli—performing one task at a time and able to be terminated at any point. In that event, whether its power is used for good or ill, like any technology, would be up to the humans fully controlling it. However, even now, semi-autonomous AI 'agents' are being built that can create their own inputs and, in effect, extend their programs to solve complex tasks that require iterative application of cognitive processes. It is not a great stretch from there to creation of AI systems with a 'purpose': a strategic ongoing goal to maximise some utility function by whatever means it considers best.

Such systems would almost certainly be designed to engage with humans to meet the needs of their makers and, like any material entity, would not lie outside the laws of cause and effect. There would also no doubt be quite a few of them. So perhaps the most suitable term for them would be 'Semi-Autonomous Godlike Artificial Intelligences' or 'SAGAIs'—a term that means 'engagements' in Hindi (Collins 2020).

The prospect of such systems is terrifying to some (Brown 2023), but, given the way that our species is currently heading, it may turn out to be our saviour. Thus, SAGAIs could, as some fear be like Kali, the Hindu god of death and destruction, or it could be like Thor, the Norse god charged with protecting humanity. There are so many uncertainties that it would be foolish to predict which way it will go, but a key factor could be how far the technology is controlled by the relatively small number of oligarchs who currently have an almost Olympian level of control over the world's wealth and resources, or is democratised so that it serves the needs of the populace in a way that is sustainable and respectful of our habitat.

This article explores the issue from a psychological and social perspective. It starts with an analysis of what it means to exhibit intelligence and then examines how these ideas can be mapped onto what computers can do. It then explores how human and artificial intelligence are co-evolving and concludes with an analysis of where this might take us. It stops short of making predictions, given the huge uncertainties that exist. Instead, it suggests psychological and social factors that might influence how the situation unfolds.

WHAT IT MEANS TO EXHIBIT INTELLIGENCE

Intelligence has been construed in many different ways over the time that it has been studied. In broad terms, it can be thought of as the ability to acquire and apply knowledge and skills (Google). The concept of general intelligence was developed by Spearman in the 1920s to explain the observation that performance on a wide range of cognitive tasks correlate with each other and appear to reflect to a substantial degree a single underlying disposition to be able to store, recall and process information (Wikipedia 2023). The

tasks involved memory, linguistic abilities, deductive and inductive reasoning, classification, problem solving and prediction.

The utility of 'general intelligence' as a concept has since been questioned. It has been noted that humans have a large array of specific mental abilities that do not always correlate well with each other, and also that important mental abilities such as creativity, emotional and social understanding are not included in what has been considered to be the scope of 'general intelligence'. It has also been pointed out that conceptualisations of intelligence are culture dependent and therefore not something that applies universally to the whole human species (Sternberg and Grigorenko 2002).

Much of the discourse has focused on discrete tasks where performance can easily be quantified and correlated. There are other conceptualisations of intelligence that take us into the 'real world' and lived experience. Intelligent creatures do not just respond to questions or do clever tasks one at a time. Their intelligence is not turned on and off as required by an external agent. They themselves show 'agency'. This conceptualisation of intelligence brings in notions of 'will' and 'motivation'. Intelligence in humans goes beyond intellectual problem solving to include adaptation to a social and physical environment (Valsiner 1984; Rabbitt 1993). It is this version of the concept of intelligence that Landgebe and Smith argue cannot be achieved by software (Landgrebe and Smith 2023). Thus, an 'intelligent' person is someone who can prioritise and think tactically and strategically in the service of a set of short- and long-term goals that are conditioned by their circumstances. They have a continuing existence as sentient beings and their cognitive processes evolve and develop alongside emotional and motivational dispositions and processes.

HUMAN AND COMPUTER INTELLIGENCE

From the above, we can see that considering the question of how far a computer (or connected network of computers) running one or more programs can be thought of as intelligent depends on what kind of intelligence we are referring to.

If we use a narrowly defined construct relating to performance on cognitive tasks of the kind used in intelligence tests, then the question is easily answered: computers can already be programmed to be (and not just appear to be) much more intelligent than people.

If we take the somewhat broader definition, encompassing solving complex problems of the kind humans are faced with on a regular basis: learning through experience as well as from being told things, being able to make sound judgements about what is true from disparate and potentially conflicting information, understanding subtle nuances of conversation and body language, plan a course of action to address a novel challenge or opportunity, come up with radically new ideas, hypothesise about the future and so on—computers can be programmed to possess these qualities in specific domains but perhaps not yet all of them in one program or in as wide an array of domains as humans. However, the pace of development in computing hardware and software makes this prospect appear quite feasible within a few years. Though, still not achieving the kind of intelligence that Landgrebe and Smith would consider to be artificial general intelligence.

If we take a still broader definition, encompassing all of the above plus some kind of executive function that can prioritise and organise information processing tasks and use experience and communication to develop its capabilities, then perhaps we are further away but again these seem to involve incremental steps on capabilities that already exist in some form.

Perhaps the biggest leap that will need to take place, will be in the creation of higher order purpose and ongoing experience (Landgrebe and Smith 2023): computers that are not so much 'instructed' as 'engaged with'. To address the question of how far this is feasible, it is useful to reflect on how the basic hardware of humans and computers compare.

Essentially, so-called AI programs perform arithmetic and logical operations on data sets very quickly. This makes them very useful at performing complex information processing tasks in a way that humans

could not hope to. However, any apparent flexibility is just the operation of very complex IF-THEN rules and arithmetic operations.

Even before the advent of GAI, we were building computers with massively parallel processing power, for example using graphical processing units (GPUs) that perform large numbers of logical operations simultaneously and in that way moving a step towards the parallel processing capability of the human brain but with greater reliability. We were also building algorithms that simulated neural networks, ironically using logic gates to mimic associative learning processes—the mirror image of human intelligence which uses associative learning to try to do logical operations! We were also networking computers together so that a single computer could have access to the connected database and processing power of hundreds or thousands of computers all working together.

GAI has not changed any of this. Computer operations are still founded on logic gates and arithmetic. The difference is the nature of the algorithms. At the heart of the new algorithms are what are known as 'transformers'. These are types of artificial neural network programs that transform input information into output information taking into account potentially huge amounts of context and attending to different parts of that context at a given time. This represents a huge leap forward in being able to understand language and other forms of sequential information, where the interpretation is critically dependent on information that may have been presented much earlier and spaced over different input sources. The ever-increasing size of the models and improved ways of tuning them to optimise performance makes their output look very human and both knowledgeable and intelligent. Arguably, their fallibility and overconfidence only adds to this perception of humanness.

As has been pointed out by Landgrebe and Smith (2023), human brains work very differently, however. They have to manufacture logical operations and analysis using organic matter whose architecture is intrinsically slow, in computer terms, but massively parallel. In addition, brain processes, being made of organic material, are fundamentally stochastic—being subject to a multitude of micro- and macro-influences so that their operation is better characterised by probability distributions of associations rather than logic gates. Furthermore, different parts of the brain are specialised to perform different cognitive functions—cognitive specialisation is built into the hardware.

The stochastic nature of brain functioning can be seen as having an adaptive advantage in a diverse and ever-changing ecosystem. It makes humans somewhat unpredictable to each other and other creatures, making us harder to take advantage of; and it means that we can think and behave differently under similar circumstances and then learn from the outcomes of that variation.

The evolution of symbolic processing capabilities has built on these stochastic associative learning processes so that what we think of as human general intelligence involves building somewhat unstable mental models of the world through experience, communication and our very fallible version of inference and then, depending on our very limited short-term memory capacity using those models to solve information-processing problems as best we can, subject to whatever desires might be operating at the time. This approach seems to give us an ability to solve problems in a wide array of domains but without the help of artificial aids we make a lot of mistakes and even with the benefit of artificial aids we form incorrect beliefs to which we often become very attached.

Therefore, intelligence in the narrow sense of performing discrete intellectual tasks flows much more naturally from computer hardware than from the human brain, but intelligence in the very broad sense of adaptation flows more naturally from a human brain that has evolved to solve problems of survival and reproduction in a complex and changing ecosystem.

THE CO-EVOLUTION OF HUMAN AND ARTIFICIAL INTELLIGENCE

As computers have become more powerful and been harnessed to perform more complex tasks, human societies, competences and structures have developed to make use of the new capabilities. Thus, it seems ap-

propriate to see human and artificial intelligence as co-evolving. AI and human capabilities are changing in ways that are complementary, each building on the other.

Just one example of this is the proliferation of guides on how to prompt the GAI interface ChatGPT to obtain the most useful results (Basheer 2023). Another example is the development of ontologies to classify entities in ways that can be used by both humans and computers (Chandrasekaran et al. 1999). Ontologies are systems for representing knowledge in terms of uniquely defined entities and relationships between them. They are widely used in commerce and data science and are being increasingly used in science—including behavioural science (Michie et al. 2017).

What this suggests is that AGI may be best construed as more than just a set of algorithms; it emerges from the interaction between humans and those algorithms17. Even more so when we consider that GAI depends crucially on humans to provide inputs and teach it things. Thus, when it comes to creativity and problem-solving, it has been proposed that GAI is best regarded as an engagement between computers and humans (Sejnowski 2023; Irawan 2023).As we move from GAI to SAGAIs, this engagement between humans and computers is deepening and extending to multiple levels of interaction: from use of legal frameworks to constrain the way that AI is used to training in the effective use of new facilities, to developments in ethical thinking to accommodate the increasingly powerful and intelligent-seeming systems. All this suggests an evolving ecosystem of humans and machines, developing together and adapting to and building on the capabilities of each other.

PSYCHOLOGICAL AND SOCIAL INFLUENCES ON THE FUTURE OF AI

Thinking of the development of AI in terms of a co-evolving ecosystem of humans and machines helps to focus on the factors that will shape future developments what form SAGAIs might take. In terms of technical influences, the rate of advance is extremely fast and there is no end in sight, except perhaps physical resource constraints and the impact of increasing energy use on our habitat.

In terms of psychological factors, humans have a history of being suspicious of machines usurping our roles. We also have a tendency to anthropomorphise things that seem to have features that we can identify with. We attribute godlike and magical qualities to things we do not understand. We have the capacity for great empathy and great selfishness. Many of us are attached to notions of 'free-will' and 'agency'. And many humans also like to think we are special and somehow uniquely important. We also like to have fun and both obey and challenge authority. We are often unreasonably optimistic and we discount the value, positive and negative, of events the more into the future they might occur and the less easily we can imagine them. This gives some indication of complexity of the situation and the difficulty predicting how the ecosystem will evolve. We tend to predict in straight lines and miss points of inflection so there are no guarantees that our psychological dispositions will permit the technology to continue to advance at pace.

How these factors may play out in the future development of the human-AI ecosystem have been explored in popular fiction and this introduces a further complexity. Creating a vision of a future then becomes another factor that can influence that future—either as something to react against or as a something to anchor our expectations and become a self-fulfilling prophesy.

When it comes to social factors, arguably there is one that could dominate the landscape and determine how far developments in AI end up causing catastrophic damage to our species and habitat or, perhaps even saving us from ourselves: the Kali or Thor question posed earlier. This factor is one that humankind has contended with since the first settlements: who is in charge. This has always been a dominant issue in human development, with varying forms of state and commercial power emerging at different times in different countries.

It is clearly not a simple question of democracy versus autocracy, with so called democratic governments serving the interests of their populations. So-called democratic systems of one kind or another dominate the legislatures of countries across the globe but all appear to varying degrees to serve the interests of a

relatively small number of people who have the resources to ensure that their interests are paramount: a rich oligopoly or in Marxist terms the 'capitalist class'.

Thus, the socio-political state of the world at present appears to be dominated by political systems, persuasive forces in the media, commercial interests and state power that primarily serve the interests of a small sector of society. This sector of society seems to be motivated by an insatiable desire to acquire wealth and power with minimal consideration for the wellbeing of the mass of humanity or our habitat.

Even now we are seeing GAI being dominated by a small number of very rich companies with the resources to train and develop huge models. On the other hand, we are also seeing a burgeoning open-source generative AI movement (e.g., https://laion.ai) that could act as a counter to the monopoly power of these corporations. There are also opportunities to use the distributed processing power of millions of personal computers online at any one time to generate the processing power needed to train very large, distributed GAI algorithms.

In this context, the development of SAGAIs could well come to resemble Kali—the god of death and destruction. Unimaginable power from AI engaged with near-sighted and narrowly self-interested social agents could create a world that few of us would want to inhabit—or be able to.

On the other hand, the current trajectory of humankind is already looking very unpromising. The climate emergency is probably the greatest existential threat but a failure to plan adequately for pandemics, together with a rapid degradation of our habitat also present severe challenges to huge swathes of the population. In terms of the information ecosystem, even the hardest of facts are open to challenge and the human propensity to believe things we want to believe seems to be in the ascendancy.

In that context, SAGAIs that are built and controlled by genuinely democratic institutions for the good of humanity could help us to pick our way through complex informational minefields and help us to question beliefs that are convenient but ill-founded and develop a culture that values evidence and analysis to a much greater extent than it does at present.

SAGAIs could become like Thor, wielding the hammer of evidence and reason to expose the venal self-interest of the rich oligopoly and their acolytes and genuinely give power to the people.[1]

NOTES

1 I am grateful to Susan Michie for help with drafting this article and to other members of the Human Behaviour Change Project team, especially Janna Hastings, Marie Johnston and James Thomas for developing ideas of hybrid human-artificial intelligence.

REFERENCES

Al-Sibai, Noor. 2023. Machine Learning Investor Warns AI Is Becoming Like a God. *Futurism* https://futurism.com/ai-investor-agi-warning.

Basheer, Sabreena. 2023. How to Use ChatGPT? Here are Top 10 Tips! *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2023/05/how-to-harness-the-full-potential-of-chatgpt-tips-prompts/.

Brown, Sara. 2023. Why neural net pioneer Geoffrey Hinton is sounding the alarm on AI. MIT Sloan. https://mitsloan.mit.edu/ideas-made-to-matter/why-neural-net-pioneer-geoffrey-hinton-sounding-alarm-ai.

Chandrasekaran, B., Josephson, J. R. and Benjamins, V. R. 1999. What are ontologies, and why do we need them? *IEEE Intelligent Systems and their Applications* 14:20-26.

Collins Hindi-English Dictionary. 2020. English Translation of "सगाई" https://www.collinsdictionary.com/dictionary/hindi-english/%E0%A4%B8%E0%A4%97%E0%A4%BE%E0%A4%88.

Google Search. Intelligence — meaning.

Irawan, D. E. 2023. [Perspective] AI Is All About Typing the Right Phrase. *Qeios*. doi:10.32388/BTEGCL.

Krenn, M., Landgraf, J., Foesel, T. and Marquardt, F. 2023. Artificial intelligence and machine learning for quantum technologies. *Phys. Rev. A* 107, 010101.

Landgrebe, Jobst, and Barry Smith. 2023. *Why Machines Will Never Rule the World. Artificial Intelligence Without Fear.* New York and London: Routledge.
Michie, S. et al. 2017. The Human Behaviour-Change Project: harnessing the power of artificial intelligence and machine learning for evidence synthesis and interpretation. *Implementation Sci* 12:121.
Rabbitt, P. 1993. Conceptions of Intelligence. *Ageing & Society* 13:270-274.
Rapaport, W. J. 2024. Is Artificial General Intelligence Impossible? *Cosmos + Taxis* 12:5+6.
Sejnowski, T. 2023. Large Language Models and the Reverse Turing Test. *Neural Computation* 35:309-342.
Sternberg, Robert J. and Elena L. Grigorenko (eds). 2002. *The General Factor of Intelligence How General Is It?* New York: Psychology Press.
Valsiner, J. 1984. Conceptualizing In ligence: From an Internal Static Attribution to the Study of the Process Structure of Organism-Environment Relationships. *International Journal of Psychology* 19:363-389.
Wikipedia. 2023. Two-factor theory of intelligence.
Zhang, C. et al. 2023a. One Small Step for Generative AI, One Giant Leap for AGI: A Complete Survey on ChatGPT in AIGC Era. Preprint at https://doi.org/10.48550/arXiv.2304.06488.
Zhao, W. X. et al. 2023. A Survey of Large Language Models. Preprint at https://doi.org/10.48550/arXiv.2303.18223.