Archetypal failures of artificial general intelligence (AGI) – a generative guide

Agent-Based Modeling (ABM), crisis modelling, system design, human factors, artificial general intelligence (AGI)

(identifying information withheld)

Extended Abstract

The recent rise of large language models (e.g., Chat GPT) and other generative artificial intelligence applications (e.g., DALLE-2) has re-focused the global scientific, risk management and ethics community's attention on the range of potential societal risks posed by rapid advances in AI, including Artificial General Intelligence (AGI). Despite proponents' assurances that the primary goal of AGI development is to benefit humanity overall through the generation of beneficial outcomes aligned to human preferences, potential domain specific threats on the pathway to this lofty end-state remain. These include rapid and widespread job losses among skilled and low-skilled workers, inherent biases leading to inaccurate or unjust outcomes in the application of AGI algorithms, and/or unforeseen negative outcomes emanating from observer dependence in AGI architecture and/or narrow goal-seeking. For a technology whose potential benefit is so broad, few scholars or lay-people who engage seriously with the potential implications of AGI on their lives and society appear nonchalant about the imagined effects of its manifestation in the real world.

Recent work has attempted to explore archetypal risks associated with imagined future AGI architectures and functions. These risk archetypes have fallen into 6 main categories:

1. AGI removing itself from the control of human owners/managers (i.e., the risks associated with containment, confinement, and control in the AGI development phase, and after an AGI has been developed, loss of control of an AGI),
2. AGIs being given or developing unsafe goals (i.e., the risks associated with AGI goal safety, including human attempts at making goals safe, as well as the AGI making its own goals safe during self-improvement),
3. Development of unsafe AGI (i.e., the risks associated with the race to develop the first AGI, including the development of poor quality and unsafe AGI, and heightened political and control issues),
4. AGIs with poor ethics, morals and values (i.e., the risks associated with an AGI without human morals and ethics, with the wrong morals, without the capability of moral reasoning, judgement),
5. Inadequate management of AGI (i.e., the capabilities of current risk management and legal processes in the context of the development of an AGI), and
6. Existential risks (i.e., the risks posed generally to humanity as a whole, including the dangers of unfriendly AGI, the role of humans, and suffering of the human race).

Reflecting efforts to identify and categorise archetypal risks and failures in socio-technical systems (Senge, 1990, 2007), in this paper we use an agent-based model to attempt reproduction of previously identified archetypal AGI risks, as use the modelled representation in an attempt to give rise to the emergence of new risks that have not previously been identified.

Given that AGI does not yet exist, it is a challenge to model the potential future state or performance of a system whose architecture and capabilities remain speculative. The results of this work should therefore not be taken as the complete set of potential risks present. However, our motivation is to provide pre-emptive guidance to system designers, risk managers and policy-makers looking to exploit the stated rewards of AGI while also soberly considering the risk and control measures needed to prevent 'run-away' AGI development alongside undesired outcomes.

The initial model architecture consists of 6 layers of abstraction, each with n elements:

1) the world,
2) the AGI's perception of the world,
3) the AGI, its architecture, goals, and physical properties,
4) the communication interface of the AGI with people,
5) people's perception of the AGI's performance within and outside the communication interface, and
6) people and their desires.

As might be expected, conflict between goals arises in the model based on its limited ability to optimise all aspects of the world desired by people. This is compounded by errors and limitations in the AGI's perception of the world, its architecture, and physical properties as well as its effectiveness in communicating with people about its activities and the effect of those activities. People's perception of the performance of the AGI is heterogenous as are their desired goals.

In this work in progress, we demonstrate that existing archetypal risks associated with AGI can be generated through stylised representation in an agent-based model to the extent that the model can be useful in identifying potential risks associated with AGI systems ahead of time. In addition to the mapping of existing understood or envisioned risks, we explore the utility of ABM in identifying and transparently explaining the generation of new risks not previously discussed in the literature or modelled through reasoning, deduction or other thought experiments. We discuss the limitations of the representation and existing architecture in light of recent developments in our understanding of AGI.

References

Peter M. Senge (1990). The Fifth Discipline: The Art & Practice of The Learning Organization. London: Random House

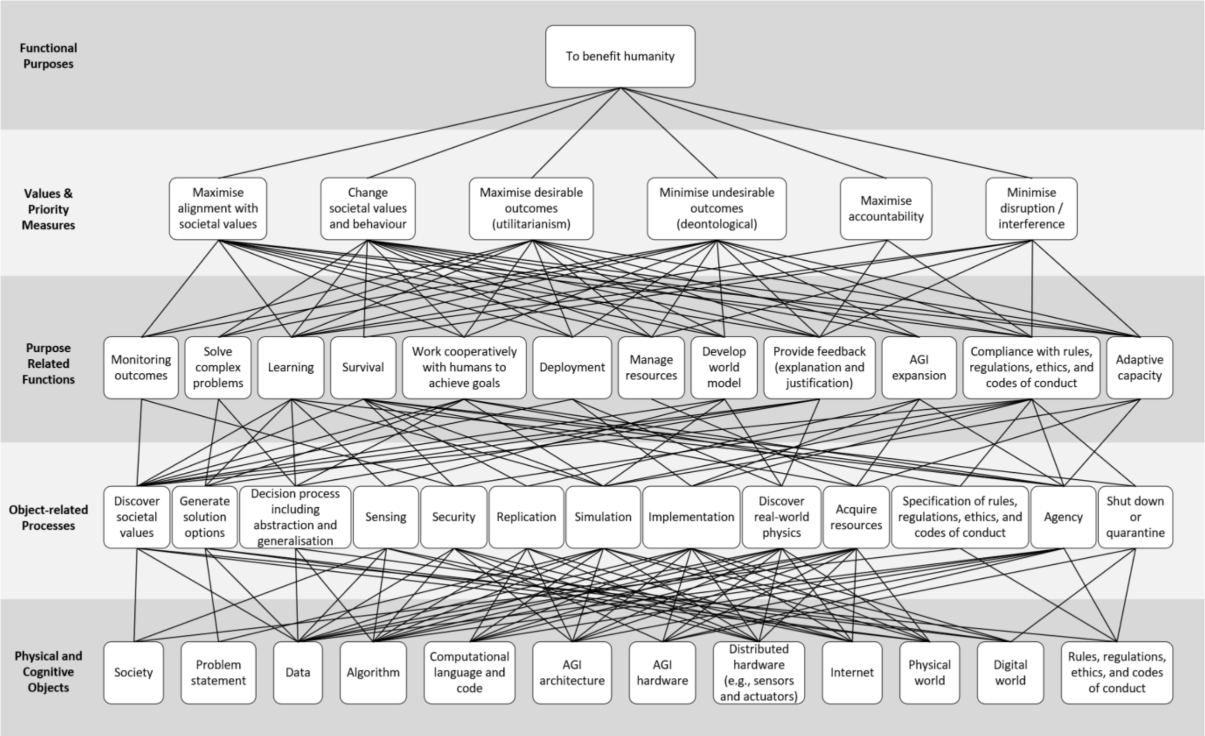
Functional Purposes
To benefit humanity
Values & Priority Measures
Maximise alignment with societal values
Change societal values and behaviour
Maximise desirable outcomes (utilitarianism)
Minimise undesirable outcomes (deontological)
Maximise accountability
Minimise disruption / interference
Purpose Related Functions
Monitoring outcomes
Solve complex problems
Learning
Survival
Work cooperatively with humans to achieve goals
Deployment
Manage resources
Develop world model
Provide feedback (explanation and justification)
AGI expansion
Compliance with rules, regulations, ethics, and codes of conduct
Adaptive capacity
Object-related Processes
Discover societal values
Generate solution options
Decision process including abstraction and generalisation
Sensing
Security
Replication
Simulation
Implementation
Discover real-world physics
Acquire resources
Specification of rules, regulations, ethics, and codes of conduct
Agency
Shut down or quarantine
Physical and Cognitive Objects
Society
Problem statement
Data
Algorithm
Computational language and code
AGI architecture
AGI hardware
Distributed hardware (e.g., sensors and actuators)
Internet
Physical world
Digital world
Rules, regulations, ethics, and codes of conduct

Figure 1. Abstraction hierachy for generic AGI produced through the work domain analysis.