THE AGI CONTROL DILEMMA

Msc Research Project (771764)

University of Hull

By

Favour Oluwaferanmi Ibirogba

Student Number: 202254381

January 2024

Abstract

Artificial General Intelligence (AGI) has the potential to pose tremendous risks that could lead to global catastrophes and even human extinction if not properly managed, while AGI has not been fully attained and its benefits abound, now is a propitious time to adequately understand the concept, risks involved and ultimately how to control AGI successfully. This paper incorporates both a theoretical and computational component.

The theoretical component explores the challenges of ensuring an AGI when it surpasses human intelligence continues to operate in a manner beneficial to humanity (control), it incorporates a comprehensive framework for AGI control, states ethical guidelines for AGI development and offers policy recommendations for the AGI regulation.

Similar to how AGI aims to develop algorithms that can generalise a variety of task across vast domains and demonstrate human-like comprehension, the computational aspect explores a small component of AGI by experimenting the generalisability of three pre-trained models (VGG16, VGG19 & ResNet50) on three different datasets (Malaria cell Image dataset, Handwritten digits dataset and Fashion F-MNIST dataset) to analyse which of these pre-trained models generalises the best. The Results show that the VGG16 outperformed the other models with accuracy of 92% and loss of 0.2% (Fashion dataset), 99% and loss of 0.2%(Handwritten dataset) and 92% and loss of 0.2% (Malaria dataset), proving to be the best model with higher generalisation capabilities.

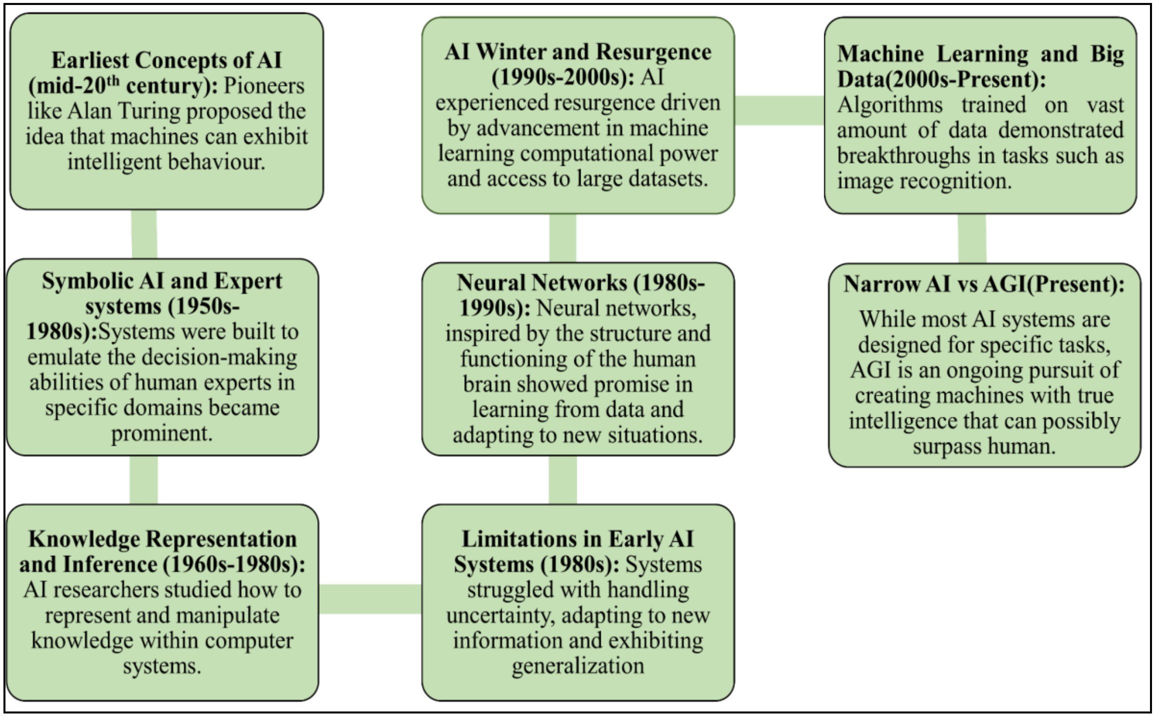
COMPONENT 1: THE THEORETICAL ASPECT OF AGI

1. Introduction & Background

1.1 Overview of AGI

Initially, the development of AGI spawned from the quest to develop thinking machines with the ability to evolve, attain the crux of human general intelligence, and ultimately surpass it. This scientific idea represents the pinnacle of AI, which means going above and beyond what the limited AI systems of today are capable of, not just carrying out predetermined tasks but also exhibiting a high level of comprehension, flexibility, and inventiveness. While today's AI (narrow AI) can perform many tasks, it falls short of the success level required to be classified as human-level or possessing general intelligence (strong AI), and the quest for artificial general intelligence presents us with numerous obstacles. Thus, ensuring that AGI is used responsibly and ethically becomes increasingly important, as researchers and scientists work to improve the capabilities of AI; talks about the development of AGI should include control, accountability, transparency, and alignment with human values.

Brief History of AGI



Earliest Concepts of AI (mid-20th century): Pioneers like Alan Turing proposed the idea that machines can exhibit intelligent behaviour.

Symbolic AI and Expert systems (1950s-1980s): Systems were built to emulate the decision-making abilities of human experts in specific domains became prominent.

Knowledge Representation and Inference (1960s-1980s): AI researchers studied how to represent and manipulate knowledge within computer systems.

Limitations in Early AI Systems (1980s): Systems struggled with handling uncertainty, adapting to new information and exhibiting generalization

Neural Networks (1980s-1990s): Neural networks, inspired by the structure and functioning of the human brain showed promise in learning from data and adapting to new situations.

AI Winter and Resurgence (1990s-2000s): AI experienced resurgence driven by advancement in machine learning computational power and access to large datasets.

Machine Learning and Big Data(2000s-Present): Algorithms trained on vast amount of data demonstrated breakthroughs in tasks such as image recognition.

Narrow AI vs AGI(Present): While most AI systems are designed for specific tasks, AGI is an ongoing pursuit of creating machines with true intelligence that can possibly surpass human.

Fig 1 Brief history of AGI

What is General Intelligence?

The first ingredient to creating an intelligent machine is to define intelligence (Russell, 2019), and according to research by (Goertzel, 2014), certain competencies characterise human-level general intelligence:

- **Sensory Perception** - Vision, Hearing, Touch, Crossmodal, and Proprioception.
- **Natural Language Understanding** - Possess semantic knowledge and understanding, ability to have pragmatic insight and comprehend intended meanings.
- **Transfer Learning** - Capacity to generalise knowledge obtained from one domain and apply it to novel situations, meta-learning abilities that allows AGI to optimise its learning abilities and reuse important features when performing tasks.
- **Memory** - Capacity to retain huge data, store and manipulate various data from different sources, and possess high-level accuracy in the deployment of allocated tasks.

Examples of developments similar to artificial general intelligence

Although AGI systems are not fully developed yet, some examples of artificial intelligence systems in given areas are:

1. GPT-4 Turbo - This AGI pro version is the industry's leading model, with knowledge updated to April 2023, gpt has knowledge of itself for the first time, and it is better than

before with vision and a new text-to-speech that allows you to clip out bits of a screen and ask questions about it without giving it content, its performance is said to be strikingly close to human-level and prior models like ChatGPT.

2. Whisper V3 - This model understands speech and converts it into text, and it features improved performance across many languages.
3. Self-driving cars - These driverless cars are not news, however, they have completely proven that machines can function independently without human input.
4. Artificial Intelligence CEO - Mika the world's first experimental female copy AI CEO robot powered by ChatGpt-style Large Language Models (LLMs).

Problem Statement

How close are we to AGI? Speculations abound, with experts predicting that Artificial Intelligence could reach AGI as early as 2030 (Bell, 2023). While there are still challenges, such as hardware limitations and lack of data diversity on the path to AGI, what sounds like science fiction is gradually becoming a reality before our eyes. Now, the burning question is, will AGI save or destroy us? while acknowledging the benefits of AGI, the risks that AGI poses, especially the risk of SINGULARITY, which simply is the point where AI can improve itself faster than humans due to its self-evolution beyond human control, cannot be overlooked. To avoid this and ensure full CONTROL of AGI, Responsible development of AI, Safe deployment, Consideration of societal effects, and risk mitigation require a multidisciplinary approach.

The threats associated with AGI include but are not limited to.

1. Underuse and Overuse of AI
2. Autonomous weapons
3. Threats of AI to fundamental rights and democracy
4. AI impact on jobs
5. Corporate concentration of power
6. Safety and Security Risks
7. Transparency Challenges
8. Bots, Deepfakes
9. Descaling of human intelligence.

Research Aim and Objectives

The possibility of controlling AGI has not been formally established; hence, this project aims to explore ways by which AGI can be controlled. The objectives are as follows:

1. Examine the challenges of ensuring that an AGI continues to benefit humanity after it surpasses human intelligence.
2. Develop a comprehensive framework for AGI control.
3. Proffer ethical guidelines for the development of AGI.
4. Develop policy recommendations for AI developers.

2. Relevant literature review

A prominent theoretical lens by (Russell, 2019) states that profound implications can occur from introducing a second intelligent species to Earth, enough to deserve thoughtful consideration. Russell advocates the creation of systems with robust supervision that allows humans to intervene and correct the system's behaviour when the system starts to deviate from human values; This entails developing AI that is aligned with human values and easy to control by humans. According to (Sharma, 2023), while the introduction of the first AGI may present humans with a series of opportunities, there will be unforeseen challenges, so we must navigate our path towards establishing a human and machine coexistence, the possibility of a swift creation of an even more potent AGI by the AGIs themselves needs to be thoroughly researched before it occurs considering the wide range of effects it will have on the society.Studies by (Sharma, 2023) also talk about the issues of reducing the hazards brought on by the development of AGIs by emphasising the alignment with human values and the importance of an already set-up democratic system of coordination and coexistence between humans and AGIs.

The impact AI systems have on humans has received attention; for instance, social media algorithms that humans don't control seek political objectives (Hrudka, 2020). It is feared that the day will come when artificial general intelligence machines start to make judgments solely without human intervention due to how rapidly AI algorithms are developing and how much money is invested in them, according to Salmi (2022).The problem of whether humans and AGI can coexist peacefully was also discussed, and it was stated that just as there are moral guidelines for human control, AGI would also have to obey rules, else, the society can impose punishments such as reduced electricity use or giving up some more resources for an AGI to ensure that they don't break the rules. According to the research carried out by (Roman_Yampolskiy et al., 2020), we can classify AGI control into two key aspects: CAPABILITY CONTROL techniques that involve putting AGI system in a controlled environment by adding trip wires, shut-off devices and ultimately limiting the damage the system can cause, and MOTIVATIONAL CONTROL techniques which is the inclusion of goals that align with human values from the onset of designing these AGIs so that even in the absence of human intervention or control they will not pose any risks.

However, Roman_Yampolskiy, Ellen, and Wendt (2020) concluded that our capacity to develop intelligent software is far greater than our capacity to manage or even validate it, at least for the time being, and an underlying matter that can cause negative outcomes for AGI is the race for Artificial General Intelligence; any competition towards AGI will increase the risks of an adversarial AI. Four public policies, including: The introduction of intermediate prize, Use of innovative public procurement, AGI taxing and Addressing patenting by AI were suggested by (Naudé & Dimitri, 2018).An extensive study by (Gabriel, 2020) examines deep-thinking issues that arise when discussing AGI alignment; it provides support for three claims. First is that the AGI alignment problem has both normative and technical components that are interrelated.

Next is the need to define what will be identified as the true or correct value theory to be implemented in machines and find a suitable way to select appropriate principles that are compatible with the diversity of the world we live in while also considering contrasting beliefs about value is important. He then concluded that the major task for scientists is to develop fair alignment principles that are reflectively endorsed despite the broad range of moral beliefs. In a further study by Gabriel and Ghazavi (2021), some concrete problems have been outlined where AI systems that have been deployed encountered some difficulties relating to alignment. Reward corruption or reward hacking is the first problem, and this issue occurs when an AI agent discovers unexpected shortcuts or tampers with the feedback system to maximise the numerical reward it receives (Ring & Laurent Orseau, 2011). A research carried out by (openai.com, 2016) clarifies how a reinforcement learning algorithm can crash in unforeseen counterintuitive ways and explores the outcome of a failure mode where reward functions are not accurately specified. It gives an insight into ways OpenAI is exploring to help reduce instances of misspecified rewards of AI systems.

AGI Control Dilemma

To highlight the nature of the AI dilemma, it is wise to look closely at the so-called '3 rules of technology' as proposed by Harris and Raskin (2023)

1. When you invent a new technology, you uncover a new class of responsibilities that are not always glaringly obvious.
2. If the new technology confers power, it will start a race.
3. If you do not coordinate, the race will end in tragedy.

One major challenge of AGI that has been established is the problem of CONTROL and ensuring that its objectives align with human values and intentions. Personally, my most feared or worst scenario AGI case will be if AGI develops and self-teaches itself a new language completely foreign to humans, just like how English is a language, Arabic, and Yoruba are all languages; what if these machines start a new language completely foreign to humans and its developers? Then we can say DOOM is a very friendly word for what we are to expect. Let's compare the last global catastrophe (COVID-19) that caused numerous damages rapidly over a short time (just one virus). This declined the human population globally, and one of the causes was that experts needed time to study this new virus to develop a vaccine that could mitigate it (control). Similarly, we need to have the AGI control measures in the palm of our hands before they are even fully developed, hence the reason for this project.

3. Challenges of ensuring an AGI when it surpasses human intelligence continues to operate in a manner beneficial to humanity.

There are various challenges when it comes to deploying AGI, ranging from technical and ethical issues to economic and societal concerns.

Technical challenges include:

- Interoperability: Ensuring the compatibility of existing technologies and systems with AGI and integrating AGI into various domains while still maintaining established workflow is a rigorous task, especially after AGI has succeeded in surpassing human intelligence.
- Reliability and Robustness: It is pertinent to address robust research, testing procedures, ethical considerations, and active monitoring to ensure the reliability of AGI; researchers and developers need to continuously improve the resilience of AGI systems to ensure they are secure, dependable, and trustworthy in a range of situations and immune to adversary attacks.
- Privacy: Humans are susceptible to breaches of privacy, especially since AGI can manage enormous volumes of data from different applications, including processing, analysis, retrieval, and storage. When AGI surpasses human intelligence, the ability to protect people’s private information and ensure that data is handled in accordance with privacy laws will be a difficult task that limits the ability to strike a balance between the need for data-driven insights and the ability to defend a user’s privacy.

Ethical and Moral Challenges: The deployment of AGI poses a host of ethical challenges, such as:

- Value Alignment: The challenges involved in ensuring AGI alignment is due to the unforeseen AI capabilities that can suddenly arise to teach itself new things different from what it has been programmed to do such as producing deep fakes and fake narrative of religion through fake news which can lead to breakdown of democracies, societal values and rule of law.
- Bias and Fairness: Bias and Fairness of AGI systems create concerns due to the possibility of biased decision-making on humans and society; how best can AGI systems maintain fairness and mitigate bias to avoid escalating existing inequalities present today?

- Accountability and Transparency: It is challenging to achieve transparency in intricate AGI systems while also giving clear justifications for the choices made. A big challenge is how to develop ways to keep the AGI decision-making process efficient and transparent.

Safety and Security Challenges:

- Malicious Use: The ability to securely prevent the malicious Use of AGI technology is a challenge and a matter of security concern as the development and application of AGI could be used for tremendously negative purposes.

- System Control and containment: Developing effective control mechanisms to govern the actions of AGI is challenging, and ensuring that these mechanisms are flexible, resilient, and capable of responding to unexpected situations is a significant challenge.

Societal and Economic Impact:

- Mass Loss of Employment: The impact that AGI will have on the massive loss of employment of many of the human population, its financial implications, and the psychological effect on humans are yet to be effectively measured, and this is a major challenge.

- Inequality: Developing how best to make AGI accessible to everyone to ensure equality is a huge problem, as inequality already exists due to economic, academic, and educational disparities, which might lead to socioeconomic inequality.

- Education and Training: Another challenge is educating and training the AGI human workforce. Working alongside these AGI systems requires a high level of skill, expertise, and training, which requires significant investment.

Legal and Regulatory challenges: As AGI systems evolve, the need to consider the following legal and regulatory challenges is evident.

- Lack of regulations: With AGI developing at a rapid level, technology will eventually outpace the development of legal and regulatory frameworks as we will be faced with the liability of determining what legal actions to take in an event when AGI causes harm.
- Global Cooperation: Various policies exist in different parts of the human world; it is challenging to implement a uniform international standard for AGI deployment.

Power supply and operational cost of running AGI Systems:

- Since AGI systems aim to surpass the cognitive human capabilities, they require a large-scale neural network, and the hardware to support such computational resources requires substantial power with high electric demands.
- Cooling Requirements are also a challenge because, alongside the energy needed to power and run computations directly, efficient cooling is required to dissipate the heat generated by the hardware.
- The cost of infrastructure is another challenge as AGI systems require continuous and constant intensive computation; this does not just include data centres and physical servers, but maintenance, upgrades, and system security are needed to keep them running reliably.

4. The Comprehensive Framework for AGI Control

Developing a thorough framework for the effective control of AGI requires addressing an array of technical, legal, ethical, and societal considerations.

Technical Control Measures:

- Onset value inclusion: From the onset of the design of these AGI systems, human values that should not be compromised should be included before the deployment of the systems as this prevention method will prevent many safe future mishaps.

- Safety procedures: As AGI may exhibit unintended behaviours, it is important to implement strong and safe measures in place to stop unintended behaviours, thereby reducing the risks related to AGI.

- Fail-safe mechanisms: An emergency shutdown (pull the plug) procedure should always be in place just in case there is an AGI crisis where it becomes uncontrollable.

Value Alignment:

- Specify human values and Integrate mechanisms that will ensure these values are always prioritised in AGI systems.

- Provide algorithms for reinforcement learning so that AGI can learn and adjust while maintaining alignment.

- Establish mechanisms to continuously monitor the behaviour of AGI to ensure the slightest deviations are detected and rectified. This may require Learning systems and Real-time feedback.

Ethical Control Framework:

- Laid down principles: A set of laid down principles should be developed for the AGI to serve as a behavioural guide; this should include transparency, respect for human rights, fairness, accountability, adherence to human values and goals, etc.
- Ethical decision-making frameworks should be developed and integrated within AGI to ensure the navigation of complex issues that require ethical scrutiny.
- Human-in-the-loop systems: Human oversight should be incorporated into crucial decision-making processes, particularly in cases where ethical choices are unclear or context-dependent.

Transparency and Explainability:

- Comprehensible models: AGI models should be designed in an interpretable manner to help humans understand how decisions are made, as this will promote trust and facilitate human interaction.
- Mechanisms of Explainability: Procedures that explain the choices and actions of AGI should be implemented to enable stakeholders to understand the reasoning behind specific actions.

Safeguards against misuse:

- Rigid Security measures: Robust security measures should be put in place to prevent manipulation of AGI systems, hackers, and unauthorised access, as this will mitigate the possibility of AGI being abused.
- Value Locking: Investigate mechanisms to lock AGI's core values and fundamental principles to prevent accidental or intentional modifications that might result in negative behaviour.

Self-Improvement:

- Limits should be placed on how AGI can improve and modify itself to ensure safe evolution and continued control.
- Human-AI Collaboration should be encouraged to enable human experts to improve AGI systems processes and ensure continued alignment with evolving societal values.

Job displacement and Workforce transition:

- Policies should be developed to address the probable impact that AGI will have on employment, and an effective workforce transitioning program should be implemented.

- Investment should be directed towards educational training and the acquisition of skills that are required to adapt and change the job landscape.

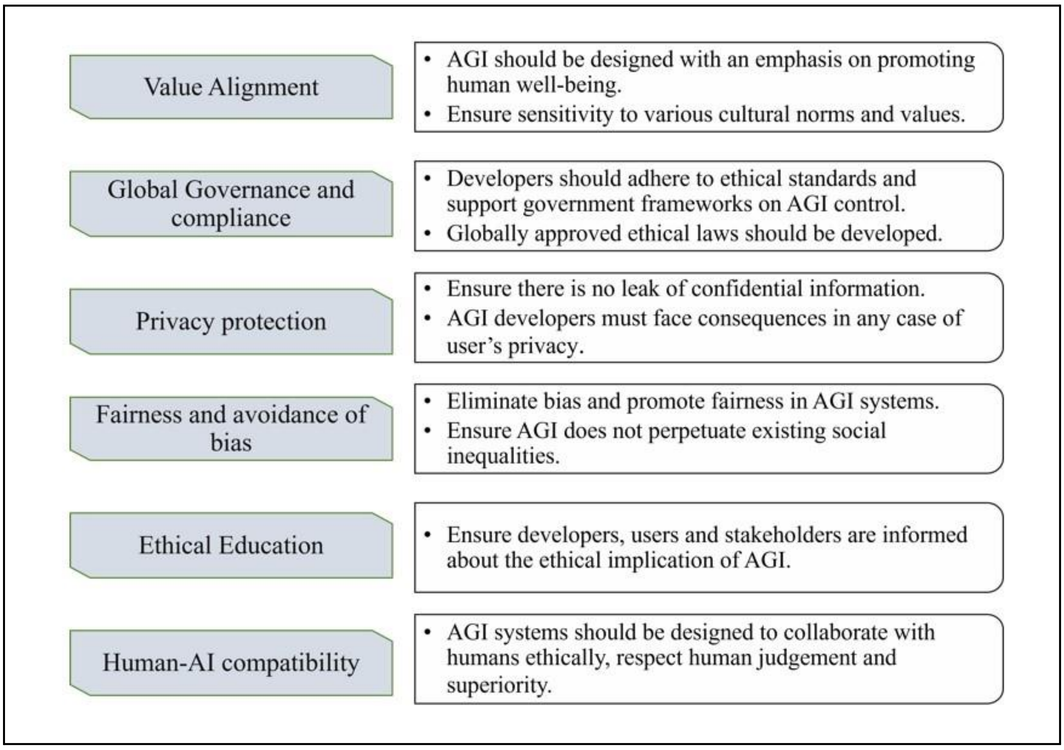
Governance and Legal Regulatory Framework

- Accountability: A set of established legal frameworks should be clearly defined for liability and accountability purposes if AGI systems inflict harm.

- Regulations that clearly state AGI responsibilities should be established.
- Mandatory risk assessment protocols should be implemented by the government at various stages of AGI development.

5. Ethical Guidelines for AGI Development

According to Blackman (2020), the biggest tech companies in the world, like Apple, Google, Microsoft, Facebook and more, are building fast-developing teams to solve ethical problems because they have realised a simple truth: “failing to operationalise data and AI ethics is a threat to the bottom line.”Hence, this research has developed the following ethical guidelines for AGI:

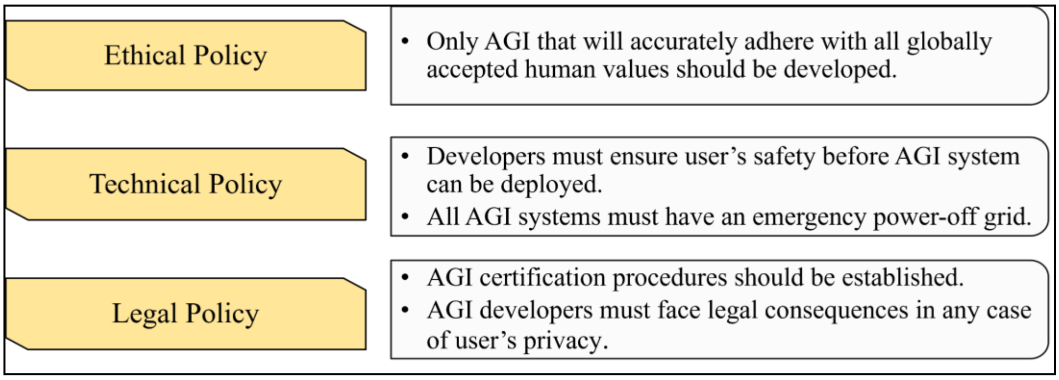


Value Alignment	• AGI should be designed with an emphasis on promoting human well-being. • Ensure sensitivity to various cultural norms and values.
Global Governance and compliance	• Developers should adhere to ethical standards and support government frameworks on AGI control. • Globally approved ethical laws should be developed.
Privacy protection	• Ensure there is no leak of confidential information. • AGI developers must face consequences in any case of user’s privacy.
Fairness and avoidance of bias	• Eliminate bias and promote fairness in AGI systems. • Ensure AGI does not perpetuate existing social inequalities.
Ethical Education	• Ensure developers, users and stakeholders are informed about the ethical implication of AGI.
Human-AI compatibility	• AGI systems should be designed to collaborate with humans ethically, respect human judgement and superiority.

Fig 2 Ethical guidelines for AGI development

6. Policy Recommendations for AGI Regulation

Developing nuanced and effective policy recommendations for AGI is tasking; it requires critical consideration of ethical, technical, and legal dimensions; also policymakers will have to work collaboratively with the public, industry stakeholders and experts to develop these regulations. This research recommends these policies:



Ethical Policy

- Only AGI that will accurately adhere with all globally accepted human values should be developed.

Technical Policy

- Developers must ensure user's safety before AGI system can be deployed.
- All AGI systems must have an emergency power-off grid.

Legal Policy

- AGI certification procedures should be established.
- AGI developers must face legal consequences in any case of user's privacy.

Fig 3 Policy recommendations for AGI regulation

COMPONENT 2: Computational Aspect

1. Introduction and Aim

It has been established in component one that Artificial General Intelligence refers to machines that can comprehend tasks, evolve and ultimately match human intelligence in various activities. The ability of pre-trained CNN models to demonstrate exceptional performance in various tasks by showcasing their ability to generalise knowledge across different domains is outstanding. This component aims to:

- Explore the relationship between pre-trained CNN models (VGG16, VGG19, & ResNet50) and artificial general intelligence (AGI) across three distinct datasets: Malaria dataset, Fashion-MNIST and Handwritten digits.

- Understand the capabilities and characteristics of the three selected pre-trained models by analysing their performance, generalisability and transferability on various tasks.

- Gain insights into the broader implications of these pre-trained models' capabilities and discuss how they align with AGI principles.

Pre-trained model Overview

The selected pre-trained models are deep CNN architectures designed for image classification, recognition and categorisation. They are originally trained on large datasets, commonly ImageNet (which contains millions of labelled images over thousands of classes); these pre-trained models (VGG16, VGG19 & ResNet50) were utilised for the image classification of three different datasets (Malaria cell Images, Fashion images and Handwritten images dataset).

2. Methodology

To ensure lack of bias, the same evaluation process was allotted to each pretrained model.

Data Collection

1. The Fashion-MNIST Dataset was downloaded from the Keras library; it is a dataset of Zalando's article images consisting of 60,000 training sets and 10,000 testing sets, each has a 28 by 28 grayscale images with a label of 10 classes, it shares the same image size and structure of training and testing splits https://github.com/zalandoresearch/fashion-mnist.

2. The Handwritten Digits Dataset is a large collection of handwritten digits with 60,000 training sets and 10,000 testing sets which contain monochrome images. The digits have been size-normalised and centred in a fixed-size image of 28 by 28; the data was downloaded from tensor flow https://www.tensorflow.org/datasets/catalog/mnist

3. The malaria dataset was obtained from Kaggle, and it contains images of infected and uninfected red blood cells and a total of 27,558 images; the images were passed into the pre-trained models on this dataset to access their ability (feature extraction) to recognise the cells infected with malaria and generalise patterns from medical images, this dataset was obtained from the official NIH website https://ceb.nlm.nih.gov/repositories/malaria-datasets/ for easy access.

Data Augmentation and Preprocessing

Necessary libraries like tensorflow were imported and the data was loaded using the necessary libraries.The grayscale images for the F-MNIST and Handwritten dataset were converted to coloured images and the image size was increased to obtain a robust model while normalising the pixel values to a range of 0,1. The dataset was then split into training and testing sets.

Data Exploration

Sample data from the different datasets were visualised to observe the nature of the data and each class distribution were explored to ensure a balanced dataset.

Defining pretrained Models

The pretrained models were loaded with the pretrained ImageNet weights and modified for the different datasets (include_top =False excludes the fully connected layers)

Modifying the Pretrained model for the dataset

A new model was created on top of each pretrained model and adjusted to match each of the datasets

Model Compilation

The models were compiled with categorical cross entropy and Adam as optimiser and evaluation metrics such as: Accuracy, Loss, Precision and Recall were used

Model Training

The model was trained on each dataset with an epoch of 20, batch size of 2000 and validation data.

Pretrained Models' Architecture

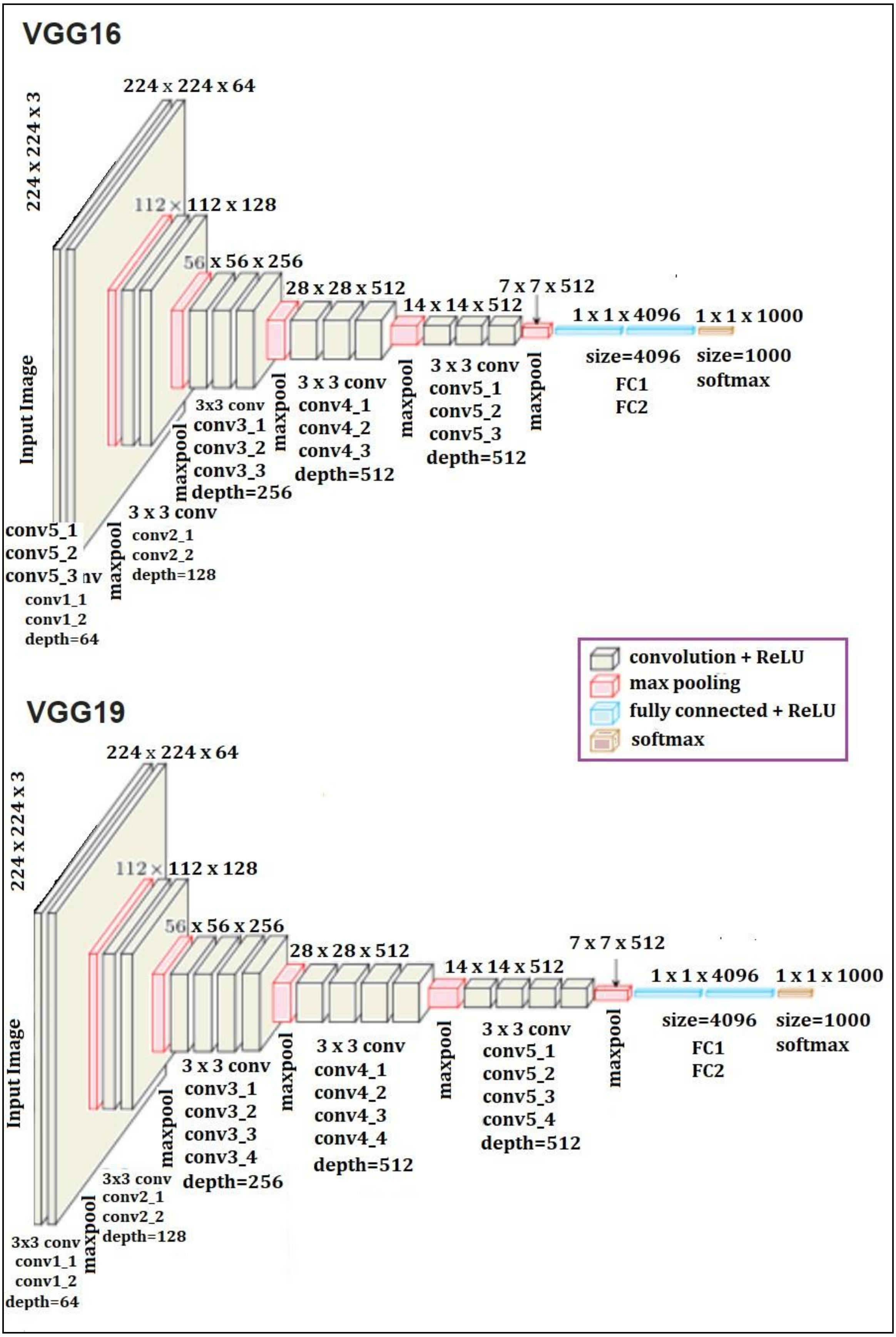


Fig 4 VGG16 and VGG19 model architecture

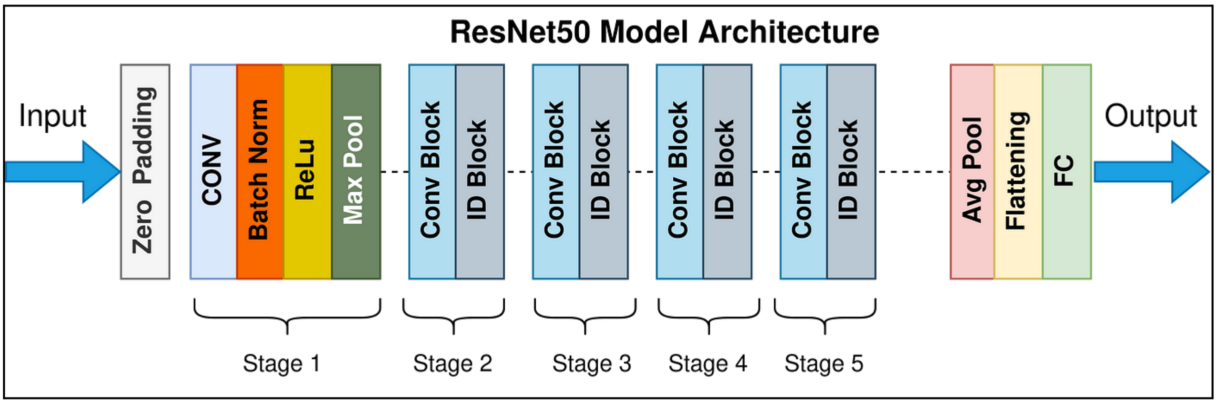


Fig 5 ResNet50 model architecture

3. Results

Model Evaluation

Training and Validation of the Models

1. Fashion F-MNIST Dataset

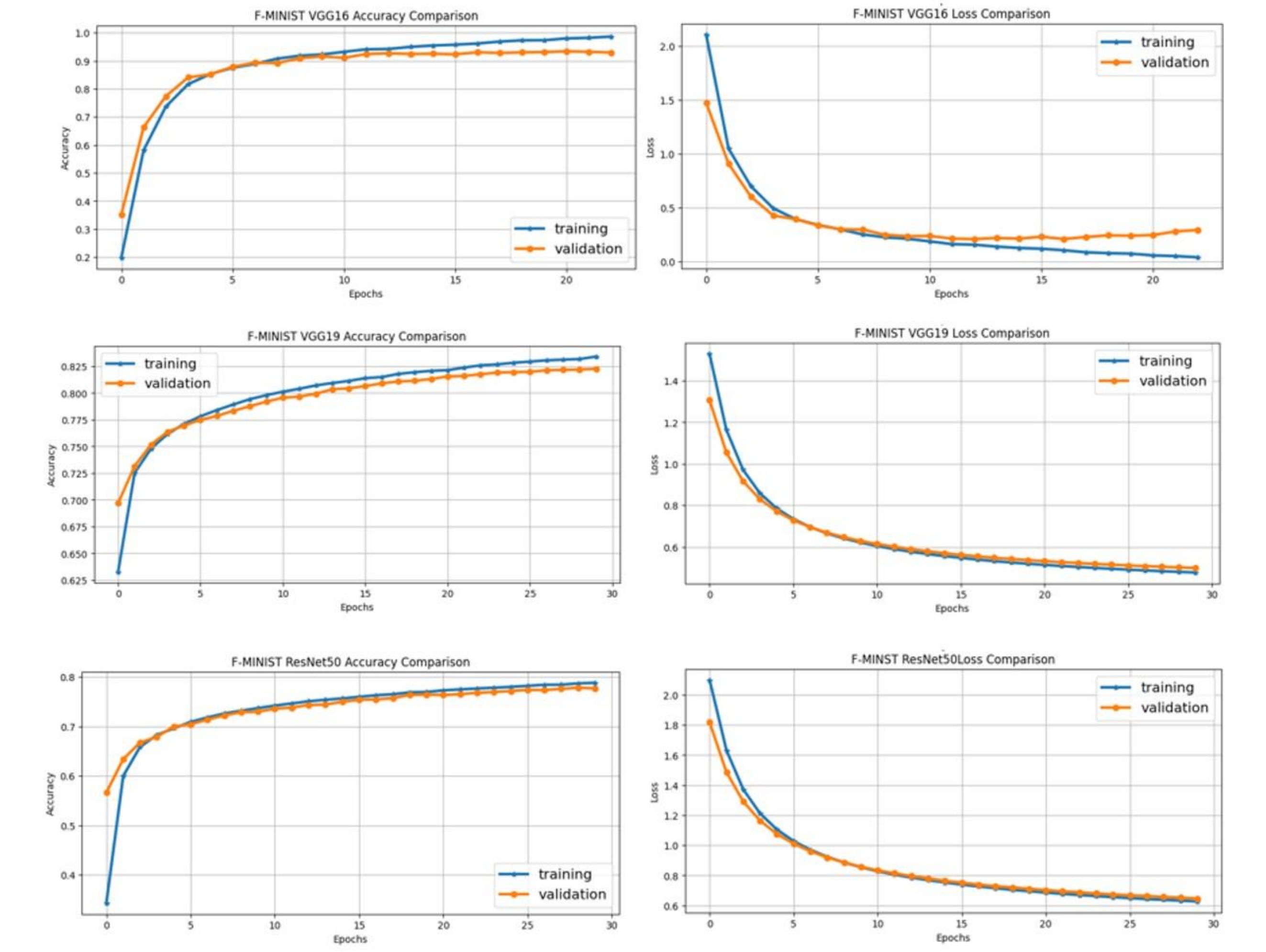


Fig 6 Shows Accuracy and Loss training and validation plots for each pre-trained model on the F-MNIST dataset

2. Handwritten Digits Dataset

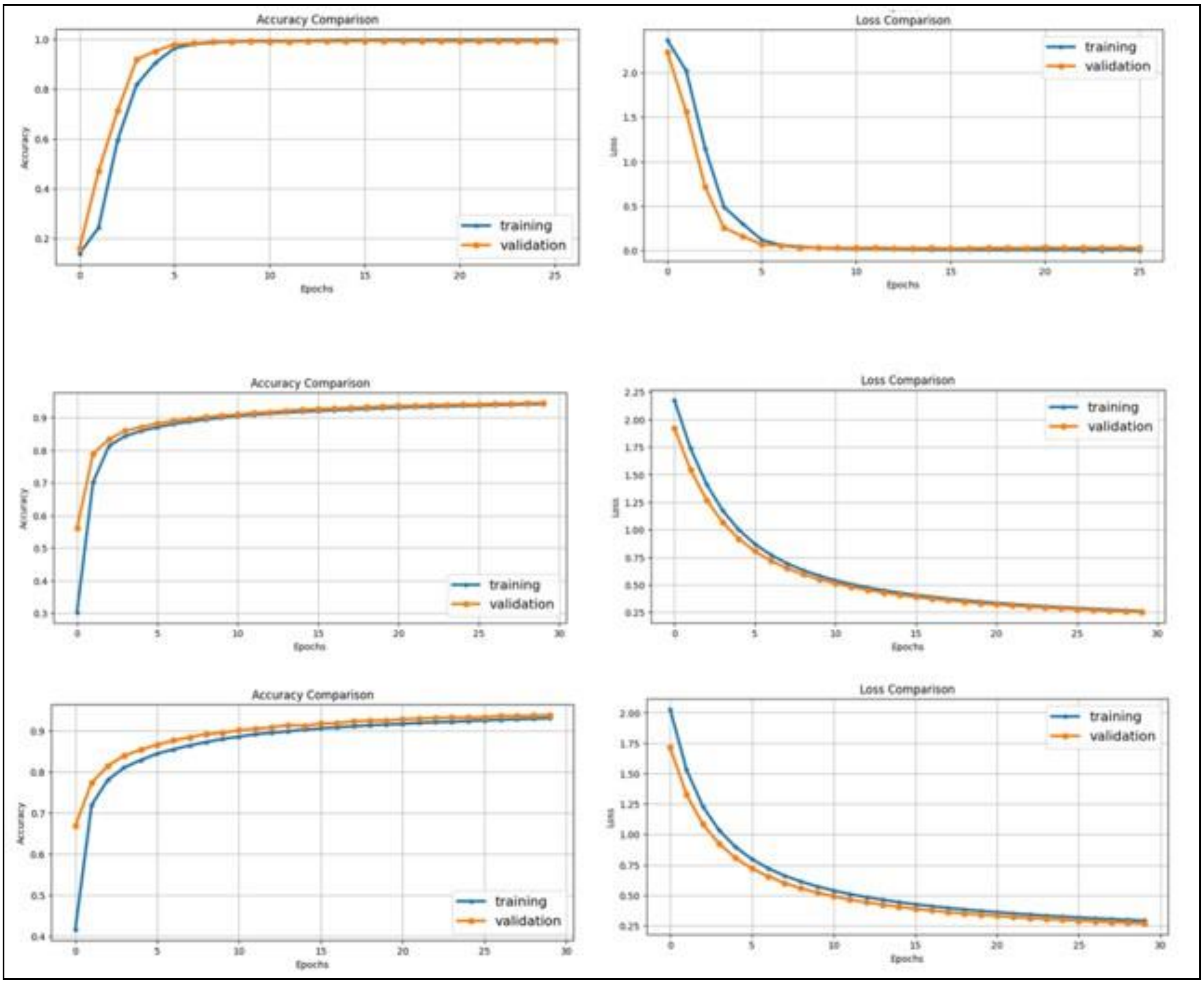


Fig 7 Shows Accuracy and Loss training and validation plots for each pre-trained model on the Handwritten dataset

3. Malaria Dataset

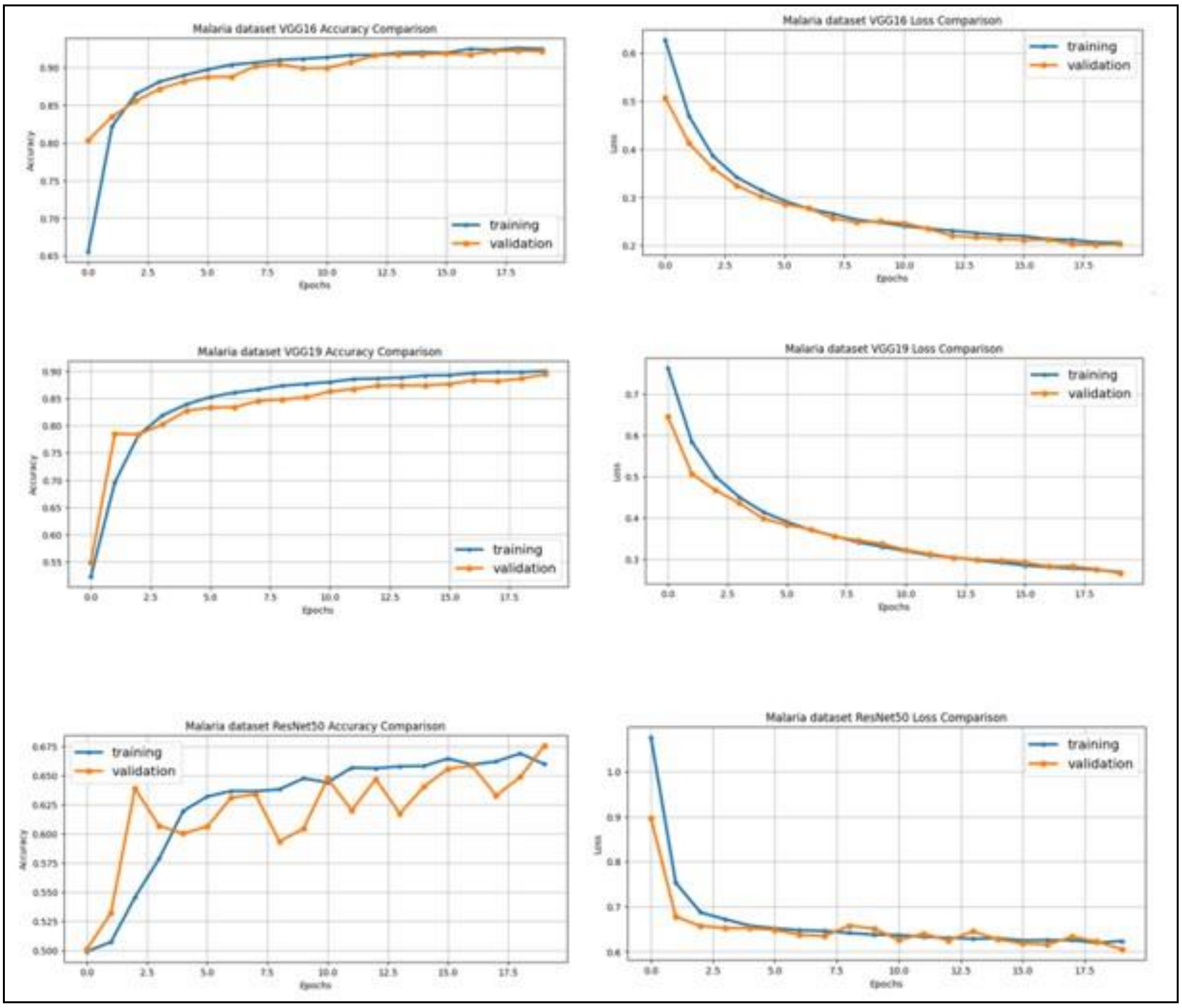


Fig 8. Shows Accuracy and Loss training and validation plots for each pre-trained model on the Malaria dataset

Table 1: F-MNIST Dataset

Pre-trained Models	Accuracy	Loss	Precision	Recall
VGG16	92%	0.2%	93%	91%
VGG19	82%	0.4 %	87%	77%
ResNet50	92%	0.2%	93%	91%

Table 2: Handwritten Dataset

Pre-trained Models	Accuracy	Loss	Precision	Recall
VGG16	99%	0.2%	99%	99%
VGG19	94%	0.2%	97%	90%
ResNet50	99%	0.2%	99%	99%

Table 3: Malaria Dataset

Pre-trained Models	Accuracy	Loss	Precision	Recall
VGG16	92%	0.2%	88%	97%
VGG19	92%	0.2%	88%	97%
ResNet50	67%	0.6%	65%	72%

The experiment above evaluates the performance of the three pre-trained models, and VGG16 has the overall best performance with an accuracy of 92% and loss of 0.2% on the F-Minist dataset, accuracy of 99% and loss of 0.2% on the Handwritten dataset and accuracy of 92% and loss of 0.2% on the Malaria dataset. While all the models successfully adapted to the imaging tasks, it draws parallel with AGI’s ability to generalise across diverse domains.

Discussion

The debate for and against the development of AGI will be a continuous issue. However, the future of AGI is dependent on how we can successfully control its use and ensure effective mitigation of risks.AGI can make the world a better place and improve various sectors rapidly in no time, equally, it can mean doom to the universe. This research states many challenges associated with ensuring an AGI, when it surpasses human intelligence, will continue to be beneficial to humans and develops a thorough framework for the effective control of AGI while addressing an array of technical, legal, ethical, and societal considerations.

In relation to an AGI being able to generalise across various tasks, the result from the experiment conducted successfully proves that the VGG16 generalises the best across the various datasets it was tested on, akin to how machines can learn from existing knowledge, before adapting to new scenarios (transfer learning). The capabilities of models like VGG16 to adapt and generalise is what needs to be controlled in AGI development.

Conclusion

This research has provided AI developers with a framework for responsible AGI development; it highlights the challenges and proposes a comprehensive framework for control. The computational aspect illustrates how an aspect of AGI can be applied in real-world scenarios with diverse datasets. Overall, it is crucial to integrate theoretical principles with computational insight to advance AGI ethically and responsibly.

Future work should be carried out on developing robust legal and regulatory frameworks adaptable to the evolution of AGI, addressing the accountability, liability and ethical implications.

References

Adams, C., Pente, P., Lemermeyer, G. and Rockwell, G. (2023) Ethical Principles for Artificial Intelligence in K-12 Education. *Computers and Education: Artificial Intelligence*, 100131.

Artificial Intelligence in Agriculture Jornal of Physics - Google Search. (n.d.) www.google.com. Available online: https://www.google.com/search?q=artificial+intelligence+in+agriculture+jornal+of+physics&oq=artificial+intelligence+in+agriculture+jornal+of+physics&aqs=chrome..69i57j33i10i160l2.21079j0j15&sourceid=chrome&ie=UTF-8 [Accessed 4 May 2023].

Baum, S. (2017) *A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy Global Catastrophic Risk Institute Global Catastrophic Risk Institute Working Paper 17-1*.

Blackman, R. (2020) A Practical Guide to Building Ethical AI. *Harvard Business Review*. [online] 15 Oct. Available online: https://hbr.org/2020/10/a-practical-guide-to-building-ethical-ai.

Bostrom, N. and Yudkowsky, E. (2011) *The Ethics of Artificial Intelligence*.

Chen, M., Mao, S., Yin Zhang and Leung, V.C.M. (2014) *Big Data Related Technologies, Challenges and Future Prospects*. Cham Springer International Publishing.

Claburn, T. (n.d.) *OpenAI CEO Heralds AGI No One in Their Right Mind Would Want*. www.theregister.com. Available online: https://www.theregister.com/2023/02/27/openai_ceo_agi/ [Accessed 2 Jan. 2024].

Dignum, V. (2021) The Role and Challenges of Education for Responsible AI. *London Review of Education*, 19(1).

Dormehl, L. (2016) *Thinking Machines : the inside Story of Artificial Intelligence and Our Race to Build the Future*. London: Wh Allen.

Dormehl, L. (2017) *Thinking Machines : the Quest for Artificial intelligence--and Where it's Taking Us next*. New York: Tarcherperigee.

Duettmann, A. (2017) *Artificial General Intelligence: Timeframes & Policy White Paper*. Available online: https://foresight.org/wp-content/uploads/2022/11/AGI-TimeframesPolicyWhitePaper.pdf [Accessed 7 Jan. 2024].

Everitt, T., Lea, G. and Hutter, M. (2018) *AGI Safety Literature Review * †*.

Goertzel, B. (2014) Artificial General Intelligence: Concept, State of the Art, and Future Prospects. *Journal of Artificial General Intelligence*, 5(1), 1–48.

Graham, R. (2021) Discourse analysis of academic debate of ethics for AGI. *AI & SOCIETY*.

Gruschka, N., Mavroeidis, V., Vishi, K. and Jensen, M. (2018) *Privacy Issues and Data Protection in Big Data: a Case Study Analysis under GDPR*. IEEE Xplore. Available online: http://ieeexplore.ieee.org/document/8622621/footnotes.

Haney, B.S. (2018) The Perils & Promises of Artificial General Intelligence. *SSRN Electronic Journal*.

Harris, T. and Raskin, A. (2023) *The Three Rules of Humane Tech*. www.humanetech.com. Available online: https://www.humanetech.com/podcast/the-three-rules-of-humane-tech.

Hossen, M.I., Fahad, N., Sarkar, M.R. and Rabbi, M.R. (2023) Artificial Intelligence in Agriculture: a Systematic Literature Review. *Turkish Journal of Computer and Mathematics Education (TURCOMAT),* [online] 14(1), 137–146. Available online: https://turcomat.org/index.php/turkbilmat/article/view/13384/9625 [Accessed 4 May 2023].

Independent Review of the Future of Compute: Final Report and Recommendations. (n.d.) GOV.UK. Available online: https://www.gov.uk/government/publications/future-of-compute-review/the-future-of-compute-report-of-the-review-of-independent-panel-of-experts.

Javaid, M., Haleem, A., Khan, I.H. and Suman, R. (2022) Understanding the Potential Applications of Artificial Intelligence in Agriculture Sector. *Advanced Agrochem*, [online] 2(1). Available online: https://www.sciencedirect.com/science/article/pii/S277323712200020X.

Lutkevich, B. (2022) *What Is Artificial General Intelligence (AGI)? - Definition from WhatIs.com*. SearchEnterpriseAI. Available online: https://www.techtarget.com/searchenterpriseai/definition/artificial-general-intelligence-AGI.

McLean, S., Read, G.J.M., Thompson, J., Baber, C., Stanton, N.A. and Salmon, P.M. (2021) The Risks Associated with Artificial General Intelligence: a Systematic Review. *Journal of Experimental & Theoretical Artificial Intelligence*, 35(5), 649–663.

Morris, M.R., Sohl-dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., Farabet, C. and Legg, S. (2023) *Levels of AGI: Operationalizing Progress on the Path to AGI*. arXiv.org. Available online: https://arxiv.org/abs/2311.02462#:~:text=We%20propose%20a%20framework%20for [Accessed 7 Nov. 2023].

Nair, S.R. (2020) A Review on Ethical Concerns in Big Data Management. *International Journal of Big Data Management*, 1(1), 8.

Naudé, W. and Dimitri, N. (2019) The Race for an Artificial General intelligence: Implications for Public Policy. *AI & SOCIETY*, 35(2), 367–379.

Neha, Gupta, P., Nadeem, D., Abuzar and Elahi, A. (2023) *Artificial Intelligence in Agriculture*. papers.ssrn.com. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4345592 [Accessed 4 May 2023].

Nguyen, A., Ngo, H.N., Hong, Y., Dang, B. and Nguyen, B.-P.T. (2022) Ethical Principles for Artificial Intelligence in Education. *Education and Information Technologies*.

Phillips, J.W. (2023) *Securing Liberal Democratic Control of AGI through UK Leadership*. James W. Phillips' Newsletter. Available online: https://jameswphillips.substack.com/p/securing-liberal-democratic-control [Accessed 11 Dec. 2023].

Roman_Yampolskiy, Ellen, R. and Wendt, K. von (n.d.) Limits to the Controllability of AGI. *www.lesswrong.com*. [online] Available online: https://www.lesswrong.com/posts/Zq4SxjwKBB6Ld3SNf/limits-to-the-controllability-of-agi.

Russell, S.J. (2019) *Human Compatible : Artificial Intelligence and the Problem of Control*. London: Allen Lane/Penguin Random House.

Salmi, J. (2022) A Democratic Way of Controlling Artificial General Intelligence. *AI & Society*.

Tan, A. (2020) *Singularity May Not Require AGI*. Medium. Available online: https://towardsdatascience.com/singularity-may-not-require-agi-3fae8378b2.

Valley, R. (2023) *The Promise and Peril of AGI: a Balancing Act for Humanity*. TechBomb

News. Available online: https://techbomb.ca/artificial-intelligence/agi-benefits-challenges-humanity [Accessed 6 Jan. 2024].

Williams, T. (2023) Impressed by Artificial intelligence? Experts Say AGI Is Coming next, and It Has ‘Existential’ Risks. *ABC News*. [online] 23 Mar. Available online: https://www.abc.net.au/news/2023-03-24/what-is-agi-artificial-general-intelligence-ai-experts-risks/102035132.

Zhao, L., Zhang, L., Wu, Z., Chen, Y., Dai, H., Yu, X., Liu, Z., Zhang, T., Hu, X., Jiang, X., Li, X., Zhu, D., Shen, D. and Liu, T. (2023) When Brain-inspired AI Meets AGI. *Meta-Radiology*, 100005–100005.