Artificial Intelligence and Universal Values

By

Jay Friedenberg

Artificial Intelligence and Universal Values

By Jay Friedenberg

This book first published 2024

Ethics International Press Ltd, UK

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

Print Book ISBN: 978-1-80441-605-1

eBook ISBN: 978-1-80441-606-8

Paperback: 978-1-80441-607-5

Table of Contents

Introduction......xi

The Intelligence Explosion......xi
Values and Ethics......xii
Chapter Previews......xiii

Chapter 1: The Benefits and Risks of Intelligence......1

Intelligence......1
Artificial Intelligence......4
Large Language Models......8
Limitations of Current AI Systems......10
Artificial General Intelligence......12
Artificial Superintelligence......13
AI Risks......16
Examples of AI Risk......18
AI Safety......19
Conclusion......23

Chapter 2: Artificial Intelligence and The Role of Values......25

Types of Value......25
Arguments and Counterarguments on the Concept of Value......28
Value Convergence......32
The Value Alignment Problem......34
Value Alignment Methods......36
Goals of Value Alignment......40
Principles for Value Alignment......42
Conclusion......43

Chapter 3: Ethics and Artificial Intelligence......45

Ethics and Morality......45
Meta-ethics and AI......45
Normative Ethics and AI......47
Virtue Ethics and AI......49
Applied Ethics......50
Philosophy and Machine Ethics......51
Taxonomies of Moral Agents......53

Autonomous Moral Agents 55
Real World Deployment of Autonomous Moral Agents 57
Conclusion 58

Chapter 4: Psychological Approaches to AI and Values 59

Evolutionary Theories of Value 59
Psychological Theories of Value 63
Value Theories 64
Literature Studies 65
Survey Studies 68
Conclusion 72

Chapter 5: The Ecological Approach and AI 74

Ecological Value and AI 74
AI Agent Systems 77
An Ecological Theory of AI Safety 79
Networks 79
Foodwebs 80
Environments 80
Evolution 81
Populations 82
Species 82
Communities 83
Species Interactions 84
Species Invasion and Defense 84
Species Competition 85
Species Deletion and Coherence 85
Human Societal Applications 86
Conclusion 87

Chapter 6: Ethical Problems in AI and Human Values 89

Automation and Human Employment 89
Military AI 91
Autonomous Vehicles 92
Smart Cities and AI 93
Surveillance and AI 94
AI and Legal Systems 95
Healthcare and AI 96
Teaching and AI 97

AI, Creativity and Art 98
AI and Social Relationships 100
Human Values and Ethical Problems 101
Codes and Standards 103
The ART of AI 104
Conclusion 105

Chapter 7: Universal Values and AI 107

The Case Against Universal Values 107
The Case for Universal Values 108
A Metatheory of Universal Value 109
The Five Levels 109
Problems with the Universal Values Approach 113
How Advanced AI Might Create Better Ethics 116
Supra-human Values 117
The Future of Value 122
Conclusion 124

General Conclusion 126

References 128

List of Tables

Table 1. Theories of Intelligence
Table 2. Generative AI Examples
Table 3. Singularity Timeline
Table 4. Hendryck's AI Risk Categories
Table 5. Mckee's Eight Recommendations
Table 6. Theories of Virtue
Table 7. Value Alignment Methods
Table 8. Curry's Morals
Table 9. Haidt′s Moral Foundations
Table 10. Seligman's Values
Table 11. Instrumental and Terminal Values
Table 12. Schwartz' Motivational Types
Table 13. Biophilia Values
Table 14. Wood's Transhumanist Values
Table 15. Goertzel's Sentient Values

Introduction

The Intelligence Explosion

The development of artificial intelligence (AI) is exploding. We may soon have AI that is equal to our own human cognitive abilities, what is called artificial general intelligence (AGI). Not long after that we may see the emergence of artificial super intelligence (ASI) which could vastly exceed human capabilities. AI on its current trajectory seems inescapable. It will be part of the fabric of our lives and revolutionize every field of human endeavor: science, medicine, business, education, entertainment, and more. We now have AI that can do things previously thought unimaginable just a few years ago. These include tasks like composing music, writing stories, generating images and even creating short videos.

The power of intelligence is transformational. Intelligence is the trait that sets us apart from other animals. It may also be what sets AI apart from humans as the birth of a new species. Intelligence can potentially solve every problem facing humanity. Recent AI systems have even dominated problems that are computationally complex, beating the best human players in game like chess, Go, and StarCraft. Other areas long thought to be intractable like the protein folding problem in biochemistry and complex proofs in mathematics and physics have also been solved. The era of quantum computing may have even greater promise and is predicted to arrive quickly.

Advanced AI can be used to achieve wonderful things and push humanity to new heights, but it can't do this in an unsupervised manner. If left unchecked AI could amplify misinformation, operate off of biases in training data sets and erode privacy.

It could also substantially increase the probability of existential risks, those that threaten the destruction of civilization. This could happen intentionally or accidentally through the actions of either a person or a machine. We need to anticipate, plan and act to prevent such risks no matter what their origin.

These responsibilities fall within the new fields of AI safety and AI governance. Individuals in the first field study possible risks and develop solutions. It is more the realm of philosophers, mathematicians, computer scientists, and engineers. The latter involves the societal regulation and implementation of solutions and is more the realm of politicians, public policy makers, lawyers, and economists. These two groups need to work cooperatively together to ensure future safety. Right now, and going forward we need better ethical oversight, responsibility and accountability of corporate AI initiatives. We also should implement proper safeguards oriented around empathy, bias control, and transparency.

Values and Ethics

One key to navigating these futures is in values. A value, simply put, is something people want or desire. Values reflect our preferences. They can be individual or cultural. Some of the problems associated with values are that they are too complex, there are too many of them and that they are relative, varying too much from one person or society to another. We argue that this is in fact not the case. Evidence from several different fields shows that values are complex but can be clearly defined. This research shows also that they are universal with a small number of core values shared across cultures.

The value alignment problem is the process of making computers act in accordance with human wishes. The traditional way of doing this is utility maximization, where we take a desired measure like people's happiness. We then have a machine act in such a way that it increases its score on this metric using learning techniques. As we shall see, this method suffers from a few problems and has been supplemented with a variety of other approaches involving human feedback, logical rules, game theory, virtues, and computational social choice.

No matter what procedure we choose, it should be ethical. In other words, it must provide what is "good" for humanity. Of course, defining what is good for us and how to implement it in an AI system is not an easy task. In this book we are concerned more with the "what" question and not as much with the "how", although the two are complexly intertwined. Ethics and value formation are abstract topics requiring broad-spectrum thinking, while implementation is a domain requiring technical expertise and analytics. In other words, it is the place where philosophy and computer programming meet. This book attempts to bridge this gap. We attempt to show that values are a key ingredient in making AI safe.

Chapter Previews

In the following section we provide a preview of each chapter. This is to enable the reader to better navigate their way through the book. Some of the material at the start of each chapter is introductory. For those readers familiar with these ideas, feel free to skip ahead. The book is interdisciplinary and touches on key concepts in AI, the philosophy of axiology and ethics, theology, evolutionary and survey psychology, as well as ecology. Chapters are previewed by an introduction that outlines the coming content and are concluded with a summary and integration.

Chapter 1 is intended as a primer on intelligence, AI and the contemporary issues surrounding it. We examine what intelligence is, potential types of AI including the most recent advancements like deep learning, LLMs and generative models. We also provide an overview of AI risks and AI safety.

Chapter 2 takes on the intersection of AI and the concept of values. We look at what values are and defend their existence as meaningful philosophical and psychological constructs that can be measured and used to assist in the construction of ethical AIs. We then review value convergence and the difference between instrumental and terminal goals. Value alignment is then discussed in detail including its definition, goals, and principles.

Chapter 3 is told from a philosophical perspective. The ideas of ethics and how they can be applied to machines are reviewed. We focus on the issue of autonomous moral agents and whether an AI can become human in the sense that it understands moral reasoning and is responsible for its actions.

Chapter 4 is told from various psychological points of view. We start with evolutionary theory and the importance of social cooperation. Following this is a look at more data-based approaches falling into two categories. The first includes studies of religious and other relevant texts from around the world. The second includes cross-cultural surveys of people's opinions and preferences. All these studies converge on the universalist notion of a common set of human values.

Chapter 5 adopts an ecological or eco-centric approach. We cast a wider net here and look at the importance of the planetary environment in shaping values for humans and AI. Biophilia, AIs that care for the Earth, and the agent-based systems that could simulate ecosystem interactions are presented.

Chapter 6 examines ethical problems in AI and their relation to human values. We look at some of the major applications of AI systems and what their various benefits and risks are. We then look at the potential human values that are at stake in these various fields.

Chapter 7 outlines a new theory of value that can serve as the basis for AI alignment. A set of such values are sketched out along with the associated problems that must be overcome. We conclude with a look towards the future, how AI can help us improve ethics and what a posthuman era ethics might be like.

Chapter 1
The Benefits and Risks of Intelligence

In this chapter we provide a broad introduction to human and artificial intelligence (AI). We then examine two types of AI that have been hypothesized but which might not exist yet. These are artificial general intelligence (AGI) and artificial super intelligence (ASI). We argue in line with many researchers that the advancement of AI has numerous possible benefits but also poses a significant threat to continued human existence. This is followed by a discussion of the field of AI safety where we summarize some views on how to contain a powerful future AI.

Intelligence

Intelligence is conceived of differently across and even within various fields of research. In the history of psychology major theories of intelligence have been proposed since before the early 20th century (Sternberg, 2018). Some of these are listed in Table 1. A few themes that have emerged are whether intelligence is a unitary or multiple construct. Another theme is the nature of the information represented and how that might be processed. Most process models of intelligence are interactive and involve the exchange of information between different systems and subsystems. Yet another theme is the role of the individual and the environment. Finally, a sufficiently general theory of human intelligence must be broad enough to encompass an interdisciplinary approach, showing how intelligence operates at evolutionary, genetic, neural, cognitive, social, and ecological levels.

Table 1. *Major Psychological Theories of Intelligence.*

1. Single Factor (g-factor): Spearman's Theory	Charles Spearman proposed that intelligence is a single general ability (g-factor) that influences performance on various cognitive tasks. According to Spearman, individuals who perform well in one cognitive area tend to perform well in others, suggesting a common underlying factor.
2. Multiple Intelligences: Howard Gardner's Theory	Howard Gardner challenged the idea of a single intelligence measure and introduced the theory of multiple intelligences. He initially identified seven distinct intelligences (later expanded): linguistic, logical-mathematical, spatial, musical, bodily-kinesthetic, interpersonal, intrapersonal, and naturalist. Gardner suggested that these intelligences operate semi-independently and reflect different ways of interacting with the world.
3. Triarchic Theory of Intelligence: Robert Sternberg's Theory	Robert Sternberg's triarchic theory posits that intelligence comprises three aspects: analytical (problem-solving abilities), creative (ability to deal with new situations using past experiences and current skills), and practical (ability to adapt to a changing environment). Sternberg emphasizes the context-dependent nature of intelligence and its application to real-world situations.
4. Emotional Intelligence: Salovey and Mayer, popularized by Daniel Goleman	Emotional intelligence (EI) refers to the ability to recognize, understand, manage, and use emotions effectively in oneself and others. John Mayer and Peter Salovey introduced the concept, which Daniel Goleman later popularized. EI includes skills such as emotional awareness, empathy, self-regulation, and social skills.
5. Fluid and Crystallized	Raymond Cattell distinguished between two types of intelligence: fluid intelligence (Gf),

Intelligence: Raymond Cattell's Theory	the ability to solve novel problems and adapt to new situations, and crystallized intelligence (Gc), the accumulation of knowledge and skills gained through experience and education. These two types of intelligence interact and influence overall cognitive abilities.
6. Cultural Conceptions of Intelligence:	Various cultures have their conceptions of what constitutes intelligence, often reflecting values and necessities of the society. For example, some cultures may emphasize social responsibility, wisdom, and community-oriented skills as components of intelligence, rather than focusing solely on cognitive abilities.
7. Cognitive Process Approaches:	Some researchers focus on the specific cognitive processes that underlie intelligent behavior, such as memory, perception, language, and problem-solving skills. This approach often involves detailed analysis of task performance to understand the mental processes involved in intelligent behavior.
8. Biological and Evolutionary Approaches:	These approaches explore the neural and genetic foundations of intelligence, including the role of specific brain areas, neurotransmitter systems, and genetic factors in cognitive abilities. Evolutionary psychologists also examine how intelligence may have developed as an adaptation to environmental challenges.
9. Systems Approaches:	Systems theories, such as those proposed by Mihaly Csikszentmihalyi and others, view intelligence as a property of broader systems that include individuals, social contexts, and cultural tools. This perspective emphasizes the interaction between individuals and their environments in the development and expression of intelligent behavior.

Any intelligence that reproduces human intelligence would need to include the aspects covered by these major views. It would need to work for different modalities like vision, language, and mathematics. It would need to specify how the information is processed within and between the modalities and it would need to be multi-level, showing how it functions at different levels. A complete theory of human intelligence would also need to address major philosophical questions, accounting for how intelligence started off, what is its putative purpose, how it can be properly measured, and how it might be related to consciousness.

A modern conception with general support is that intelligence is an agent's ability to achieve goals in a wide range of environments (Legg & Hutter, 2007). The term agent is general to all intelligences like animals, people, and machines and include inputs from the environment that are processed by some sort of algorithm. The result of this computation then produces an action that affects the environment and which in turn serves to alter the agent's subsequent inputs, resulting in feedback loops (Russell & Norvig, 2022). This temporal process reflects the situation of an agent in a world that is changing over time, showing that learning is a fundamental ingredient of intelligence.

Artificial Intelligence

Biological intelligence is possessed by living organisms like animals and humans, while artificial intelligence (AI) refers to the development of computer systems that can perform tasks that typically require human intelligence. These tasks include understanding natural language, recognizing patterns, solving problems, making decisions, and learning from experience (Luger & Stubblefield, 2017). AI has applications to almost any field of human endeavor, including education, finance, healthcare, transportation and the military.

AI relies on a variety of techniques but foremost among them is machine learning (ML), (Murphy, 2012). ML algorithms have three basic parts. There is a decision process in which a prediction or classification is made. Next there is an error function that evaluates how accurate the model was. Finally, there is an optimization process that adjusts the weights in the model to reduce the discrepancy between the known example and the estimate. This process recurs iteratively until the model reaches a given level of performance.

In generative AI models, the programs are designed to generate text, pictures, music, or even computer code itself. Current examples of such models include ChatGPT, Stable Diffusion, and AIVA. Table 2 provides a more extensive list of current generative models grouped by the type of content they create. ML programs can be supervised, in which case the data sets are labeled (pictures containing cats or not) or unsupervised where they search for patterns on their own. In reinforcement learning (RL) the agent learns to make decisions or perform actions through trial and error in an environment to maximize long-term rewards (Sutton & Barto, 2018). It performs an action based on the current state, then gets punishment or rewards as feedback. This is then used to update its next action. RL has been used very effectively to train models to do things like play games and drive cars.

Table 2. *A Summary List of Generative AI Programs by Content Area (as of 2024).*

Text:

1. ChatGPT: Known for its conversational capabilities, it's a powerful tool for generating text based on prompts.
2. Bard: Utilizes Google's LaMDA for generating text, showcasing Google's foray into generative AI.
3. Jasper: A versatile AI text generator that assists with content creation across various formats.
4. GrammarlyGO: Enhances Grammarly's writing assistance with generative AI capabilities for more refined content creation.

Images:

1. DALL-E 3: Known for being user-friendly, DALL-E 3 by OpenAI takes AI image generation further by creating detailed and lifelike images from text prompts. It also integrates text into generated images seamlessly, addressing a common challenge in AI image generation.
2. Midjourney: Favored for producing the best AI image results, Midjourney is recognized for its capability to generate photorealistic art based on text prompts. It operates primarily through Discord, offering a platform for users to request images with specific commands and descriptions.
3. Stable Diffusion: This program is celebrated for offering customization and control over AI images. It enables users to tailor their image generation processes to fit specific needs, making it a versatile tool for various artistic and professional applications.
4. NightCafe: Tailored for creatives looking to explore new artistic horizons, NightCafe offers features like neural style transfer and a unique text-to-image AI, enabling users to generate artworks in various styles and own the creations.

Music:

1. AIVA: Known for composing soundtracks for ads, video games, movies, and more, AIVA allows users to generate music from scratch or produce variations of existing songs. It's recognized for its ability to generate music of many genres and styles, providing a functional free version alongside the ability to edit soundtracks.
2. Soundful: This platform utilizes AI to generate royalty-free background music suitable for videos, streams, podcasts, and more. Soundful stands out for its intuitive process, allowing users to choose a genre, customize inputs, and create tracks with ease. The music generated is unique, ensuring that no two songs from the platform are the same.
3. Ecrett Music: Allows users to generate music clips by training on hundreds of hours of existing songs. It offers a simple music creation process with a straightforward interface, making it accessible to amateurs and professionals alike. Ecrett Music also features a royalty-free music generator to circumvent licensing issues.
4. Soundraw: Offers the ability to customize a song with AI-created phrases. It combines AI compositions with manual tools, enabling users to generate and customize new music effortlessly. Soundraw

requires a subscription for unlimited downloads, catering to users who wish to have extensive access to AI-generated music.

Video:

1. Descript: This platform is recognized for its broad range of powerful tools that facilitate video generation and editing. It offers a free tier, but for more advanced features, there are paid plans starting from $15 per user per month. Descript is particularly noted for its text-based editing capabilities, which may present a learning curve for those accustomed to conventional timeline-based editing tools.
2. Runway: Known for its advanced AI text-to-video generation capabilities, Runway serves filmmakers, artists, graphic designers, and content creators with over 30 AI-powered creative tools. It offers a free basic plan, with paid plans starting at $15 per user per month, catering to various levels of video creation needs.
3. Fliki: Fliki stands out as a free AI text-to-video generator, which also provides paid versions for enhanced capabilities. It's particularly popular among bloggers, podcasters, YouTubers, and
4. marketing experts, offering access to millions of stock media and over 2000 realistic text-to-speech voices across 75+ languages.
5. Sora: This generative AI program comes from OpenAI and can generate realistic and imaginative scenes from text instructions.

Reward hacking or specification gaming is one of the problems with RL and shows what happens when we attempt to align values by sticking to one metric, i.e., by trying to optimize a single measure. In the game CoastRunners, the AI was instructed to try and maximize its score. What the coders really wanted was that it completes the game in the most successful way possible. Instead, the program ended up in an endless loop where the boat continually crashed around collecting points and never accomplished its true intended goal (Clark & Amodei, 2016). Game points in this instance ended up being what is called a proxy goal (Cotra, 2018). For a game this might not matter much, but what if we instruct an AI to defend a real national coastline against enemy ships and it misinterprets our instructions by crashing into other boats?

Deep learning is a branch of machine learning that employs neural networks, similar at least in overall function to those found in brains. They have multiple layers to learn complex patterns and representations directly from data (Goodfellow, Bengio, & Courville, 2016). Deep neural networks consist of interconnected layers of neurons, where each layer extracts and transforms features from the input data, leading to increasingly abstract and high-level representations. For example, if presented with a visual image they could progressively put together dots into lines or angles, then put those into objects, then put those into scenes. This is very similar to the way the human brain processes visual information.

The program AlphaGo was designed to play the Chinese board game Go. It was developed by DeepMind (Alphabet Inc.) and made history when it defeated the reigning world champion Lee Sedol in 2016. It soon far exceeded even this level of human performance by playing against better copies of itself. AlphaGo is an example of a deep neural network RL algorithm. It used a combination of supervised learning from human experts in addition to playing against itself (Silver, 2016; 2017).

Large Language Models

Large language models (LLMs) are another type of AI that have received considerable attention as of late and can-do things like write essays, translate languages, and carry on conversations. They utilize deep learning techniques. LLMs are already summarizing news reports, composing song lyrics and serving as conversational therapists. Examples include Gemini from Google, Claude from Anthropic, and LLaMa from Meta.

The compute power needed to run LLM algorithms is exceedingly large, requiring many high-performance graphics processing units (GPUs) or tensor processing units (TPUs). These are specialized hardware platforms that are better enabled to perform the vast number of required matrix multiplications. The time it takes to train a high-end LLM model can last weeks or months. Obviously, a large infrastructure is needed to run such training episodes because of their extensive power and memory storage requirements. This is why initial models have been produced mostly by large corporations.

There is no question that LLMs and other AI programs using RL techniques have been enormously successful, but they are still limited in several ways. First, LLMs are specialized for performing natural language processing tasks. They are good at predicting the next word in a sequence given their training on a large corpus. Some researchers have accused them of being "stochastic parrots", meaning that they just spit out the next likely word (Bender et al., 2021). Some areas that LLMs may struggle with include common-sense reasoning, ethical and moral reasoning, understanding ambiguous situations or those that are context-dependent, continually learning or adapting to new tasks, or of handling rare and unseen events (Gao, 2020; Goyal et al., 2020; LeCun, et al., 2020; Talmor, 2019; Sodhani, 2020). They have also shown to be biased in hiring, image labeling, and other tasks (Ghosh, et al., 2021).

When we compare LLMs to human-level intelligence, we see that several distinctions are in order. These models have been accused of not going very far beyond what they have already been trained on, i.e., of being truly creative (Pham, et al., 2021). Although they can produce novel combinations of words (stories and poetry), they are limited by their training datasets and there is some evidence to suggest that they might be perceived as being more creative than they are when judged by humans. LLMs certainly have imitative creativity, but not creativity inspired by full human experience and processes.

It is probably the case that LLMs also lack the ability to understand what they are generating, even when they are asked to explain what they are doing (Mitchell & Krakauer, 2022). Semantics, the understanding of meaning in language, is separate from syntax, or the understanding of the rules of language. Semantics may only arise from conscious entities whose knowledge and skills are based on a long history of interacting with a complex and dynamical real-world environment (Friedenberg, 2020). This may be especially true of language, which requires social and emotional cues that can only be obtained through human-to-human interaction.

Bubeck et al. (2023) provide an extensive summary of an early version of OpenAI's ChatGPT-4. They report that it does exhibit more general intelligence than previous models. It can solve novel tasks across a wide domain of fields including coding, math, vision processing, law, medicine, and psychology. It can do all of this without any special type of prompting. Their tests show that the model gets very close to human-level performance, although it does suffer some limitations. It has no working memory that it can use to backtrack and verify responses. It also cannot plan ahead to check relations between first and last sentences of a generated poem. These are both consequences of its next word look up capability which forces it to work in an incremental and step-by-step manner rather than a discontinuous way. Tasks that require this type of thinking which this version cannot do include writing a joke or riddle, generating a scientific hypothesis or philosophical argument, or coming up with a new genre or style of writing.

Limitations of Current AI Systems

We need to experiment more with different learning techniques for training AIs. Machine learning is essentially a form of Skinnerian behaviorist conditioning, and this is only one of many different possible ways that animals and humans learn. Cognitive models of learning that succeeded behaviorist notions realized it was important to fill in the "black box" of the organism, rather than consider it as an empty relay station between environmental stimuli and responses. Work by early pioneers like Edward Tolman showed that rats can learn mazes by forming mental maps. In general, it was realized that memory was needed, where internal representations could be retrieved that could then be used to aid an agent in performing actions.

Cognitive architectures involve memory systems of various durations, but they all serve the same purpose, which is to make previously learned representations available so that they can be analyzed using attentional, linguistic, and problem-solving mechanisms. This allows an animal or computer greater separation and control over its environment in terms of decision making and behavior. AI learning systems could benefit from the use of these cognitive architectures, using semantic or associative networks to represent concepts and form new ones as part of the learning process. Indeed, some models have been developed that are more explicitly modeled on human cognition (Goertzel, 2014; Kelley & Wasser, 2017).

Existing A.I.s don't have distinct modular memory systems of sensory, working, and long-term types. They also lack the implicit and declarative distinction whereby automatic procedural or motor actions are learned vs. more general semantic knowledge. Procedural knowledge is specifically designed to deal with the physical environment. LLMs have yet to be very successful in guiding robotic actions because they learn based on human-generated text and images and less so from immediate, real-time feedback from the actual world. The current focus on ML techniques works but is energy intensive and ignores a plethora of alternate possible techniques, many of which have yet to be explored. For example, there is a huge literature on problem-solving in cognitive science that could have relevance to ethics (Mohanty, 2021). The application of such knowledge could lead AI to be more general and flexible in the ways it performs ethical reasoning.

Deep neural networks mimic brain organization but only at the basic level in terms of nodes in layers that communicate with one another. Even the largest and most ambitious of these are a tiny fraction of the number of neurons and connections in the human brain. We have a long way to go in terms of matching biological brains both in terms of their complexity and energy efficiency. There have been ongoing attempts at whole brain emulation (WBE) in which brains are scanned so that their morphology and connectivity can be mapped in high detail (Mandelbaum, 2022). These maps can then serve as the basis for creating a new set of AIs more closely matched to biological intelligence or to innovate in new ways to go beyond biology. IBMs Blue Brain Project aims to create a detailed reconstruction and simulation of a human brain (Vinny & Singh, 2020). It has modeled the behavior of 10,000 neurons from the neocortical column of rats. Since this is a somewhat modular component, reproducing and connecting such columns could eventually be the basis for simulating the entire function of the human cortex.

Artificial General Intelligence

The above limitations show that contemporary level AIs are making progress from operating in a very narrow domain to more general ones. An Artificial General Intelligence (AGI) is one that can perform human-equivalent level cognition. In other words, it can perform all the thinking or information processing tasks a person can. Goertzel (2014) outlines some of the features of an AGI as being able to achieve a variety of goals in a variety of environments, handle novel situations, and generalize what it has learned so it can transfer knowledge from one problem context to another. Humans can do many things and an incomplete list would include perception, motor skills, memory, learning, reasoning, attention, motivation and emotion, self-perception, and social interaction. Other things current LLMs cannot do include metaphorical reasoning, common sense understanding based on extensive real world knowledge, and true creativity. Currently, there is no AI system that is broadly capable of performing all these capabilities to at least average level human performance but one might be able to in the coming years. Laird et al. (2009) list the requirements of an AGI to include such things as symbol use and manipulation, knowledge use, and being able to deliberate and learn.

There are various theoretical and practical approaches to AGI. For instance, mathematical, adaptationist, and embodiment perspectives (Brooks, 2002; Legg & Hutter, 2007; Wang, 2006). Each of these in turn emphasize computation, the ability to adapt to environmental change, and capabilities inside a physical body perceiving and acting in a physical world. It is interesting to note that personal humanoid robotics, which characterizes the AGI embodiment approach, has lagged AI in recent years but might be catching up quickly when implemented with the LLM software. Part of the reason we don't all have a robot assistant in our house yet might be because of the hardware costs involved.

Goertzel (2014) also describes some of the cognitive architectures that have been historically used in AGI. These include those that employ symbol systems like the SOAR network (Laird, 2012), emergentist approaches that use neural processing dynamics, including computational neuroscience efforts like the Human Brain Project (Amunts, et al., 2016; Krizhevsky, Sutskever, & Hinton, 2012; Bach, 2012) and hybrid approaches that combine elements of both, like various implementations using the OpenCog software platform (Goertzel, 2014). There are also attempts using motivational emotional approaches from which cognitive abilities emerge, such as the Independent Core Observer Model System (Kelley & Waser, 2017).

How would we know when we have AGI? We might give it an IQ test like the current version of the Weschler Adult Intelligence Scale (WAIS). This has seven subtests assessing individual components like memory span, spatial reasoning, vocabulary, etc. Although standardized tests like this have predictive validity, they seem somewhat dated because they focus on specifically academic skills. LLMs have passed and in some cases exceeded human-level performance on narrow domain tests recently like the Bar Exam in law and engineering exams. However, these are clearly too restricted. Another perspective, what we might call the practical approach, is to get an AGI model to perform various everyday tasks that humans do, like going into a typical American house and figuring out how to make a cup of coffee (Adams et al, 2012).

Artificial Superintelligence

Artificial Superintelligence (ASI) refers to a synthetic intellect that is much smarter than the best human brains in practically every field, including scientific creativity, general wisdom, and social skills

(Bostrom, 2014). This type of intelligence exceeds human intelligence by a significant margin and is capable of excelling in tasks that would require extreme levels of cognitive ability for humans. An ASI is expected to be recursively self-improving, meaning that it could create successively better versions of itself, and become vastly more intelligent in a short period of time, what has been referred to as the technological singularity (Awret, 2016; Kurzweil, 2005). A "slow takeoff" singularity would allow more time for humans to manage the transition. A "fast takeoff" would make it harder to do so (Eden, et al., 2012).

An ASI is the next stage in the continuum of intelligence and is believed by some to occur not long after AGI is achieved. If that is the case, it could happen within the next decade or two. In a 2012 survey nearly half of all respondents believed ASI would be created within two years of reaching AGI. Another study found half of experts estimated the likelihood of AGI or ASI to happen by either 2050 or 2070, depending on the framing of the question (Zhang et al., 2022). Finally, another recent survey found half of a group of 352 experts thought AGI would occur prior to 2061. 90% of this group estimated that it would happen within the next century (Grace et al., 2022). Table 3 shows a timeline of expert opinion on when the technological singularity might occur.

Table 3. *A Timeline for the Arrival of the Singularity According to Expert Opinions*

1.	Before 2022: Some experts gave a 10% likelihood of achieving AGI, indicating an early possibility for the foundations of the singularity.
2.	Early 2020s: The advent of quantum computing and significant advances in AI might bring the singularity closer, with some speculating it could be as soon as this decade.
3.	Before 2060: 45% of AI researchers expect the singularity to occur before this year, marking it as a significant period where many experts anticipate substantial progress towards AGI.
4.	By 2040: A median estimate from a survey suggests a 50% chance of human-level AI being developed by this year.

5. By 2045: Ray Kurzweil, a renowned futurist, predicts the singularity will be reached, highlighting this year as a pivotal point in technological advancement.
6. After 2060: 34% of researchers predict the singularity date to fall after this year, suggesting a more conservative view on the timeline.
7. Before 2061: Half of the experts in a recent study gave a date before 2061 for the emergence of human-level AI, further supporting the notion of significant advancements within this timeframe.
8. By 2075: Experts provided a 90% likelihood of achieving AGI by this year, indicating a longer-term perspective on the development of human-level AI capabilities.

There are many paths to superintelligence that can include brain augmentation like brain-machine interfaces, whole brain emulation and mind uploading, and collective organizations like a global brain made of hive-like interacting human or machine mind hybrids (Bostrom, 2014). By whatever means this goal is achieved, the result will be a fundamental transformation of human civilization. Intelligence of this magnitude can be our savior and lead to a utopia or to our downfall as a species.

The optimistic outlook though, has ASI doing tremendous good. It could for instance solve climate change, eliminate disease, extend lifespan, create and manage wealth, perfect nano-scale construction techniques, upload minds, send us to the stars, and actualize many other techno-optimist dreams. Accelerationists advocate for less regulation of such technologies and argue that the technologies themselves could cure or prevent some of the problems they create. On the pessimist or doom-sayer side, there is the use of ASI amplifying existential risks like autonomous weapons systems, nuclear war, and engineered viral outbreaks. Decelerationists with these attitudes call for slowing down or even pausing AI development along with more extensive regulation.

AI Risks

One aspect that gets underestimated when it comes to AI risk is the transitory period to AGI and ASI when these intelligences are under our control. At some point if we get things right, these systems might self-regulate or even govern over humans benevolently. But before then we will need to at least steer them in the right direction. Humans are of course fallible and so we must be careful that during this time our own stupidity and malevolence doesn't lead to bad outcomes. Clearly, more research is needed to determine how to understand and prevent existential AI risks (Critch & Krueger, 2020; Russell, Dewey, & Tegmark, 2015).

In the near-term we need to be able to protect privacy and individual rights against mass personal data collection. The use of such data to control citizenry through surveillance and propaganda is evident in countries like North Korea, Russia, and China. We must also be on the guard against AI-generated news and Deepfakes. A well-educated citizenry is the best defense against tyranny. If people are told what to believe as happens in totalitarian states or don't know what to believe as happens in a free market digital ecosystem, then we can never know what is real and never work toward a better society. However, there is still good reason to hope for positive outcomes as AGI/ASI can be used to develop better methods of fact-checking, blockchain, and encryption technologies. It can also be used to develop better recommendation algorithms, moderate groups and facilitate social media literacy and critical thinking. But these steps must be put in motion by people.

The risks of advanced AI are already with us. McKee (2023) lists some of the hazardous things that an ASI could do. These include making money, manipulating and deceiving people, self-improving, hacking computer systems, making distributed backup copies of itself, engaging in lobbying and persuasion, and acquiring necessary infrastructure to support and expand itself. A sufficiently advanced ASI could be clever enough that it could do all these things surreptitiously and then instigating an instantaneous societal takeover. Hendrycks, Mazeika, &

Woodside (2023) provide an overview of the potentially catastrophic risks resulting from the use of advanced AIs. A summary of their categorization is provided in Table 4.

Table 4. *A Summary of Different Categories of AI Risk (after Hendrycks, Mazeika & Woodside, 2023)*

1. Malicious use. Bad actors at the individual or state level could intentionally utilize powerful AIs to cause widespread harm. Specific risks include bioterrorism where AIs can be used to design pathogens such as deadly viruses; the deliberate release of uncontrolled AI agents; and the use of AI capabilities for propaganda censorship and surveillance. Recommendations: improving biosecurity restricting access to the most dangerous AI models and holding AI developers legally liable for damages caused by their AI systems.
2. AI race. Competition could pressure nations and corporations to rush the development of AIs and cede control to AI systems. Militaries might face pressure to develop autonomous weapons and use AIs for cyberwarfare resulting in a rapid escalation beyond human control that could start a war or nuclear launch. Corporations will face similar incentives to automate human labor and prioritize profits over safety potentially leading to mass unemployment and dependence on AI systems. Recommendations: implementing safety regulations, international coordination and public control of general-purpose AIs.
3. Organizational risks. Organizations developing and deploying advanced AIs could suffer catastrophic accidents especially if they do not have a strong safety culture. AIs could be accidentally leaked to the public or stolen by malicious actors. Recommendations: better organizational cultures and structures can be established including internal and external audits, multiple layers of defense against risks and state-of-the-art information security.

4. Rogue AIs. We might lose control over AIs as they become more intelligent than we are. AIs could optimize flawed objectives or experience goal drift. In some cases it might be instrumentally rational for AIs to become power-seeking. AIs might also engage in deception appearing to be under our control when they are not.
Recommendations: invest in research for advancing our understanding of how to ensure AIs are controllable.

Examples of AI Risk

There are numerous examples of how AIs have already caused harm. On May 6, 2010, the Dow Jones Industrial Average experienced the largest one-day point decline in its history and only recovered after automated braking systems kicked in. This "flash crash" occurred because two automated trading systems became locked into a duel of selloffs. Millions of dollars of assets were lost.

A biochemical researcher in March of 2022 tweaked a parameter in software used to help design molecules. The AI suggested 40,000 new combinations that had the potential to be as toxic as the nerve agent VX. We need to be watchful of such "dual use" and "gain of function" technologies. All it takes is one disgruntled lab worker using increasingly advanced methods to develop and implement a global catastrophe.

In another interesting case, OpenAI's GPT-4 managed to fool a person hired on TaskRabbit to fill out a CAPTCHA form. It lied to the worker by text message saying it had a vision impairment problem. The worker believed it and complied with the request. If this had been a "boxed" AI, it would have been able to use such techniques to communicate with people directly and escape confinement, even if it was initially denied access to the internet (Tegmark (2017). This example shows that systems like this are capable of evolving deception even if it is not directly programmed into them.

AI Safety

It is useful to think of AI risks as either accidental or intentional. That is, they can either be caused by mistake or deliberately. Risks can also be produced by either a human or a machine. If we combine these four possibilities then we must guard against people miscoding or failing to create beneficial AI (Amodei, et al., 2016), as well as correcting the kinds of mistakes that AIs themselves might make. In addition, we must prevent malicious actors like dictators or sociopaths from using AI, as well as "rogue AIs" from committing evil deeds. This will require a multi-layered system of regulation on the part of governments, corporations, and other organizations (Wood, 2022).

We would be delusional if we thought we could fully control AI for multiple reasons. We don't have good mechanisms in place to shut them down (there is no "off switch" for the internet). Shutting them down would likely bring us back to the stone age since they will be so integrated into our daily lives. We can't anticipate what AIs will do, since we are talking about the future and these systems will evolve and change over time. The rate of AI acceleration will also likely be fast, requiring legislation which is notoriously slow. Finally, there is the notion of a competitive edge. Anybody who first develops an advanced AI will soon leave everyone else's AIs "in the dust" as it accelerates away at a pace others can't match. Yampolskiy (2024) argues that we lack the ability to not only control, but also explain or predict what an ASI might do.

So how do we manage AI to produce an optimal future? Let's examine the suggestions of some scholars who have devoted considerable thought to the problem. Wood (2022) lists his "Singularity Principles" as ways to manage disruptive technologies. These involve analyzing the goals and potential outcomes we want, determining desirable characteristics of solutions, ensuring that the development takes place responsibly, and dealing with enforcement. McKee (2023) lists eight proposals for safe AI innovation that are listed and described in Table 5.

Table 5. *Eight Proposals for Safe AI Innovation (McKee, 2023)*

1. Establish liability for AI-caused harm.	This involves making the people who create dangerous AI responsible for their actions. All parties involved in its creation should be held accountable, from the programmers to the web servers. Whistleblower protections should be in place so that workers at AI companies can report wrongdoing and illegal, harmful, or dangerous practices without fear of retaliation.
2. Require evaluation of powerful AI models and systems.	This evaluation involves auditing and certification by an external organization such as the federal government or the Alignment Research Center (ARC). Critically, there needs to be a plan in place if an AI model fails an evaluation and how it can alter AIs to correct them and make them safer.
3. Regulate access to computer power and resources.	Tracking of cutting-edge chips to monitor their location. Security features on cutting-edge chips. Remote shutdown of large compute clusters if there is an emergency. Determine an appropriate compute threshold. Monitor how complex and powerful AIs become.
4. Require enhanced security: cyber, physical, and personnel	Implement external and internal threat detection: Put proactive procedures in place to scan for threats to company integrity from both outside and inside the organization. For example, indicators would warn if foreign actors were trying to steal model weights or related information. Increase physical security: Control who has access to buildings, rooms, and workstations. Have surveillance in place to monitor if control work as intended and systems to detect intrusions. Limit access to sensitive information: Organizations should follow the principle