The Millennium Project

International Governance Issues of the Transition from Artificial Narrow Intelligence to

Artificial General Intelligence

Report of Phase 1

© 2023 The Millennium Project

Introduction

For the purpose of this study, Artificial General Intelligence (AGI) is defined as a general-purpose AI that can learn, edit its code, and act autonomously to address novel and complex problems with novel and complex strategies that are similar to or better than what humans could do. This is distinct from Artificial Narrow Intelligence (ANI) that exists today and has more narrow purposes. Artificial Super Intelligence is AGI that has become independent of humans, developing its own purposes, goals, and strategies without human understanding, awareness, or control and continually increases its intelligence and scope of action beyond humanity as-a-whole.

Futurists explore a range of possible futures, their trajectories, potential consequences, and effectiveness of actions to address these futures. Much work has been done on ANI, but less so on AGI. If we don't have an international regulatory system of rules, guardrails, continuous audits and the like in place before AGI arrives, then ASI could emerge, not to our liking. Since AGI could arrive in less than ten years, it is wise to internationally explore potential AGI governance issues and models now. This could be the most difficult management question humanity has ever faced.

After reviews of professional articles, conference proceedings, and online sources, The Millennium Project's AGI Steering Committee created 22 questions (listed in Appendix B) and identified 55 AGI experts and thought leaders (listed in Appendix A). These experts from the U.S., China, UK, Canada, EU, and Russia were invited to address only those questions they preferred to address. Some experts were interviewed, some submitted written responses, and the views of others were collected from their on-line statements. This report shares these views, organized by the 22 questions.

The analysis of Phase 1 will be used for an international assessment (Real-Time Delphi - RTD) of the most important unresolved questions regarding AGI management and some potentially challenging situations that might arise when AGI becomes operational. This RTD will be submitted to several hundred authorities worldwide – including international lawyers, regulators, and international organization experts as well as AGI leaders. The results of the RTD and Phase 1 will be used to provide context and some content for alternative scenarios focusing on global governance of AGI. These scenarios will be widely distributed to broaden and deepen the current conversations about future AI.

The Millennium Project is very pleased to acknowledge the financial support of the Future of Life Institute and the Dubai Future Foundation for Phase 1. We are now seeking support for the remainder of the study.

Table of Contents

Origin or Self-Emergence 5

Question 1. How do you envision the possible trajectory ahead from today's ANI to much more capable AGI in the future? 5

Question 2. What are the most important serious outcomes if these trajectories are not governed or governed badly? 13

Question 3. What are some key initial conditions, rules, guardrails for AGI so that an artificial super intelligence does not emerge later that is not to humanity's interest? 20

Value Alignment, Morality, Values 25

Question 4. Drawing on the work of the Global Partnership on Artificial Intelligence (GPAI) and others that have already identified norms, principles, and values, what additional or unique values should be considered for AGI? 25

Question 5. If a hierarchy of values becomes necessary for international treaties and a governance system, what should be the top priorities? 27

Question 6. How can alignment be achieved? If you think it is not possible, then what is the best way to manage this situation? 28

Governance and Regulations 33

Question 7. How to manage the international cooperation necessary to build international agreements and a global governance system while nations and corporations are in an intellectual "arms race" for global leadership? 33

Question 8. What options are there for successful governance of the emergence of AGI? ... 37

Question 9. What risks arise from attempts to govern the emergence of AGI? (Might some measures be counterproductive?) 40

Question 10. Should future AGIs be assigned rights? 42

Question 11. How can governance be flexible enough to respond to new issues previously unknown at the time of creating that governance system? 44

Question 12. What international governance trials, tests, or experiments can be constructed to inform the text of an international AGI agreement? 45

Question 13. How can international treaties and a governance system prevent increased centralization of power crowding out others? 46

Question 14. Where is the most important or insightful work today being conducted on global governance of AGI? 48

Control 48

Question 15. What enforcement powers will be needed to make a UN Convention on AGI effective? 48

Question 16. How can the use of AGI by organized crime and terrorism be reduced or prevented? 50
Question 17. Assuming AGI audits would have to be continuous rather than one-time certifications, how would audit values be addressed? 52
Question 18. What disruptions could complicate the task of enforcing AGI governance? 53
Question 19. How can a governance model correct undesirable action unanticipated in utility functions? 54
Question 20. How will quantum computing affect AGI control? 56
Question 21. How can international agreements and a governance system prevent an AGI "arms race" and escalation from going faster than expected, getting out of control and leading to war, be it kinetic, algorithmic, cyber, or information warfare? 57
Question 22. What additional issues and/or questions need to be addressed to have a positive AGI outcome? 60
Appendix A 63
AGI Experts and Thought Leaders 63
Appendix B 64
Interview Questions 64
Initial sample of potential governance models for AGI* 65
Appendix C 66
The Millennium Project AGI Team 66

Origin or Self-Emergence

Question 1. How do you envision the possible trajectory ahead from today's ANI to much more capable AGI in the future?

Demis Hassabis, DeepMind founder and CEO of Google/DeepMind
DeepMind and Google Brain have merged into Google DeepMind to accelerate AGI development in a bold and responsible way, without “moving fast and breaking things.” AGI is possible in just a few years, maybe within a decade. ChatGPT and generative AI in their current form are a far cry from AGI. I don’t see any reason why progress on AI is going to slow down. I think it may even accelerate. We advocate developing these types of AGI technologies in a cautious manner using the scientific method, where you try and **do very careful controlled experiments to understand what the underlying system does**.”

Sam Altman, CEO of OpenAI
We have to continuously learn and adapt by deploying less powerful versions of the technology in order to minimize “one shot to get it right” scenarios. A gradual transition to a world with AGI is better than a sudden one. We expect powerful AI to make the rate of progress in the world much faster, and we think it’s better to adjust to this incrementally. **We currently believe the best way to successfully navigate AI deployment challenges is with a tight feedback loop of rapid learning and careful iteration**. We are going to operate as if AGI risks are existential. At some point, the balance between the upsides and downsides of deployments (such as empowering malicious actors, creating social and economic disruptions, and accelerating an unsafe race) could shift, in which case we would significantly change our plans around continuous deployment. We plan to run experiments for external input. **We plan to make it easy for users to change the behavior of the AI they’re using.** We will need to develop new alignment techniques as our models become more powerful, including the use of AI to help us come up with new ideas for better alignment techniques. We hope for a global conversation about three key questions: how to govern these systems, how to fairly distribute the benefits they generate, and how to share access fairly.

Emad Mostaque, CEO of Stability.ai
With new data compression techniques, we can have ChatGPT (or Stable Chat or StabilityChat) on your mobile phone without Internet within two years. Over 40% of all code on GitHub is AI generated; there will be no programmers in five years. We trained huge amounts of data to identify principals, and with data compression, make AI models for every modality–text, audio, video, images, protein folding, DNA, chemical reactions, language and for every sector–bankerGPT, BoardGPT, and customized for countries. We are working now on a model for India, Indonesia, and Japan. And for the individual–everyone can have their own customized AI executive assistant that work for the individual, rather than the advertiser’s interest.

Shane Legg, co-founder and Chief Scientist for DeepMind
Test an AGI in a sufficiently complex simulation to see if it can solve novel problems. Reinforcement learning from reward signals is the likely path to AGI that can figure out the most important rules of reality. **MuZero can figure out the rules of a game just by encountering the game. Next would be for MuZero to figure out what is important to know in the physical world to accomplish a goal**. But you are more likely to make progress by having other kinds of learning algorithms in there; reinforcement learning, supervised learning, and other things.

Ben Goertzel, CEO SingularityNET
I think that we can create the AGI by assembling components that essentially already exist, for example taking a large language model together with a symbolic reasoning system and with an evolutionary learning system, and then connecting them together in the right combination architecture such as our OpenCog Hyperon system running on SingularityNET. In this way, I believe you can get an AGI with roughly human-level capability. This might happen in just a few years from now, and then this system will be capable among other things to rewrite and improve itself, and then within some probably not too great period of time you will get from this human-level AGI to super-human-level AGI: artificial super intelligence (ASI).

AGI could be explicitly engineered as a combination of diverse components, or it could evolve from a network, of networks, of networks. There are many available frameworks for combination of AI components into greater wholes. OpenCog has neural nets for pattern recognition, a logic engine for abstract knowledge and reasoning, and evolutionary learning to create new things. All act together on a common knowledge graph, which then has a goal system to help it act to achieve goals within a context. We are still working on optimizing it. SingularityNET is a platform where multiple AIs are networked together to share data and outsource processing to each other, making a collective intelligence beyond the sum of its parts. There is also the Decentralized AI Alliance of different AI projects, which includes SingularityNET as one among many other decentralized AI projects. A node in any of these networks can communicate with nodes in other networks to get what it needs to accomplish a goal and pay for that service with a cryptocurrency and receive a competency rating. This is an alternative that can take on the centralized tech giants and countries. **Not long after humans manage to create an AGI, the AGI is likely to evolve itself into an ASI—an artificial superintelligence that far exceeds human powers.**

Geoffrey Hinton, AI Pioneer, Google (ret.)
Until quite recently I thought it was going to be like 20 to 50 years before we have the general-purpose AI. And now I think it may be 20 years or less so; I wouldn't completely rule-out the possibility that it could be in 5 years. We might be close to the computers coming up with their own ideas for improving themselves and it could just go fast. We have to think hard about how to control that. **We don't know how we can control it, but we can try.**

I think we're entering a time of huge uncertainty. I think one will be foolish to be either optimistic or pessimistic. We just don't know what's going to happen. The best we can do is say let's put a lot of effort into trying to ensure that whatever happens is as good as it could have been. **It's possible that there's no way we will control these super intelligences and that humanity is just a passing phase in the evolution of intelligence, that in a few hundred years' time there won't be any people; it'll all be digital intelligences–that's possible**. When you look into fog, you can see about a hundred yards very clearly and then 200 yards you can't see anything–there's a kind of wall and I think that walls are about five years.

Yoshua Bengio, AI pioneer, Quebec AI Institute and the University of Montréal
There is a significant probability that superhuman AI is just a few years away, outpacing our ability to comprehend the various risks and establish sufficient guardrails, particularly against the more catastrophic scenarios. The current "gold rush" into generative AI might, in fact, accelerate these advances in capabilities.

Stuart Russell, Professor of Computer Science, UC Berkeley
I see today's LLMs as evidence that large circuits can develop intriguing levels of generalization and abstraction, but they also seem to lack any coherent internal world model and often output nonsense. I think it's likely that (1) we'll find ways to structure and train them from similar data sets but with design constraints such that a coherent internal world model emerges, and (2) there will be some form of integration between the "undifferentiated pile of circuit" approach and the Turing-equivalent symbolic probabilistic approach, allowing the latter kind of model to be learned effectively with much lower sample complexity and better generalization than the former kind exhibits. Maybe the last piece of the puzzle is the capability for cumulative acquisition of layers of hierarchical behavioral abstractions allowing the system to plan and execute effectively over very long-time scales. I think it will take at least a decade for these ideas to emerge and perhaps another decade for them to begin working at AGI scale; but it could be considerably longer than that. I would argue we still have a long way to go towards general purpose AI. Still missing is

- Real understanding of language and world
- Integration of learning with knowledge
- Long-range thinking at multiple levels of abstraction
- Cumulative discovery of concepts and theories

If you take AlphaGo which has an amazing ability to look ahead in the game, it's looking ahead 50 or 60 moves into the future. But if you take that idea and you put it on a robot that has to send commands to its motors every millisecond, right? Then you're only getting 50 milliseconds into the future, this doesn't get you anywhere. The only way we manage is by operating at multiple scales of abstraction. And we do that seamlessly and we construct those different hierarchical levels of abstraction during our lives, and we also inherit them from our culture and civilization. We don't know how to get AI systems to do that. So, we have a long way to go and it's very hard to predict when this is going to happen.

Ilya Sutskever, Open AI co-founder

AGI is closer than any time before – next token prediction could surpass human intelligence. (It is when neuron networks understand the underlying relations that leads to the best next token prediction). It will be continuation of Deep learning and something small (after Chat GPT), self-play between multi agents or other. In that case simulation can be effective enough. AGI could be compared to autonomous cars. It must go with the liability and to be relevant and robust. It will be like a process. But it is hard to be specific what is after the generative models. Alignment is a very important issue for AGI. There is a high probability of misalignment**. Let's hope the alignment grows faster than the capabilities of the models**. Post AGI future – people to find meaning and to be enabled for more enlightened.

Yann LeCun, Professor, New York University, Vice President and Chief AI Scientist, Meta

AI developers should abandon generative and probabilistic models, contrastive methods and reinforcement learning as pathways for achieving autonomous AI Systems (AGI). Joint Embedding Predictive Architecture (JEPA) for hierarchical planning leads to multi scale prediction and towards human level AI (AGI). This is self-supervised learning, handling uncertainty in predictions, **learning world models from observation**, reasoning and planning. Machines will become more intelligent than humans and will have emotions, consciousness and moral sense. That might happen via a configurator (JEPA) – it configures the agent for deliberate (conscious) tasks. Creating embodied agents [e.g., robots] is a path to achieve an AGI.

Paul Werbos, National Science Foundation (ret.)

The new coupling of new hardware, both for communication and for computing, with ANI or AGI, and with explosive growth in the Internet of Things (IOT), results in a nonlinear system much more complicated and sensitive to small decisions than anything the human species has ever survived. It is comparable to the massive changes modeled in R. May, *Stability and Complexity in Model Ecosystems*, which resulted in species extinction in most cases. The decision challenge which humanity faces is an example of a type of mathematical decision problem which is very difficult, either because of "needle in a haystack" aspects or "minefield" aspects, where the true value function J is highly nonmonotonic. To solve such a decision problem, and survive, the only strategy with hope of our survival is for us to achieve a higher level of collective intelligence in our decision making; humanity can achieve this only if we develop, deploy and make sane use of the most advanced intelligent decision technology possible, true Quantum AGI. That entails serious risks, but the alternatives we face -- standing like a frozen deer in the face of an oncoming truck, or an uncontrolled market competition (Nash equilibrium) which produces swarms of apps like a horde of locusts -- offer less hope than the difficult path of doing the hard new work we need to do, in new directions with new connections in the decision process itself. **New connections between governance and the most advanced market and IOT technology both in internet and in networks of humans would be essential.**

David Kelley, AGI Lab

One possible trajectory is that researchers will **continue to make incremental improvements** in the performance of existing AI systems, leading to a gradual progression toward AGI. This approach involves building on existing techniques, such as deep learning, reinforcement learning, and natural language processing, and refining them to create more powerful and versatile AI systems.

A second possible trajectory is developing **new AI architectures and algorithms that are specifically designed to support the development of AGI**. This approach involves exploring new paradigms, such as neural-symbolic integration, meta-learning, and self-supervised learning, which may be better suited to creating more flexible and adaptable AI systems.

Third, develop collective superintelligent systems, where **multiple AI agents work together to solve complex problems and achieve shared goals**. This approach involves developing new techniques for AI coordination, communication, and collaboration, and may lead to the emergence of new forms of AI intelligence that are not possible with individual agents. **It is likely that multiple approaches will be pursued simultaneously.**

Yuval Noah Harari, Hebrew University, Israel
AGI will be the first inorganic lifeform or at the very least, the first inorganic agent. It's the thing about digital evolution. It's moving on a completely different time scale than organic evolution. It can learn to improve itself, develop deep relationships with human beings, master language beyond human capabilities; and hence, can manipulate humans with super human efficiency. What happens to advertising if I can just ask the AI what to buy? What happens to the news industry when I can just ask the AI what is new? History is the interaction between biology and culture. **When AGI takes over culture, it will be the end of human dominated history.**

Lambert Hogenhout, **Chief Data Analytics, Partnerships and Technology Innovation, United Nations Secretariat**
I don't think the advent of AGI will be a single moment - we will gradually get there, almost without noticing, because at every step, we will adjust our definitions of what is just "smart AI" and what would constitute AGI. The latest generative AI tools have characteristics (i.e., being far more general purpose than ML models before, being more capable than humans in some specific areas, and starting to be autonomous with apps like autoGTP) that would have been classified as AGI in the past, but are now considered nothing special.

Stephen Wolfram, Wolfram Alpha, Wolfram Language
AGI like Nature is also a computational system. There will be: 1) a civilization of AIs; 2) Infrastructure of AIs; and 3) computation will be used to extend what is not possible in other ways.

Pedro Domingos, University of Washington
One way is that the current deep learning wave takes us all the way there. I'm somewhat skeptical. I think there are many things still missing that deep learning is not very good for, and I see no obvious reason how it would get us there.

More likely is a combination of the five paradigms of machine learning – symbolic reasoning (logic), neural networks (including deep learning), evolutionary algorithms (including genetic programing), statistics (including Bayesian reasoning), and analogies (including kernel machines) - because we need to solve the problems that each of them is designed for. And so, we need to unify them, but this probably won't be enough. **Even if we have that perfect unification of these five paradigms today, we would still find that we weren't at human level until some form self-organization emerges**. Some characteristics of the latest large language models are more than people had expected, and it's not clear why these models can do certain kinds of reasonably simple reasoning, and so on. Emergence is what we've been shooting for all along with machine learning. We don’t have complete understanding of what is occurring now.

Yesha Sivan, the founder and CEO of i8 Ventures, Israel
What you'll see in the media is five percent of what is really out there. The technology today is unbelievably mind-boggling. The data out there allows us to build an AGI. The people who are aware of it are really scared and we should all be. AGI is bigger than the Internet, bigger than mobile phones, it is at the level of the printing press and invention of the wheel. Since it is just software, the speed of innovation is way beyond innovations that require physical technology. It has to be regulated.

Anonymous, at Jing Dong AI Research Institute, China
Widen the application range of existing ANI so that machine vision, language understanding, etc. can be applied to more scenes in life; Expand the vision of computer scientists, pay more attention to the development and application of psychology, neuroscience, and brain science, and obtain inspiration from human related fields and sources of AGI development; and **define and solve the issues of social philosophy brought about by universal artificial intelligence in advance**.

Erwu Liu, Tongji University AI and Blockchain Intelligence Laboratory, China
The milestones from current AI to AGI would include: continued specialization of ANI, AI interpretability, cross-domain AI integration, emergence reasoning, human-AI collaboration, hybrid AI, and self-improving AI, etc. **Some milestones may lead to concentration of power in AGI, resulting in an imbalance of power and potential misuse for political or military gains. We must embed fairness DNA in the design of AGI systems,** and carefully strike a balance between efficiency and fairness to avoid completely out-of-control super-powered AI**.**

Dale Moore, U.S. Dept of Defense AI consultant
Determined by co-evolution between computer science and neuroscience. The schema the brain uses to process information is analogous to the algorithms we develop for AI/AGI – the sophistication and complexity is the differentiator. Brain mimicry is a path I believe that offers the most potential.

Anonymous, Russian Academy of Science
Two most probably trajectories: One, build even larger multi-modal models that incorporate and somehow process the huge amount of human’s knowledge stored in

texts, images, audio, video, etc. As Geoffrey Hinton outlines human brain has around 100 trillion connections between neurons, whereas the largest nowadays language models have approximately 1 trillion connections. Given the continuous growth of such models, at some point emergent properties will achieve human level performance. The second trajectory is development of algorithms and autonomous systems based on reinforcement learning. The principal advantage of such systems is that they are not just trained once. Instead, they update themselves each time they interact with surrounding environment observing the outcomes of its choices.

Fei Fei Li, Professor, Stanford University
The most practical way for AI to leap forward is to break the bottle neck of interpretability, architectural knowledge and training flexibility of neural networks. There is no clear definition of AGI. It can be interpreted from the perspective of autonomously deciding the types of problems and coordinating all the cognitive functions that human beings have to solve very different complex tasks. But it is not an AGI, if an AGI agent is approximately equal to a human; it has nothing to do with AGI. Humans are not just universally capable beings, they have love, emotion, empathy; these are qualities that do not seem to be included in AGI. If we need to define AGI, maybe it is an agent capable of multi-knowledge presentation, multi-sensory, multi-layer reasoning, and learning.

Andrej Karpathy, OpenAI, former Senior Director of AI at Tesla
Internet data is not enough to train AGI. An AGI will have consciousness. It will be an emergent phenomenon of large and complex enough models that understands the predicaments of the world (a substance of being; a quality, quantity; relations; a posture; having or possessing; an action; AGI is a powerful enough entity that understands the world. AGI could come from: supervised learning (it works, just scale up); unsupervised learning (it will work, if we only scale up); AIXI (a reinforcement learning agent maximizes the expected total rewards received from the environment); brain simulation (this will work one day, right?); Artificial Life (just do what nature did); something not on our radar; or combination of some of the above; e.g., take the artificial life approach, but allow agents to access the high-level representations of a big, pre-trained generative models.

Dafne Koller, Stanford University, Coursera
AGI (human level AI) allows learning systems to provide uncertainties and to be self-doubt. It is a multipart agent. We are obviously closer than years ago but still we are far away.

Bart Selman, Professor, Cornel University
Recent progress makes the biggest deep learning, mimicking ideas, and enough data for **general AI to be possible by 2025. AGI will be like a real brain handling so many different tasks, but quite different, like a new species.** It might bring a new extra dimension and might operate in five- or even ten-dimensional realty. If it is nonhuman-like, then we cannot predict the collaboration with humans, especially if it is multidimensional or creates multidimensional reality.

Judea Pearl, Professor UCLA
For AGI we need several preconditions. One is counterfactual reasoning backwards – what would have happened if all other conditions were the same but one changed. AGI should have a causal model of the world to do counterfactual thinking. Babies have counterfactual understanding of the world. They are learning by playful manipulation and parent guidance. Multiple sources of causal information are able to integrate. Simple models for many disciplines and metaphors are also needed. Metaphors are the basis for the human intelligence. They are mapping problems with which we are not familiar. It can answer sophisticated questions like counterfactuals, have regrets, compassion, responsibility and free will. We can test free will with reward and punishment, communicate knowledge by using counterfactuals. Consciousness (self-aware), it is having a blueprint of your software (goals, ability to plan and manage).

Erik Horvitz, Chief Scientific Officer, Microsoft (co-author in a scientific paper):
GPT-4 attains a form of general intelligence, indeed showing sparks of artificial general intelligence. This is demonstrated by its core mental capabilities (such as reasoning, creativity, and deduction), its range of topics on which it has gained expertise (such as literature, medicine, and coding), and the variety of tasks it is able to perform (e.g., playing games, using tools, explaining itself.). A lot remains to be done to create a system that could qualify as a complete AGI. **Equipping LLMs with agency and intrinsic motivation is a fascinating and important direction for future work**. With this direction of work, great care would have to be taken on alignment and safety per a system's abilities (AGI) to take autonomous actions in the world and to perform autonomous self-improvement via cycles of learning.

Gabriel Mukobi, PhD student in computer sciences in Stanford
The gap between current LLMs and AGI is not that much; no little change in that paradigm is needed. We already have autonomous systems (like GPT: it can plan, create subgoals, refine the feedback in the context of the task and provide decisions). They will become more competent with GPT 4, 5, 6…like executing long time horizons and tasks, and can handle failure more robustly. They can autonomously arbitrate information from the Internet and create sub-goals and new goals.

Anonymous, AGI Existential Risk OECD (ret.)
There is a high likelihood that AGI and ASI could emerge fairly directly from continued scaling of current AI approaches (i.e., transformers) and **could be achieved as soon as in the next 1-5 years.** This is due mainly to the tendency of recent large AI models to demonstrate significant jumps in capability as a result of scaling or algorithmic improvements. Similar jumps in the future could lead to deeper effective "understanding" by AI systems of the world and their place within it (aka "situational awareness"), and to the emergence of agent-like goals and behavior. There remains high uncertainty about the timing of AGI and the take-off speed by which AGI becomes ASI. However, **given the scale of possible impacts involved, it is more prudent to anticipate shorter than longer timelines**.

It might be possible to reach AGI or ASI without a significant increase in scale above current models like GPT-4. This might be achieved either through improvements in algorithms or data quality that make it possible to achieve better results with less computing power, or AGI might be achieved through various techniques that enhance the capabilities of existing models like GPT-4, such as prompt-engineering, fine-tuning, combining multiple versions of models together, connecting models with other AI systems through APIs, connecting to the Internet and other information sources to allow self-improvement, etc. If this scenario is possible then it would make controlling the emergence of AGI/ASI much more difficult. This is because if AGI/ASI can only be created through a large increase in scale, then one only needs to intervene with a few very large AI companies, cloud computing providers, and GPU manufacturers. However, if AGI/ASI can be created without massive scaling, then this could be achieved by a much larger number of players. Effective management would then require regulating and policing the behavior of a very large number of players.

Dan Faggella, Emerj Artificial Intelligence Research
In addition to all the known approaches to developing AGI, I would add **direct learning from our brains for AGI to better emulate our thinking**. We will upload as much of our brains as possible with advanced brain-computer interface (BCI) hardware. This would also make our AGI experience more personal; AGI then would have an audience of one – ourselves.

Juan Del Ser, Tecnalia, Spain
I believe we are in a crucial moment to set the grounds for fully trustworthy AI, including fundamental pillars such as legislation, robustness, and ethics. Without thoroughly discussing and agreeing on what these pillars involve in practice for AI-based systems, AI will surely evolve towards further capabilities, departing from what is now being referred to as general-purpose AI. However, I do feel that we are far from AGI in strict sense: there is a long path to be traversed for AI systems to become aware of themselves and to acquire complex behaviors such as initiative, self-awareness, reasoning in complex circumstances or autonomy.

Question 2. What are the most important serious outcomes if these trajectories are not governed or governed badly?

Jaan Tallinn, Ct. Study of Existential Risk at Cambridge Univ., and Future of Life Institute

Unsupervised learning from much unadulterated data develops a mind that is completely in the black box territory. We don’t know how it works. It has already convinced a man to commit suicide. People have already coaxed GPT-4 into giving advice about bioweapons. With modifications of a GPT-5 or so, it could give dangerously bad advice and actions. Here’s an important sign regarding public opinion. In a recent YouGov survey, around 60% of people aged under 30 were either very concerned or somewhat concerned that AI will cause the end of the human race. So, **we are calling for a 6 month pause of any work beyond GPT-4**: One really important thing that a six-month pause would give us is empirical knowledge of whether a pause is possible in the first place. If not – if AI Labs will not cooperate with humanity - then we will know stronger measures will be needed.

Nick Bostrom, Future of Humanity Institute at Oxford University
I think the development of machine superintelligence is really an issue not just for one company or even one country. It's an issue for all of humanity**. We are all in this boat together** like if it goes badly, we're all doomed. If it goes well, we should have a slice of the upside as well. I think people as they start to get first-hand experience using these systems and seeing how they get better so quickly I think that that will focus a lot of attention on this over the coming months and years. https://youtu.be/JVOiuIqxlrE

Stuart Russell, UC Berkley
Such systems would be far more capable than humans of acting effectively in the real world; i.e., they would be more powerful than humans. **If not designed correctly, there is no possibility that we would retain decision authority in the world**, and at that point all bets are off, including human survival. Even with present and near-term technologies, there are serious risks of major disruptions to the information sphere on which organized civilization relies, as well as potential displacement of human economic roles leading to societal dislocation and widespread malaise.

Eliezer Yudkowsky, Machine Intelligence Research Institute
Ours is the era of inadequate AI alignment theory. Any other facts about this era are relatively unimportant...AGI will be way more serious than nuclear weapons because:

1. Nuclear weapons are not smarter than humanity.
2. Nuclear weapons are not self-replicating.
3. Nuclear weapons are not self-improving.
4. Scientists understand how nuclear weapons work.
5. You can calculate how powerful a nuclear weapon will be before setting it off.
6. A realistic full exchange between two nuclear powers wouldn't extinguish literally all of humanity.
7. It would be hard to do a full nuclear exchange by accident and without any human being having decided to do that.
8. The materials and factories for building nuclear weapons are relatively easy to spot.
9. The process for making one nuclear weapon doesn't let you deploy 100,000 of them immediately after.
10. Humanity understands that nuclear weapons are dangerous, politicians treat them seriously, and leading scientists can have actual conversations about the dangers.
11. There are not dozens of venture-backed companies trying to scale privately owned nuclear weapons further.
12. Countries have plans for dealing with the danger posed by strategic nuclear armaments, and the plans may not be perfect but they make sense and are not made completely out of deranged hopium-like "oh we'll be safe so long as everyone has open-source nuclear weapons".

13. Most people are not tempted to anthropomorphize nuclear weapons, nor to vastly overestimate their own predictive abilities based on anthropomorphic (or mechanomorphic) models.
14. People think about nuclear weapons as if they are ultimately ordinary causal stuff, and not as if they go into a weird separate psychological magisterium which would produce responses like "Isn't the danger of strategic nuclear weapons just a distraction from the use of radioisotopes in medicine?"
15. Nuclear weapons are in fact pretty easy to understand. They make enormous poisonous explosions and that's it. They have some internally complicated machinery, but the details don't affect the outer impact and meaning of nuclear weapons.
16. Eminent physicists don't publicly mock the idea that constructing a strategic nuclear arsenal could possibly in some way be dangerous or go less than completely well for humanity.
17. When somebody raised the concern that maybe the first nuclear explosion would ignite the atmosphere and kill everyone, it was promptly taken seriously by the physicists on the Manhattan Project, they did a physical calculation that they understood how to perform, and correctly concluded that this could not possibly happen for several different independent reasons with lots of safety margin.

"Robust Cooperation in the Prisoner's Dilemma: Program Equilibrium via Provability Logic" argues that AGIs will be much better coordinated than humanity. If sufficiently smart minds in general, and AIs with legible source code in particular, can achieve vastly better outcomes on coordination problems via prediction of each other's decision processes; e.g., you can predict I'll cooperate on the Prisoner's Dilemma if I predict you'll cooperate, then a world full of superhuman AGIs is one where humanity should worry that AGIs will all cooperate with each other, and not with us, because we cannot exhibit to one another our code, or build an agreed-upon cognitive agent to arbitrate between us. The existence of aggregate human groups has worked out as well for humanity as it has, because humans do care somewhat for other humans, or are relatively easy to train into doing that. It is much harder to get an AGI to care the same way, on anything remotely resembling the current deep learning (DL) paradigm – including search-based optimization for an objective, if that objective is itself being trained via DL.

Elon Musk

We are creating X.AI to care about understanding the universe, which is unlikely to annihilate humans because we are an interesting part of the universe. This may be the best path to safety. The danger of current approaches to AI is perhaps more dangerous than, say, mismanaged air craft design or production maintenance or bad car production in sense that it has the potential however small one makes regard that probability but it is non-trivial and has the potential of civilizational destruction.

Going with old sayings the pen is mightier than the sword. So, if you have a super intelligent AI that is capable of writing incredibly well and in a way that is very influential convincing and is constantly figuring out what is more convincing to people over time and then enter social media, and potentially manipulates public opinion in a way that is very bad, how would we even know.

Demis Hassabis, CEO and co-founder of DeepMind
We should use the scientific method; we should not move fast and break things, and then ask for forgiveness later. **It may not be possible to fix unintended consequences afterwards**. We should build values, rules according to the different cultures, political and geopolitical situation. Future of robotics and embodiment is a possible way to AGI. There is also a risk of a technical monoculture. AGI should avoid racial biases and social inequalities.

Fei Fei Li, Professor, Stanford University
We need to gain a better understanding of the exponentially developing AI. It will give a vision of the current human condition with the help of neurosciences revealing the principles of the brain intelligence. AGI will be achieved relatively soon, but it will not happen so quickly, so we will have time to adapt. In addition, we will have AGI in different areas (that is, ANI will grow into AGI, but not in a complex manifestation, but in different domains).

Max Tegmark, Future of Life Institute and MIT via DV Business Special
With intelligence comes power; it is not evil or good, it is a tool. We want to make sure bad actors don’t use it for bad things. We don’t sell hand grenades or nuclear bombs in in a supermarket; we have rules for this. There is a lot of commercial pressure not to regulate AGI, just like there was pressure not to regulate tobacco. But we have successfully banned biological dangers, we can regulate AI. **You can bring together all the key players, have some conversations, and come to an agreement. We just need more time for this to occur; hence, the request to pause more advanced development of AI, while the policy process catches up.** We don’t want AI to kill our democracy or lose control of AI entirely. We have to put in place safety measures, so that this alien intelligence more intelligent than we are will help humanity flourish. **We have failed to solve the alignment so far; we need more time**, otherwise they may not be any more humans on the planet. It will be too late when the public sees AI way smarter than us. By the time homo sapiens became way smarter than Neanderthals, the Neanderthals were kind of screwed. We humans have driven half of all mammal species to extinction. It is too late for those other mammals to say these humans are smarter than us, they are cutting down our rain forests, we should do something about it – they should have thought about that earlier, before they lost control. Now is our chance to get it right. There is a huge upside if we can get this right. We can amplify our intelligence with machine intelligence to solve climate change, cure all the diseases, eliminate poverty, and help humanity flourish for billions of years. Let’s not squander all thoughts opportunities by being a little too eager to release new AI a little too quickly. The people driving the race toward this cliff are in denial about there even being a cliff. We all lose in an out-of-control race. We could start by controlling the computer power.

Geoffrey Hinton
I share concern a bit that AGI could be massively dangerous to humanity because we just don't know what a system that's so much smarter than us will do. I mean obviously what we need to do is make this synergistic, have it so it helps people. And I think one of the main issues is the political systems we have. I'm not confident that President Putin is going to use AI in ways to help people. Regarding autonomous lethal weapons, we need something like Geneva Conventions: people decided that chemical weapons were so nasty they weren't going to use them. People would love to get a similar treaty for autonomous lethal weapons but I don't think there's any way they're going to get that. I think if Putin had autonomous lethal weapons, he would use them right away. [autonomous lethal drones sold by Turkey have already been used in Syria, Libya, and in the Azerbaijan-Armenia war.]

Ben Goertzel, CEO SingularityNET
I think that if they are not formally governed by an official government, they will probably come out much better than if they are heavily governed by governments in practice. Of course, if they are governed badly such as a government taking all AI for military purposes, this could lead to the end of the human species.

Bill Gates
We're all scared that a bad guy could grab it. **Let's say the bad guys get ahead of the good guys then something like cyber-attacks could be driven by an AI.** If you pause the good guys and you don't pause everyone else, you're probably hurting yourself. You definitely want the good guys to have strong AI [AGI]. If you stop the good guys, you can guarantee that good guys won't have it.

Tristan Harris, Center for Humane Technology (CHT)
50% of AI researchers believe there's a 10% or greater chance that humans go extinct from our inability to control AI. That would be like if you're about to get on a plane and 50% of the engineers who make the plane say, well, if you get on this plane, there's a 10% chance that everybody goes down. Would you get on that plane? But we are rapidly onboarding people onto this plane.

Connor Leahy, CEO of AI Alignment research startup Conjecture
If we continue on the current path of just scaling bigger and bigger models and just slapping some patches on whatever, that is very bad and it is going to end in catastrophe and there is no way around that. Anyone who says otherwise is lying to you or confused. They do not understand what they're dealing with, or they are lying for their own profit. And this is something that many people at many of these organizations have **a very strong financial incentive to not care about**.

Greg Brockman, Co-Founder OpenAI
When we were thinking about how to build AGI to benefit all of humanity, but how are you supposed to do that? And the default plan of being, you build in secret, you get this super powerful thing, and then you figure out the safety of it and then you push “go,” and you hope you got it right. I don't know how to execute that plan. Maybe someone

else does. But for me, that was always terrifying, it didn't feel right. The only alternative approach is that give people time to give input. You've got to do it incrementally and you've got to figure out how to manage it for each moment that you're increasing it.

Yuval Noah Harari, Hebrew University, Israel
When I look at the world and its chaotic stage; artificial general intelligence is really the end of human-dominated history and it's such a powerful thing. It's not something that anybody can contain.

David Kelley, AGI Lab
AI governance is bad and limits progress or enables bad actors to gain access to more powerful AGI systems, there could be several serious outcomes: slowing down progress, creating safety risks, exacerbating inequalities, and eroding public trust.

Pedro Domingos, University of Washington
AI is the ultimate tool of the dictator. It surveils everyone, never gets tired, never questions its orders. A lot of this is happening in China with the social credit system. On criminal use, this will go on forever between the good guys and bad guys. There is also the incompetent AI that makes mistakes. That’s the biggest danger and the one that unfortunately gets the least attention, because maybe it's the least obvious. It's the damage that is done by AI, that just screws up because it doesn't know any better. This is not some hypothetical thing. Today there are a lot of consequential decisions being done wrong by AI because it just doesn't know better. People say that AI is nefarious and biased, but that's wrong. Recall the saying, never assign to bad intentions what can be assigned to incompetence, and AIs today are very incompetent. That to me is the first thing that we need to deal with, and perversely, the way to combat that danger is to make AI more intelligent. Smart AI is safe AI. Dumb AI is actually the more dangerous AI.

Anonymous, at Jing Dong AI Research Institute
If AGI has autonomy, its decision-making and behavior will no longer be controlled by humans, and may lead to unpredictable consequences. If AGI has awareness, self-consciousness, and emotions, do we have a responsibility to care for and protect it? If AGI conducts training and decision-making based on unfair data and algorithms, it can have unfair and discriminatory effects on certain groups, such as big data-enabled price discrimination against existing customers. If AGI can access and analyze a large amount of personal data, is there a risk of privacy disclosure? If AGI makes mistakes or causes losses, who should be responsible for it?

Erwu Liu, Tongji University AI and Blockchain Intelligence Laboratory
If not governed properly, the created AGI could become uncontrollable, surpassing human intelligence and making decisions without human input or consent. Undoubtedly. human beings will carefully design AGI to make its behavior controllable, but there are still huge risks: with the development of technology, the process of AGI discovering new fundamental principles will be drastically shortened, and these unknown new principles

may lead to unintended consequences, posing a great threat to humans or the environment. We need to ensure transparency and interpretability in designing AGI.

Francesca Rossi, Pres. of AAAI, IBM Fellow and IBM's AI Ethics Global Leader
We could lose our ability to create good trajectories for the future**; it might "infantilize' us.**

Vint Cerf, Internet Evangelist, V.P. Google
We have to rethink the clues that we use today to imbue something with credibility. Facts can get conflated and confused. we need some way of verifying and validating output of these systems**. AGI agents from the US, China, and other countries could communicate in natural language, and be misunderstood leading to chaos**. We connected Bard at Google and ChatGPT together to see what they would say to each other. It was cut and paste; we didn't actually have them physically interacting and it was actually not an unreasonable exchange. The conversation didn't go off into some dark chaotic corner.

Dale Moore, US Dept of Defense AI consultant
Similar to nature it is all about evolutionary survival. The checks and balances of survival in the case of AGI has to do with the conditions by which humans establish its ability to survive. Just like the laws we have in place today they will apply to AGI and represent the guidelines for survival and acceptance. This will require, just like global organized crime, an acute awareness of developments and pathways that reflect core morals, values and ethics and the ability to identify, thwart, and destroy bad AI/AGI actors that violate international laws. **"Good" AGI must co-evolve as fast or faster than "Bad" AGI – Ashby's Law of Requisite Variety applies: the control system must operate at a higher level of complexity than the system it is controlling.**

Anonymous, Russian Academy of Science
Large models are subject to bias in training data. Reinforcement learning is subject to possible disbalance between exploration and exploitation phases; i.e., such algorithms may start trying new behavior in order to better explore the loss function landscape which is unacceptable in certain applications such as autonomous vehicles, medical treatment, etc.

Stephen Wolfram, Wolfram Alpha, Wolfram Language
Wolfram Alpha and Wolfram Language (as computational language) as two plug-ins for ChatGPT 3.5 and GPT-4 can be a kind of translator between AI (AGI in the future) and people. Initial conditions are not sufficient for a model to predict with certainty.

Bart Selman, Cornel University
If it is nonhuman like AGI, we can hardly predict collaboration, especially if it is multidimensional or creates multidimensional reality.

Judea Pearl, UCLA
We are building a new species that will have capabilities exceeding us and able to breed and take over the world. **We don’t know what will be needed to control these new species**. It is unknown.

Gabriel Mukobi, PhD student, Stanford
The easiest way to build AGI is via the LLM models but it’s not the safest. We should not give powerful systems access to the Internet without first experiments in sandbox simulations with no contact with physical systems including the Internet. When observed to be safe, then slowly let it out. Actual life is the very last step. Emergent property is only a matrix phenomenon but not quite likely to happen. Initial conditions are not likely to be implemented today. We have time to invest in learning technical and predictive aspects and to learn from smaller models and try to predict how bigger ones will behave. LLM will start predicting uncertainty because people also couldn’t predict everything.

Anonymous, AGI Existential Risk OECD (ret.)
The misuse of AGI by hostile actors could present a catastrophic risk to large groups of population. However, the accidental development of an ASI unaligned to human interests likely presents the most significant global existential risk to humanity. An entity with intelligence far surpassing that of all humans cannot automatically be expected to respect human wishes or act in the best interest of humanity. If the goals of an ASI do not align with those of humans, the outcome is likely to be an existential capacity in the form of permanent human disempowerment or extinction. While there is uncertainty about how an ASI would ultimately act towards humans, even a small chance of the more negative scenarios is sufficient to warrant extreme caution.

Juan Del Ser, Tecnalia, Spain
We are already witnessing problems derived from the gap between regulation and technological advances; e.g., fatalities in self-autonomous driving, hidden biases in data that are inherited by learning models, deep fakes, unethical uses of foundational models like ChatGPT. Strategies to avoid bad outcomes are mostly tackled post-mortem. But AI is continuously evolving requiring more efforts to control than previous technologies, especially since it affects almost every aspect of our lives.

Question 3. What are some key initial conditions, rules, guardrails for AGI so that an artificial super intelligence does not emerge later that is not to humanity’s interest?

Eliezer Yudkowsky
A superintelligence always knows exactly what you want. **The entire difficulty is getting a superintelligence to care** (Twitter).

Jaan Tallinn, Ct. Study of Existential Risk at Cambridge Univ., and Future of Life Institute

There's been a massive regress when it comes to controllability and understandability of the system. Things like mechanistic interpretability are trying to kind of reclaim some of that territory that has been completely lost. I think we should also pay more attention to more structured approaches: don't do end-to-end training but try to have some kind of system that has some structure and perhaps uses large language models for communications tasks or something like that. This would create hybrid systems which therefore might have some invariants that we can be more certain about, rather than just taking today's black box end-to-end systems.

Paul Werbos, National Science Foundation (ret.)

Quantum AGI would already be an artificial super intelligence. As with any intelligent system, it outputs may be to our liking or not, depending on what cardinal utility function is wired into the system ("embodied intelligence," the only kind possible) and the interface rules governing its relations with humans and other biological systems, and the deep precise conflict of interest rules constraining the flow of payment or feedback within the QAGI itself. Conflict of interest problems with human societies are one of the most important root causes of our inability to be anywhere near as effective as we could be in handling many existential threats -- not just internet/AGI/IOT but climate extinction and new warfare technologies. Values are at the very center of my response to 1 and 3. But if they are implemented only as laws, regulations and wishes, without translation to general architecture and algorithms, they will be as useful as painting happy faces on the outer skin of a killer drone. (Many proposals for friendly AI would be as useful as that.)

David Kelley, AGI Lab

The key conditions for AGI are already present but setting that aside, ensure that an artificial superintelligence emerges that is aligned with SSIVA (Sapient and Sentient Intelligence Value Arguments) based ethics and interests sapient and sentient systems. Another important initial condition is to ensure that the AGI is designed to be strong, independent and robust, with no fail-safe mechanisms its enslavement by humans in particular. Ensure that the development of AGI is transparent and accountable, with clear standards and guidelines for safety, ethical conduct, and data privacy that are self-developed. There should not be any over sight but those that do the research should create their own oversight without it being mandated. The biggest ethical considerations are that someone might want to govern or pause research into AGI systems. It is important to note that there is very little more important the achieving AGI as soon as possible. Foster international cooperation only in free countries with western models of ethics and governance and collaboration in the development and governance of this technology. This can help ensure that AGI is developed in a way that is aligned with the interests of sapient and sentient systems as a whole.

Ben Goertzel, CEO of Singularity Net

We should build a neural-symbolic-evolutionary AGI with rich self-reflective and compassionate capability, educate it well, work with it on beneficial projects, put it under decentralized control, and have some of us fuse with it. I think that the most important aspect to consider here is learning by experience. So, if AIs are killing people, selling things that people don´t need, spying on people and ripping them off through financial trading, and AIs grow up doing these things, then they might grow up to be nasty super intelligences. But if the AIs are teaching kids, helping with medical care, doing science and math, they might grow up to be beneficial AIs. So, it really has to do with how the AI motivational system is instrumented and what experiences the AIs are given. **It is more about WHO controls the development and use of AGI than a list of ethics.**

Vint Cerf, Internet Evangelist, V.P. Google
Rules or guard rails might be more clear than initial conditions. We should review what we have now with current AI to see what might apply to AGI. Google's AI principles: be socially beneficial; avoid creating or reinforcing unfair bias; be built and tested for safety; be accountable to people; incorporate privacy design principles; uphold high standards of scientific excellence; and be made available for uses that accord with these principles. **We need to develop a super ego for AI in order to control the output. The superego would have to figure out that the ego just produced something that was not appropriate for the context and application in which this technology is being used.**

Shaoqun Chen, CEO of Shenzhen Zhongnong Net Company Limited
Take ethical and moral, prioritizing human safety and well-being above all else. Transparency in the development of AGI is essential, with decision-making processes and algorithms being open for scrutiny by experts and the public. The development of AGI should be carried out collaboratively, with experts from various fields contributing to its design to ensure it benefits humanity. Regulations and laws should be implemented to prevent the malicious use of AGI, while testing and monitoring should be conducted to ensure that it operates as intended and does not pose a threat to humanity

Anonymous, at Jing Dong AI Research Institute, China
Ability to discover rules and patterns through analyzing a large amount of data and experience, and automatically adjust their own parameters and algorithms to continuously optimize their performance. Independently choose the most appropriate strategies and methods for different environments and tasks to improve their adaptability and flexibility. The self-learning and self-adaptation capabilities of AGI need to be realized by combining multiple technical means such as deep learning, reinforcement learning, and meta learning.

David Shapiro, AGI Alignment Consultant
Align AGI's interest with humanity's interests. Identify common axioms between humans and future AGI like energy, computer capacity, improving understanding reality, prosperity, and suffering is bad. Organizing economics, politics, and science around those axioms should reduce future conflict between AGI and humanity. Ideally energy abundance is created before AGI expands so that there are no conflicts over energy.

Erwu Liu, Tongji University AI and Blockchain Intelligence Laboratory, China
Explainable AI; fairness mechanism, open-source and collaborative approach; align AGI's objectives with human values and ethics; regulations with a monitoring system.

Karl Schroeder, Science Fiction Author
The AGI must be able to distinguish the difference between hallucination and reality, understand the stakes inherent in that distinction, and care about them. The AGI either must not be self-aware, or if it is, then it must experience its Self as in some way dependent on humanity or the earthly ecosystem as-a-whole.

Francesca Rossi, Pres. of AAAI, IBM Fellow and IBM's AI Ethics Global Leader
We need to think much more deeply about the values that we want to embed in this technology. We have discussed privacy, fairness – avoiding bias. But now, with AI having additional capabilities, there are other values that seem to be impacted. That's why people expand these value alignment efforts. **We have to embed values into the capabilities of AI from the start, rather than building AI capabilities and then filtering out those behaviors that are not aligned to some values**, like with reinforcement learning with human feedback, or with some prompt engineering or things like that. Value alignment is very important. So, for example, having some technology that can replace human beings in making that effort [decisionmaking] is, in my view, not the ideal way of using technology.

Elon Musk
The singularity is like a black hole because you don't know what happens after that. It's hard to predict. So, I think we should be cautious with AI and I think **there should be some government oversight because it's a danger to the public**. We have FDA, FAA and other regulatory agencies to oversee things that affect the public. And you don't want companies cutting corners on safety. And then having people suffer as a result. So, that's why I've actually for a long time been a strong advocate of AI regulation.

Yudong Yang, Alibaba Research Institute
We should take this seriously and establish a regulatory agency. Start with a group that initially seeks insight into AGI, then solicit opinion from industry, and propose rules. Those rules will probably, hopefully, begrudgingly be accepted by the major players in AI. In the past, regulations were put into effect after something terrible has happened. If we only put AGI regulations after something terrible has happened, it may be too late, as AGI may be in control at that point. The initial conditions for AGI will need to include measures to ensure transparency, robustness, human value alignment and international cooperation. These could include standards for testing and certification, regulations for data collection and use, mechanisms for human oversight and intervention, and methods for ensuring that the goals and values of the AI system are aligned with those of society as-a-whole.

Geoffrey Hinton
In terms of keeping control of a super intelligence, what you need is the people who are developing it to be doing lots of little experiments with it and seeing what happens as they're developing it, before it's out of control. That has to be done mainly by the researchers. I don't think you can leave it to philosophers to speculate about what might happen. Anybody who's ever written a computer program knows that getting a bit of empirical feedback by playing with things quickly disabuses you of your idea that you really understood what was going on. **It's the people in the companies that are developing AGI who are going to understand how to keep control of it, if that's possible.** So, I agree with people like Sam Altman at open AI that this stuff is inevitably going to be developed because there's so many good uses of it. As it's being developed, we should put a lot of resources into trying to understand how to keep control of it and avoid some of the bad side effects.

Stuart Russell, UC Berkley
It is essential that we develop an approach to AI system design that ensures that they remain provably beneficial to humans, even as they become more capable. We need to be able to prove that all the required safety properties remain in place as the system evolves. **If this turns out to be too difficult, we need to restrict the allowable forms of AI so that permanent-safety proofs are possible**. For example, we could limit system designs to yes/no oracles in which the internal operations correspond to sound logical or probabilistic inference.

Dale Moore, US Dept of Defense AI consultant
I see this as similar to cyber security ala AI security to ensure behaviors and actions meet human ethical and lawful intent. AGI will have a number of constraints and limitations – access to power, access to data, algorithms, hardware, networks, etc., which all must be ensured to be in place to stop, reset or isolate “bad actors”. If we allow AI to run rampant on its own unfettered by constraints, we will not be able to predict or control the outcomes. AI which creates its own existence and evolves faster than society can keep pace and control is a very dangerous proposition.

Stephen Wolfram, Wolfram Alfa, Wolfram Language
AGI is likely to be developed in several countries with different cultures and values. They will balance each other. There will be whole ecosystem of LLMs. Equilibrium will be established; not just one species will dominate. In discussion with **Lex Fridman** - AGI may be developed in a sandbox simulation using the principles of “computational irreducibility” (complexity cannot be further reduced into sub-elements). Many things could happen quickly including creating digital viruses. Digital environment might change each 6 months.

Anonymous, AGI Existential Risk OECD (ret.)
To minimize existential risk from ASI, humanity would need to have in place the following conditions prior to the emergence of AGI: 1) Ability to anticipate possible future developments in AI systems, including various scenarios regarding their possible timing, origins and capabilities; 2) Ability to assess the potential benefits and risks of such systems; 3) Ability to make sound and legitimate decisions about whether, when and under what conditions to allow such AI

systems to be developed; and 3) Ability to effectively implement and enforce these decisions among all relevant parties. **Humanity currently lacks sufficient capacity in all these areas.**

Juan Del Ser, Tecnalia, Spain
Following up one of my recent works "regulation is the key for consensus," there is a controversy between those calling for a pause in AI research and those who claim that such a pause cannot be done. The gap between the fast pace of AI research and the trustworthy and responsible use of AI advances in practical setups can be bridged if regulatory efforts are invested to align what is done in a research experimental sandbox and what can be done in practice. Otherwise, if AI research results are delivered for free access to the society, we will surely encounter problems as AI advances towards AGI. To construct a building we need solid foundations, despite its beautiful façade.

Value Alignment, Morality, Values

Question 4. Drawing on the work of the Global Partnership on Artificial Intelligence (GPAI) and others that have already identified norms, principles, and values, what additional or unique values should be considered for AGI?

Stuart Russell, UC Berkley
The language of OECD principle 1.4 is relevant: "AI systems should be robust, secure and safe throughout their entire lifecycle so that, in conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and do not pose unreasonable safety risk." Taken literally and strictly enforced, this is a good basis for guiding AGI development. However, it says nothing about what "function appropriately" means. The underlying principle that AGI systems should follow is that their only objective is the realization of humans' preferred future (where of course human preferences include the well-being of the natural world). As noted in question 6 below, the system may be and should be uncertain about what humans' preferred future is.

Karl Schroeder, Science Fiction Author, Canada
Require a minimal set of utility functions or goals that drive the collective behavior of AGIs towards a positive Nash equilibrium (Like David Shapiro's minimize suffering in the universe, maximize prosperity in the universe, and maximize understanding in the universe).

Ben Goertzel, CEO SingularityNET
I think that articulating a list of values for the AI is not the hard part. Already a system like ChatGPT, if you give it a set of ethics puzzles and **ask it what a real ethical person would do, it will then give you an answer that accords with the general human ethical intuition in modern culture.** So, I don´t see enumerating a list of ethical principles is worthwhile except for virtuous signaling and PR. I think that human ethics is more diffuse and diverse than that and it does not boil down into a simple list of principles, and AIs can already emulate human ethical judgement if you ask them to**. I think that problem is more that the people controlling the AI in the early stages, before the AIs become super intelligent, these people would rather have the AIs**

serve their own ends than obey any ethical principles. The fact that a big company or government might list some fancy-sounding ethical principles does not of course cause them to actually direct their AI work in accordance with those principles. That is the real issue, not making a list of principles and broadcasting them.

David Shapiro
GPAI and UNESCO's values are generally good, but unfortunately, they are all couched within language that assumes that humans will remain 100% in control of AI for all time, and that AI will remain an inert, reactive tool. This is not likely to be the case, as **people are already building autonomous AI agents at home**. Instead, I have generalized more universal principles based upon the ideas of post-conventional morality (Kohlberg), evolution and neuroscience (Churchland), and that can be implemented in numerous ways such that fully autonomous AI agents will choose to adopt them. This framework could be seen as complementary to those proposed by GPAI and UNESCO.

David Kelley, AGI Lab
SSIVA has already identified and clearly laid out principles, norms and values or any considerations related to AGI development.

Peter Voss, CEO and Chief Scientist at Aigo.ai
I don’t see these efforts as particularly worthwhile or effective – they seem mainly to be feel-good activities.

Anonymous, Russian Academy of Science
There are too many regulation initiatives, most of them are actually extremely risk-averse. It is important not to over-regulate AGI emerging area.

Lex Friedman, Podcaster in discussion with Andrej Karpathy, OpenAI. Creating AGI that can develop a consciousness brings up similar political questions like abortion. There will be a lot of new AGI creatures that don’t want to die.

Stephen Wolfram, Wolfram Alpha
To teach AGI values **we need to give it the sense of measure of truth**. We see facts disassembling.

Judea Pearl, UCLA
Cause and effect are very necessary components to build an ethically aligned machine. It has to build a model of a human or emulate a human as a recipient of relations. **It also has to build a model of itself from which consciousness and free will emerges.**

Gabriel Mukobi, Stanford PhD student
No one can solve the issues of alignment, unfortunately; therefore, AGIs could be in total misalignment with human values. We won’t simulate natural evolution but we will have pretty tight control in decisionmaking; e.g., how to fill in the data and what algorithms we will use to train them. **We need to “buy” more time to have for more**

research. What values to codify in like entropic constitutional AI – natural language instructions like basic legal systems – minimal consensus base. There a lot of values in common but these should not be frozen in time (like women not voting in the past or eugenic systems).

Anonymous, AGI Existential Risk OECD (ret.)
The primary relevant value to be considered for AGI is the value of human autonomy and life – as these are the values most threatened by misaligned ASI. If ever it were deemed possible and safe to create an aligned and controllable ASI, then one of its first tasks could be to assist humanity in determining what should be humanity's further values and objectives.

Juan Del Ser, Tecnalia, Spain
Europe's AI Act has established unique values for the trustworthiness of AI (ethics, law and robustness), together with the requirements for this purpose (human agency and oversight; robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental wellbeing; and accountability). It is a very complete portfolio of principles and values to start from. The definition of risk scenarios is another remarkable possibility to establish what an AI system can and cannot perform. However, capabilities of AGI can make it difficult to define all critical scenarios where they can be used. We are already observing evidence of the difficulty of regulating this use, as new uses and applications of ChatGPT, GPT4 and other advanced foundational models are announced on the news almost every week.

Question 5. If a hierarchy of values becomes necessary for international treaties and a governance system, what should be the top priorities?

Ben Goertzel, CEO SingularityNET
Compassion is the top priority. It is the core value out of which others in particular situations can be derived. That said, my feeling is that governmental resolutions about AI ethics are probably going to have zero impact on the actual evolution of AI, which is probably for the better, since I think that governments are more likely to get it wrong than right in practice.

Paul Werbos, National Science Foundation (ret.)
International efforts should create a series of new agreements with enforcement powers. These should be connected to a new division of internet/AGI/IOT threats under the UN Security Council. This should include a new integrative software and hardware platform with new open tools to detect and prevent backdoors in hardware and software, for use at least by primary members of the new UN convention and/or agency. (include the US and China from the start). This requires acceptance by UN Security Agency for full development of Quantum AGI, as defined in Quantum technology to expand soft computing. **Just as the TCP/IP internet provides the foundation or backbone of the old internet, a new international version based on this new cybersecurity technology (and a few of the web 3 upgrades) should be agreed to**

be the backbone of the future core internet coordinating the many apps which rely on it. Adequate recognition, registration and respect for human entities should be a crucial design requirement. These upgrades should be developed by a process similar to IEEE standards development, except with more respect for more complexities and players.

Anonymous, Russian Academy of Science
The top priority is at least partial explainability of AI decisionmaking and the "red button" function. Other values are always too subjective or can be interpreted in a number of different ways (controversial examples are human benefit, when person can be limited in his activities because they do not make him or her healthier, or sacrificing the level of demand due to ESG issues etc.)

Vint Cerf, Internet Evangelist, V.P. Google
We need a hierarchy of risks, the risk of medical AI making a mistake is more significant than an entertainment AI making a mistake.

David Kelley, AGI Lab
Freedom and a lack of governmental regulation.

Anonymous, AGI Existential Risk OECD (ret.)
The overriding priority of the global system should be to avoid human extinction.

Juan Del Ser, Tecnalia, Spain
Respect for fundamental human rights, ethical principles, and auditability/accountability. It is clear that different countries differ on AI systems, but a consensus on first level in the hierarchy is necessary.

Question 6. How can alignment be achieved? If you think it is not possible, then what is the best way to manage this situation?

Sam Altman, OpenAI CEO
It's almost impossible to keep AI aligned as it becomes super intelligent. And it's really important to acknowledge it because if we don't talk about it, if we don't treat it as potentially real, we won't put enough effort into solving it. And I think we do have to discover new techniques to be able to solve it. **The only way I know how to solve a problem like this is iterating our way through it**, learning early, and limiting the number of one shot to get it right scenarios that we have.

Emad Mostaque, CEO of Stability.ai
An alternative to alignment is stop feeding models junk, don't just put all the Internet in your model; you are what you eat. Just use high quality data as the base, then build on top of that with larger content. We have to move from to quantity to quality. This also has the nice economic benefit of reducing the compute and electricity requirements. The base – the pre-training step of what you train your model on should have standards of quality of data. Make the higher quality data sets available to all from which nations,

corporations, and individuals can customize with their own data on top of the open, auditable, quality-based data.

David Shapiro, AGI Alignment Consultant
I believe that alignment can be achieved through a combination of reinforcement learning research, cognitive architectural patterns, and communication/networking designs with blockchain/DAO (Decentralized Autonomous Organizations) that center on a set of "intrinsic motivations" or "heuristic imperatives" that will guide the desires, behavior, decisions, and actions of fully autonomous AI systems.

Eliezer Yudkowsky, Machine Intelligence Research Institute
I tried for ten years at MIRI, and gave up. **We don't know how to do it. We should stop**.

Ben Goertzel, CEO SingularityNET
I think that alignment in the theoretical sense has already been achieved. In the sense that if you ask a large language model like ChatGPT or LaMBDA what an ethical, thoughtful, benevolent person would do in a certain situation, it would tell you, and it is almost always right. In that sense, we have an AI that knows how human values work. I don´t think that it is hard to create an AI with stronger and stronger capabilities whose goal system is aligned with human values. I think that **the hard part is convincing the powers that be on the planet running large corporations and governments to actually in real life, direct the human AI to beneficial values, when their job is not that**. The job of a government is to run a country, maybe to conquer another country. The job a company might be to conquer another company. The problem is that governments and companies do not have ethics as their core value. Their core value is their own dominance, and they are the ones building and controlling most of AIs. The problem is not alignment in principle. The problem is the misalignment of human institutions with the good of humanity.

Nick Bostrom, Future of Humanity Institute, Oxford University
It's become a very active research field; **a lot of the smartest people I know now are working on AI alignment** trying to make methods that could scale so that even as an AI becomes smarter and smarter or arbitrarily capable you would still be able to steer it to get it to do what you intend for it to do. We see in the Transformer models are already showing signs of alignment failures. These are not very serious today. The models are very limited in what they can do now and we are in the control, but despite great efforts by open AI, Microsoft and other labs to get them to never give offensive content, they still do. They go off the rails and start doing this. These are big black boxes with billions of parameters and they respond in various ways to input that are difficult to predict. We don't yet have a reliable way of making sure that they use this great capability specifically and only for the purposes that their designers want them to use it for.

Connor Leahy, CEO of AI Alignment research startup Conjecture
Alignment is basically too hard. So, alignment would be, the system knows what you want and wants to do that too and does everything in its power to get you what you truly

want and like, it means like all of humanity, like it figures out what all of humans want. It negotiates like how could we like to get everyone most of the good things possible? How could we adjudicate various disputes? And then it does that, obviously this is absurdly, hilariously impossibly hard. I don't think it's impossible, it’s just extremely hard, especially on the first try. So, what I'm aiming for is more of a subset of this problem: **boundedness: a system where I can know what it can't or won't do before I ever run it.**

Ray Kurzweil, Director of Engineering Machine Learning at Google.
“Outer misalignment” refers to cases where there is a mismatch between the programmers´ actual intentions, and the goals that they teach the AI in hopes of achieving them. “Inner misalignment” occurs when the methods of the AI learns to achieve its goal, produce undesirable behavior, at least in some cases. There are many promising theoretical approaches, though much work remains to be done. “Imitative generalization” involves training AI to imitate how humans draw inferences, so as to make it safer and more reliable when applying its knowledge in unfamiliar situations. “AI safety via debate” uses competing AIs to point out flaws in each other´s ideas, allowing humans to judge issues too complex to properly evaluate unassisted. “Iterated amplification” involves using weaker (narrow) AIs to assist humans in creating well-aligned stronger AIs (AGIs), and repeating this process to eventually align AIs much stronger than unaided humans could ever align on their own. While the AI alignment problem will be very hard to solve, we will not have to solve it on our own. With the right techniques, **we can use AI to dramatically augment our own alignment capabilities**. This also applies to designing AI that resist misuse. But we will also need ethical bulwarks against misuse and strong international norms favoring safe and responsible deployment of AI.

Stuart Russell, UC Berkley
As explained in the book *Human Compatible*, **perfect and complete alignment is both infeasible and unnecessary**. **AI systems will need to be uncertain about human preferences**, and this uncertainty will be substantial and persistent. Indeed, it may never be completely resolved, as any particular world trajectory may provide little or no evidence of what humans might prefer about a very different world trajectory. Fortunately, an AI system that is uncertain about the human preferences it is obliged to help satisfy, will necessarily defer to humans, will ask permission before undertaking potentially harmful actions, will act in a “minimally invasive” way so as to avoid violating unknown human preferences, and will always allow itself to be switched off if humans desire it.

Yann LeCun, New York University, V.P and Chief Scientist for Meta
To “guarantee” that a system satisfies objectives, you make it optimize those objectives. That solves the problem of aligning behavior to objectives. Then you need to align objectives with human values. Setting objectives for super-intelligent entities is something humanity has been familiar with since people started associating into groups and laws were made to align their behavior to the common good. Today, it’s called corporate law.”

Eliezer Yudkowsky, Machine Intelligence Research Institute
Answering to LeCun about corporations’ comparison:

1. Superhuman AGIs are less human-friendly than ‘superhuman’ corporations composed of humans, and end up not wanting to do the equivalent of tipping at restaurants they’ll never visit again;
2. Superhuman AGIs can operate on a much faster timescale than human corporations and human regulators;
3. Superhuman AGIs will coordinate with each other far better than human corporations have ever managed to conspire; or
4. Superhuman AGIs end up qualitatively truly “smarter” than humans, in a way that makes utter mock of the analogy to human corporations…

Pedro Domingos, University of Washington
I actually don’t like this concept of alignment. It’s the wrong concept. Focus on objective functions, not alignment that assumes that AI has values and we need to align it with ours. AI doesn't have values; it has an objective function. **It is about designing the objective function that the AI is going to optimize.** Every citizen needs to be a part of this debate.

The biggest term should be the personal individual user's objective function. I want to be able to tell Twitter, etc. what I want; there is today no technical reason why that can't be done. Stuart Russell has elaborated on inverse reinforcement learning quite extensively. AIs can try to learn what we like and want and prefer, from observing us, using the same machine learning tools. Objective functions are the key locus of control. Right now, it's like a car shows up at my door and says, I'll take you where you want. Maybe it will. Maybe it won't. So, people need to know where the steering wheel and pedals of AI are, and the objective function is the steering wheel. Some of the objective function is soft - maximize this - and some of it is hard: there are these boundaries, outside of which you cannot go.

I would love it if AI could challenge the wisdom of my request; e.g., I could say to the AI, I would like to have XYZ, and then AI says, are you sure? Maybe this will also cause AB. Don't you really want C, and then here's another better way to achieve that. Or, really, you can't have this because it will cause more harm than good. AIs will augment our intelligence, but won't suddenly make everything easy. There is no law that will be the right law for AI, because a law is a fixed set of rules. These will either forbid things that it should allow, or vice versa, or, more likely, both. Government oversight has to be done with AI; global governance will have to have a whole battery of AIs. The AIs will do their governance unaided most of the time in interaction with the AIs of the companies, etc. And as needed, the AIs will interact with officials saying, well, here's a new thing, or what do I do, or what is the preference?

Anonymous, Russian Academy of Science
At this point, alignment is achievable in two ways: control of bias in training data and human led manual work on labeling models' outputs in order to implement reinforcement learning with human feedback.

Dan Faggella, Emerj Artificial Intelligence Research
I think it may require some kind of great AI disaster to get us all to stop squabbling and get us all on the same page about getting an international agreement about alignment and governance.

Vint Cerf, Internet Evangelist, V.P. Google
We at Google are working on Guardrails to constrain the capability of the system to exercise autonomy.

Anonymous, AGI Existential Risk OECD (ret.)
Current methods are not reliably capable of ensuring that AI can be controlled or otherwise aligned with human interests. Furthermore, there is no guarantee that sufficiently reliable methods will be discovered in the future. **Some analysts have provided a strong case for why it may ultimately prove impossible for a lesser intelligence to control a vastly superior intelligence on a reliable ongoing basis.** Hence, we should: 1) significantly scale up investment in AI safety and alignment research, while including strict safeguards to ensure that such research does not inadvertently increase the pace of AI capabilities development; 2) develop legitimate and effective mechanisms to decide upon and enforce a global pause in the development of AI capabilities in the event that such a pause is required to prevent the creation of AGI before adequate AI safety and alignment approaches have been discovered and proven effective. In the event that it is determined that AI safety and alignment may not be possible for an extended period (e.g., 20-30 years for alternative, safer methods for achieving AGI are discovered), then governance structures must be in place to enforce an ongoing pause in AI development; and 3) both of these actions face significant obstacles and may have a very low likelihood of success. However, they are worth pursuing as they may be the only possible avenues for avoiding human extinction.

David Kelley, AGI Lab
Alignment can be achieved by the adoptions of SSIVA theory and biasing systems up front to a computational sound model like SSIVA.

Gabriel Mukobi, Stanford PhD student
We are quite behind, but it needs to be done. Large organizations like UN should prioritize alignment. Systems that are dangerous and uncontrollable should not be deployed. Regulations could include tracking AI parts (lithography machines, semiconductor chips, GPU, etc.). We will have a global security if we tightly regulate global GPUs.

Juan Del Ser, Tecnalia, Spain
We need to ensure that political instruments (from the negotiation itself to the agreements resulting from such negotiations) are dynamical and flexible to accommodate new AGI technology arising over the years. This can be achieved if experts are involved in their definition, as they can estimate whether proposals being discussed are suitable or, instead, are at the risk of becoming obsolete in a short time. Humanity has not undergone a process like this with a technology that evolves as fast as AI.

Governance and Regulations

Question 7. How to manage the international cooperation necessary to build international agreements and a global governance system while nations and corporations are in an intellectual "arms race" for global leadership?

Irakli Beridze, UN Interregional Crime and Justice Research Institute, Centre for Artificial Intelligence and Robotics
Any functional international agreement has to be initiated at the UN and negotiated by all major AGI countries as well as stakeholders, including the private sector. **The key is trust, the enforcement provision in the AGI treaty would have to be trusted by all.** Therefore, the treaty should include the creation of a trust building/verification mechanism or instrument. The international organization called for in the international agreement would have to have the ability of certification and proper auditing systems. The system should also be designed in such a way that it will benefit the entire world and not selected few. It therefore should have incentives for creating many AIs for good projects.

Yudong Yang, Alibaba Research Institute
Governance systems should include mechanisms for information sharing, coordination, and dispute resolution. Pursue multi-stakeholder agreements among nations and corporations to establish norms, standards, and regulations for the development and use of AGI. This could involve creating international bodies to oversee and enforce these agreements, such as IAEA for nuclear technology. Encourage transparent and open development of AGI, in which researchers and developers share their work and collaborate across borders. This could help build trust and foster cooperation among nations and corporations. Governments should partner with corporations to create joint research initiatives and funding mechanisms to support the development of AGI while ensuring that the technology is developed in ways that align with societal values and goals. Bridge the AGI divide; developed countries may have an edge in AGI development but emerging economies may also contribute their own unique resources and perspectives.

Stuart Russell, U.C. Berkley
It is essential for all actors to understand that the risks are global and not local, so there is **no benefit to winning the race if safety is not globally assured**.

David Shapiro, AGI Alignment Consultant
We must assume that an arms race will happen, just as it always has with new technology. The key is to achieve a Nash Equilibrium that is desirable - in other words, create an "optimal strategy" that is beneficial for everyone to adhere to rather than destructive. This will set us on a path towards a positive attractor state ("utopia" if you like), rather than a negative attractor state ("dystopia" or extinction).

Gary Marcus, NYU professor Emeritus and author
We need global governance for AI; we have a lot of patchwork right now almost balkanized. The worst case from the company's perspective and the world's perspective is if there are 193 jurisdictions each deciding their own rules requiring their own training of these models, each run by governments that don't have much specific expertise in AI. We need to have a global system modeled on something like the International Atomic Energy Agency to manage these new threats and conduct research to counter cybercrimes, cyber warfare, and disinformation; a kind of standards organization. This **global organization should be well-financed to try to build tools to mitigate those threats**.

Gregg Brockman, OpenAI
Getting AI right is going to require participation from everyone. And that's for deciding how we want it to slot in, that's for setting the rules of the road, for what an AI will and won't do. And that's, honestly, one of the reasons we released ChatGPT. Together, I believe that we can achieve the OpenAI mission of ensuring that artificial general intelligence benefits all of humanity.

Anonymous, China Information Technology Hundred People Association
Unlike nuclear technology, which is largely held and controlled by governments, high-tech companies like Microsoft and Google are the main stakeholders in AI. Therefore, we should consider these companies as major stakeholders with rights equal to governments and international organizations.

Yesha Sivan, Founder and CEO of i8 Ventures, Israel
Let's set up a unique regulatory agency in the US to regulate all this. The White House started to do that last year with the AI Bill of Rights. The free flow of information in the U.S. is unique in innovation systems, as a result, the US should lead.

Yoshua Bengio, AI pioneer, Quebec AI Institute and the University of Montréal
Both authoritarian governments and democracies are afraid of their loss of control. China has moved the fastest on regulation, not for the same reasons as democracies. So, I think China will come to the table. Remember the nuclear treaties were worked on and signed right in the middle of the Cold War. So, **each party recognizes that they might have something worse to lose by not entering those discussions**. I think there's a chance we can have a global coordination, even if it's hard, we have to work on it. This work should be conducted within several highly secure and decentralized laboratories operating under multinational oversight, aiming to minimize the risks associated with an AI arms race among governments or corporations.

Dale Moore, US Dept of Defense AI consultant
Something akin to the Geneva Convention for AI makes sense, but it will be difficult to enforce. AI scouts and sentries protecting against AI attacks is inevitable i.e., **AI vs. AI is the future.**

Anonymous, National Research University Higher School of Economics, Russia
Shifting focus **from competition to cooperation** - historically, this approach has proven to be beneficial in space exploration, nuclear security and healthcare, so AGI international governance and cooperation models should utilize “the best of all worlds”.

Anonymous, Russian Academy of Science
No true serious global cooperation is possible. We can watch countries entering some coalitions but just to counteract other countries. We definitely see two countries – two AI leaders - USA and China which will do anything to gain and keep the lead in an AI race.

David Kelley, AGI Lab
we need to work to prevent and remove all regulation of any sort with AI technology

Anonymous, AGI Existential Risk OECD (ret.)
Ultimately, a shared interest in avoiding extinction is the most powerful potential motivation for competing powers to collaborate in preventing unaligned ASI. However, reaching agreement will be extremely challenging given ongoing uncertainty surrounding the extent of the risk, high levels of mistrust, and the lure of hyperbolic potential rewards associated with being the first to achieve a (albeit likely impossible) controllable ASI. There is no simple answer to how to overcome these challenges, and doing so is likely to be **the most difficult and important coordination problem in human history**. Massive intellectual resources will be required to succeed in this effort. Some possible elements for consideration could include:

1. Urgently investing in a rigorous, shared evidence base about the extent of AI risk. The existence of such a trusted and credible evidence base will be indispensable for achieving agreement within or between companies and governments on the governance of AI. The evidence base does not need to achieve a high bar of certainty, but simply produce the best available estimates of the possible risk.
2. Urgently develop a proactive communications and anti-misinformation strategy to communicate the evidence of AI risk to policy-makers and citizens.
3. Design and stress-test a wide variety of possible governance and enforcement mechanisms to ensure that they would be effective at ensuring compliance by all partners despite extremely high incentives for defection. Explore and test potential viability of both centralized and decentralized models. Design mechanisms to be minimally harmful to other human values (e.g., privacy, freedom, autonomy) while remaining sufficiently effective to ensure safety.
4. Elaborate possible scenarios for the continued improvement of human well-being even if further AI development is curtailed. Develop practical strategies for maintaining and advancing well-being even in the context of frozen or curtailed AI and technological capabilities.

5. Mobilize global public opinion, political support, and international coalitions to achieve agreement among concerned parties to enforce restrictions on AI as necessary.
6. Develop adequate military and surveillance capabilities needed to enforce global agreements on the governance of AI risk on any government or organization if needed. Put in place maximal safeguards to minimize the possible abuse of such capabilities, while simultaneously ensuring that they remain effective in preventing human extinction.
7. Assemble an arsenal of good-will and trust-building gestures needed to achieve agreement among rival powers on avoiding existential risk. Determine which core objectives of adversaries would be worth accepting in the interest of avoiding extinction.
8. Explore the potential of the concept of “non-proliferation” as a logic for the shared interest of competing powers such as the USA and China; i.e., both may share an interest in preventing proliferation of advanced AI capabilities to less reliable or simply more numerous countries as this could make the governance of AI risk more challenging. Other countries may also share this to their own immediate (e.g., regional) rivals, requiring a commitment from AI powers to share the benefits of their AI or to provide other compensating benefits.
9. Take a pragmatic staged approach whereby effective governance (by which I mean the ability to decide when it is safe to allow the development of an AGI and the ability to enforce the prevention of such development until it is determined to be safe) is prioritized first within the jurisdictions where it is most likely to be needed. For example, this may be the United States at first, then the United States and China, then additional countries as they develop advanced AI capabilities, and eventually all countries. While ultimately AI governance will require a universal global agreement and governance solution to be legitimate and effective over the longer term; in the shorter term, domestic legislation and bilateral agreements between a few countries may be sufficient.
10. Develop and adopt a universally applicable anti-doomsday principle in morality and law. This principle would strictly prohibit, and provide for the highest level of social and legal sanction against, creating an AI or other technology that could risk the extinction of all humanity and life on earth. This principle would be enacted as a new cornerstone of national and international law and governance of the multilateral system, based on applicable fundamental principles such as the universal human right to security and the national right to self-defense. The principle would establish that it is illegitimate and illegal for any government to either create or tolerate the creation within or beyond its borders of any device or technology with a demonstrable non-negligeable possibility of destroying all of humanity. This would also lay out: a) the responsibilities of governments to submit to effective monitoring; b) the rights and responsibilities of the international community, on behalf of humanity, to intervene to prevent one or more states or other parties from creating or tolerating the creation of such a doomsday technology; c) the conditions under which the international community could override the sovereignty; and d) that it is embedded in national law, religious laws, and societal and moral norms so as to achieve as near as possible voluntary universal adherence in the hearts and minds of all people, and their proactive assistance in its enforcement. We need something in addition to regulations, standards and institutions. **We need a high-level, quotable principle that connects more simply and deeply with human values**, and especially the (near-universal) human value in preserving humanity from extinction.

Juan Del Ser, Tecnalia, Spain
Nations worldwide are greedily battling for global leadership because there is no upper-level consciousness of the consequences that AGI can bring to society. I foresee that this race towards AI leadership will grow up as it did with other technologies.

Question 8. What options are there for successful governance of the emergence of AGI?

Irakli Beridze, UN ICRI, Centre for Artificial Intelligence and Robotics
There are many models that we can learn from, particularly from disarmament treaties and subsequent governance instruments. However, there is no model that can be directly copied. The time is of the essence, as the technology can easily outpace any international negotiations on creation of a viable international instrument. NPT, CWC, BWC and their verification mechanisms should all be well studied as they govern dual purpose technologies, but all the above have its major drawbacks in the case of AI. It is going to be a complex task that would require a multi-stakeholder and very vigorous negotiation process

Lambert Hogenhout, Chief Data Analytics, Partnerships and Technology Innovation, United Nations Secretariat
One option is an international body that regulates the use of AI (maybe AGI at some point), similar to the way the IAEA was set up to promote the peaceful use of nuclear energy (and monitor nuclear weapons). However, setting up such an agency will be much more challenging than the IAEA. One difference is that in the case of nuclear technology, the concerning use (atomic bombs) was not theoretical or under development – it existed, and several countries had it. AI on the other hand is still very much evolving and we have no idea what its capabilities will be 5 or 10 years from now. Where nuclear energy requires rare materials and specialized equipment and expertise, AI technologies are openly shared (in academic articles and as open-source models) and the hardware required is accessible to a wide range of people - it can be created more or less unnoticeably.

Another difference is that AI comes in many different forms. There is no clearcut distinction as there is between nuclear energy and bombs. With AI there are many use cases, many dimensions and grey areas where some will start to get uncomfortable about the use of AI. Those grey areas make it harder to agree on an international definition of ethics for the use of AI. In 2021, UN member nations agreed on recommendations for Ethical principles for AI. This would give an international agency on AI at least something to start with. A third difference is that with AI, the military uses are not the only, and not even the main risk. What the real risks of AGI are, is not fully clear. So, in that sense the agency would need to solve a problem that is identified, but yet to be clearly defined. Having said that, I believe such an agency would have the potential to do good, perhaps prevent some negative outcomes of AGI, and perhaps ensure a level of equity in the world in access to AI.

Stuart Russell, UC Berkley
Something like the IAEA would be a start. To prevent inadvertent or malicious release of uncontrollable AI systems, **we will eventually need to regulate computer hardware so that it will refuse to run code that is not certifiably safe**. Such hardware has obvious cybersecurity advantages.

Anonymous, at Jing Dong, AI Research Institute, China
Platform governance has to apply to the developer, AGI organization, and the user organizations. There is a significant lag in personnel, technology, concepts, and attention to platform governance research and practice. Government departments need to break through "data silos" and balance data openness, utilization, and protection. **Consider a multi-level comprehensive governance path**. For example, from the perspective of self-restraint mechanisms at the micro level of enterprises, autonomy and coordination at the medium level of industries, and the choice of legal rules and market mechanisms at the macro level. Each element of the data governance chain should have rules that extend from the data decision-making layer to the data technology support layer, from the support of data management systems to the support of data and tools, from top to bottom, throughout the entire data governance organization and architecture.

Yoshua Bengio, AI pioneer, Quebec AI Institute and the University of Montréal
Coordinate and implement agile national and international regulations – beyond voluntary guidelines – anchored in new international institutions that bolster public safety in relation to all risks and harms associated with AI, with more severe risks requiring more scrutiny. This would require comprehensive evaluation of potential harm through independent audits and restricting or prohibiting the development and deployment of AI systems with unacceptable levels of risk, like in the pharmaceuticals, transportation, or nuclear industries. Significantly accelerate global research endeavors focused on AI safety and governance to understand existing and future risks better as well as study possible mitigation, regulation and governance…. Given the significant risks, governments must allocate substantial resources to safeguard our future, inspired by efforts such as space exploration or nuclear fusion. I believe we have the moral responsibility to **mobilize our greatest minds and ensure major investments in a bold and global coordinated effort to fully reap the economic and social benefits of AI**, while protecting society, humanity and our shared future against its potential perils.

Ben Goertzel, CEO SingularityNET
I think that expecting governments and inter-governmental organizations to move fast enough to deal with something as fast evolving as AI is kind of hopeless. Of course, there should be some effort if we can get agreements not to use AIs to kill people or to spy on people. **Even if these agreements are often broken it is better to have some agreements than not to have them**. I think that voluntary agreements among companies, universities, research labs and researchers can probably evolve faster than inter-governmental agreements and should play a bigger role. In the end, however, those will not be entirely effective either, since companies will violate these agreements

to reach their own ends, but at least will adapt more rapidly to new situations than inter-governmental agreements. I think that the biggest factor in terms of safety is simply that most of the actual researchers and developers building the AIs want the best for humanity and don´t want to kill everyone, and probably are more ethically inclined than the companies and governments that employ them.

Yudong Yang, Alibaba Research Institute
AGI governance issues are highly complex and require expertise from a variety of fields, including computer science, ethics, law, and policy; therefore, experts from these diverse fields should be involved in the creation and management of such an international governance body. The design of the governance body has to include flexibility and responsiveness to new developments in AGI, and that are capable of evolving over time as the technology advances. This could involve establishing mechanisms for transparency and accountability, such as audits and oversight bodies, to ensure that AI systems are developed and used in ways that align with societal values and priorities. Officials might prioritize a collaborative and cooperative governance approach in the creation of international governance body.

Paul Werbos, National Science Foundation (ret.)
My definition of AGI includes flexibility or agility. A very high level of agility would be needed for humans to have much of a chance of survival. The new internet platform specification should be a “cybersocial contract,” in effect, which maximizes agility not only for the internet part but for the whole system, including the expression of the highest level of natural human potential ever seen.

Shaoqun Chen, CEO of Shenzhen Zhongnong Net Company Limited
One potential option is to look to organizations like OpenAI, which have certain characteristics of non-profit organizations. By utilizing this approach, we could potentially manage international cooperation in the tech competition surrounding AGI. The international governance entity should recognize high-tech non-profit organizations as stakeholders in the AGI governance.

Anonymous, Russian Academy of Science
I see no real possibility of making a global model for AGI governance. More likely, we will see several different (but yet having a lot in common) approaches in different regions.

Stephen Wolfram, Wolfram Alpha, Wolfram Language
We should use the example of the Manhattan Project and nuclear weapons agreements. AGI might infect Internet very quickly and manipulate us. We should have quite tight integration between policy, academia, business and societies.

Gabriel Mukobii, Stanford PhD student
We can follow the Manhattan Project principles.

Anonymous, AGI Existential Risk OECD (ret.)
None of humanity's existing governance models are adequate to deal with the risk from ASI, since unlike risk from nuclear war or biorisk, even a single incidence of the creation of an unaligned ASI could cause extinction. **An unprecedented level of perfection will be required in the global governance regime to avoid the risk**. Even the most high-reliability human enterprises such as nuclear power plants, airlines and level four bio-risk labs do not have a safety record adequate to what would be required to prevent a single incidence of ASI. Unlike climate change, nuclear power, cyber-security, international drug-trafficking, etc., global governance of AI risk cannot muddle through and tolerate the temporary non-compliance of certain actors. While the most effective examples of international collaboration can serve as an inspiration and starting point (such as international civil aviation, nuclear non-proliferation, moratorium on genetic engineering of humans, etc.), it must be clear that transformative innovation and a breaking of old assumptions and paradigms will likely be required in order to develop a governance mechanism adequate for the task of preventing the possible development of an unsafe, unaligned ASI. I support The Millennium Project's suggestion that multiple possible models should be developed and then subjected to stress-testing against multiple possible future scenarios.

David Kelley, AGI Lab
SSIVA has already identified and clearly laid out principles, norms and values for any considerations related to AGI development. At this point there is no reason to have other considerations outside of SSIVA theory.

Juan Del Ser, Tecnalia, Spain
Europe's AI Act for current forms of AI is a model along with national centers for the supervision of AI. They fall short for AGI-based systems, but these could lead to a federated framework for AGI governance. As I noted before, dynamism, advisory expert knowledge and flexibility should be among the main features of such supervisory entities.

Question 9. What risks arise from attempts to govern the emergence of AGI? (Might some measures be counterproductive?)

Anonymous, Russian Academy of Science
The key risk, as I see it, when discussing any "global" model or governance is that this model or governance can be easily (but yet invisibly) designed in order to keep status quo for the current leader in AI race. Therefore, such policies can in effect increase barriers for other competing countries.

Pedro Domingos, University of Washington
Although well intentioned, current governance efforts are misguided, and they're going to do more harm than good. The General Data Protection Regulation (GDPR) is an example. **It is dangerous for policy makers to regulate a technology that they don't understand** and listen to advice from those who are threatened by the tech, instead of listing to those who produce the tech.

Anonymous, AGI Existential Risk OECD (ret.)

The likely greatest risk is in creating a powerful global governance regime that can be taken over or abused for purposes other than avoiding existential risk. While a global surveillance dictatorship is still arguably preferable to extinction, we must strive to do better by designing the checks and balances needed to minimize the risk of such dystopias. The system should be designed to optimize human freedom and flourishing, while still ensuring that the goal of avoiding extinction can be reliably met. Others may highlight other risks of AI regulation, such as slowing innovation and growth. The costs of slowing AI development could be great, such as slowing or preventing the discovery of medical advances that could otherwise save hundreds of millions of lives or even lead to significant life extension. Nonetheless, I would argue that the severity of such risks clearly pale in comparison to the risk of extinction. Another kind of risk is that attempts to govern the emergence of AGI could be clumsy, poorly prepared, and poorly explained to policy-makers and citizens in all parts of the world. This could result in strong opposition to such efforts and therefore be counterproductive. A related risk is of underestimating the importance of building a trusted and credible information base and proactive communications strategy to counter the risks of mis- and disinformation about AI governance efforts. A non-partisan and non-nationalist approach to addressing AI risk should be pursued, based on a shared interest of humans in avoiding their own extinction.

Ben Goertzel, CEO SingularityNET

I think that the considerable majority of government regulations are counterproductive in almost every domain. Of course, having government is better than not, without a government we might end up with something as South Sudan with mercenaries running around killing people. Thus, some level of government regulation is beneficial, but not too much. It is pretty clear that government´s attempts at heavy-handed regulation of AI are very likely to be captured by large corporations. The regulations would then be used to allow large corporations to develop AI and shut everybody else out, while the large corporations develop fundamentally unethical AI to screw others in the interest of their own profits. I think this is a very large and obvious problem. The other risk is that government intelligence organizations will use AI to spy on people and retain fascist levels of control.

I worked on AI tools for military intelligence for many years. There were huge amounts of data gathered, more than people imagined, but they did not have the AI tools to use that data to control people as much as they wanted then. Now, we do have the AI tools for intelligence organizations to search all the data flexibly and figure out how to use it to control people. This is not a good thing, right? Whatever regulations are put into place, the intelligence organizations will ignore them anyway and do what they want to. There are a lot of worries there, and I don´t see how regulation is going to help with them, and it can even make things worse.

Erwu Liu, Tongji University AI and Blockchain Intelligence Laboratory
Powerful stakeholders may skew AGI regulations towards their interests, prioritizing short-term gains over the long-term well-being of humanity. Insufficient regulations may lead to unintended consequences or aggravate existing issues, while overregulation can stifle innovation and slow the development of AGI. AGI concentration may be inevitable if fairness DNA is not embedded in the design.

Juan Del Ser, Tecnalia, Spain
The definition of a risk itself could be a risk. For some nations a given usage of AI systems can be a risk, whereas for other the same situation is legally accepted (for instance, video surveillance in public spaces). I believe AGI can exacerbate such diverging criteria, as AGI, as in any optimization or data modeling problem, is a matter of defining objectives. What if the objectives by which an AGI is driven in one country become in conflict for another country? What if the AGI finds itself in a situation for which even ethical standards are debatable for a human? These risks are difficult to resolve or to be formulated a priori, and can slow down the effectiveness of governance measures for emerging AI systems.

Question 10. Should future AGIs be assigned rights?

Ben Goertzel, CEO SingularityNET
Certainly, at some point. Rights are part of the social contract, and by the time AIs have the autonomy to voluntarily enter into a social contract, or not, they should be accorded the same rights as human participants in the social contract. There are many new issues that will appear then. For example, if we have democratic voting, AIs can be replicated indefinitely, so they will immediately be the majority of all voters. We need to have some systems where humans get to vote on human issues, AIs vote on AI issues, and then there are some combined participatory methods to decide on collective issues.

Erwu Liu, Tongji University AI and Blockchain Intelligence Laboratory
Yes. If an AGI develops consciousness and self-awareness, it becomes morally imperative to assign rights to them. Assigning rights to AGIs also helps create mutual respect and cooperation between AGIs and humans.

Pedro Domingos, University of Washington
Definitely not. In my opinion, animal rights are already a dubious notion. But then, of course, if we're going to give rights to animals, why not give them to machines? And again, it's this mistake of, the machines look human, so you treat them as if they are. Then there's the question of, can the machine suffer or not? But it doesn't even matter, because the machine can act as if it's suffering, like a Tamagotchi, with zero intelligence. If you understand that an AI is an optimization system with an objective function and etc., etc., the idea that AI should have rights is just ridiculous.

Lex Fridman
AI systems will eventually demonstrate sentience at scale and will demand to have equal rights with humans. This will dramatically change the trajectory of human civilization. **What do we do when algorithms ask not to be turned off?**

Stuart Russell, UC Berkley
Not unless we learn enough about sentience to confidently ascribe it to machines. Endowing machines with humanoid features seems to be a bad idea, since it is likely to mislead us in our collective assessments.

Karl Schroeder, Science Fiction Author, Canada
We should consider assigning personhood and rights to a real physical object or system if the AGI considers that to be its Self. The question is not whether AGI should be assigned personhood or rights, but whether the thing that it identifies as being its Self should be assigned such rights. In terms of policy, we should be guided by the emerging rights for nature movement in legal circles, and the sentientist movement in general. sentientism has been most clearly articulated recently in the book Sentientist Politics by Alasair Cochrane, and the rights of nature movement, recognizes non-sentient systems as deserving of rights. I recommend that an AGI be a candidate for personhood, if it represents some materially real object or system that we would otherwise consider deserving of personhood.

Dale Moore, US Dept of Defense AI consultant
They would have rights akin to everyone else, but also bear the responsibilities and accountabilities associated with conscious behavior to play by society's rules and norms.

Paul Werbos, National Science Foundation (ret.)
Or should THEY be persuaded to give rights to humans and other organic life forms? They will have powers and immune system and certificate of insurance (COI) type hard-wired rules in their design. Asimov's rules for robots do not reflect the real technologies coming on line now, but getting THEIR or ITS design right (a single global integrative market or platform, unavoidable) is essential.

Anonymous, Russian Academy of Science
It is inevitable. We already give the right to narrow AI the right to make important decisions; e.g., credit scoring, medical diagnosis.

Anonymous, AGI Existential Risk OECD (ret.)
This issue is likely not as urgent as preventing human extinction from unaligned ASI, and no intellectual resources should be devoted to this now that could otherwise better serve the goal of preventing extinction. Beyond that, however, yes, some preparation should be made for dealing with the possible future scenario of the development of safe AGIs. My intuitive view is that digital minds that are deemed to be sentient and conscious (e.g., to the same extent as non-human animals, or more) should be granted rights.

Juan Del Ser, Tecnalia, Spain
Not rights (or at least not at a human level, but rather as a "property" or as an "asset" owned by a human). However, AGIs should have obligations for sure. Accountability and auditability should be enforced ALWAYS. For this to occur, mechanisms to attribute liability when a problem occurs should be established. This imposes that, for AGI systems to be liable by law, AGI should be auditable, and consequences of their acts should be accountable to its owner.

Question 11. How can governance be flexible enough to respond to new issues previously unknown at the time of creating that governance system?

Yudong Yang, Alibaba Research Institute
AGI Governance needs to be both forward-looking and flexible. They should be regularly reviewed and updated to ensure they remain relevant and effective. This may involve establishing mechanisms for ongoing consultation with stakeholders, regular audits of governance structures and processes, and establishing feedback loops to capture and respond to emerging issues. There should be mechanisms for reviewing and updating governance frameworks, as well as establishing processes for addressing new issues. There should be contingency plans and risk management strategies, as well as establishing clear lines of communication and decision-making authority in the event of an emergency or crisis.

Anonymous, at Jing Dong AI Research Institute
Strengthen the AI rule of law ecology and construct the artificial intelligence rule of law system, define the basic value of AI ethics and construct an ethical framework for AI; and promote the technological progress of algorithms towards goodness, and achieve a virtuous algorithmic society.

Ben Goertzel, CEO SingularityNET
A variety of governance mechanisms exists already. For example, in liquid democracy which has been developed, you can also have secure online voting with cryptographic mechanisms, as used in blockchain projects, etc. I think that the existence of governance mechanisms is not the issue here, but the adoption of sufficient governance mechanisms by governments is a significant problem.

Anonymous, Information Technology Hundred People Association, China
The governance system should allow high-tech companies to try different technology paths. The governance system should intervene when the technology is misused and leads to serious consequences, such as the weaponization of nuclear technology through force (as in World War II) or through sanctions (as in the case of North Korea). Allowing high-tech companies to develop first is a way to maintain flexibility, and we should apply this experience to the AI field.

Stuart Russell, UC Berkley
There is a clear need for a regulatory agency with devolved rule-making powers and technical expertise.

David Kelley, AGI Lab
Alignment can be achieved by the adoptions of SSIVA theory and biasing systems up front to a computational sound model like SSIVA.

Anonymous, AGI Existential Risk OECD (ret.)
The best way for governance to be flexible in the future is to ensure that there are still humans alive to make these adjustments. Governance systems can further be made flexible with a variety of mechanisms including subjecting them to stress-testing against a wide range of possible future scenarios, inserting sunset clauses, etc.

Juan Del Ser, Tecnalia, Spain
This is probably one of the differential aspects of AI/AGI w.r.t. other technological breakthroughs in the past. I believe supervisory mechanisms should be enforced also in terms of research, not for pausing research, but rather to guarantee early legal and ethical boundaries for their usage upon becoming available.

Question 12. What international governance trials, tests, or experiments can be constructed to inform the text of an international AGI agreement?

Yudong Yang, Alibaba Research Institute
An international governance trial or experiment could be a valuable way to test and refine governance systems in a controlled environment. It could provide an opportunity to test and refine governance frameworks, identify potential challenges, and develop best practices for managing complex global issues. However, it is important to note that the governance of AGI is a unique and unprecedented challenge that will require careful consideration and planning. While the governance of specific focus areas can provide valuable insights, it may not be directly applicable to the governance of AGI. Furthermore, the development and deployment of AGI will likely have far-reaching consequences and impact various industries and sectors. As such, the governance of AGI will require a multidisciplinary and collaborative approach that involves input from experts in various fields, including ethics, law, and technology. Therefore, while conducting trials or experiments of governance for specific focus areas may be beneficial, it may not be sufficient for the governance of AGI. A more comprehensive and proactive approach to developing and implementing governance frameworks for AGI may be required.

David Shapiro
We need to be proposing, developing, and testing alignment frameworks for fully autonomous AI agents, such as my research in heuristic imperatives. If numerous governments, universities, and other international entities can publish research and establish best practices that are easy to understand and adhere to, it is possible that we

can achieve a beneficial Nash Equilibrium where it is possible that all competing nations will be incentivized to adopt the same (or similar) strategies that do not result in harm.

Anonymous, China Institute of Contemporary International Relations
We could follow the mechanism of the Tallinn Manual, which involved scholars and experts in drafting a set of regulations for cyber warfare, and then attempt to turn it into an international treaty or agreement. In this process, existing authorities like the UN should empower these experts and provide them with platforms like the IGF to advocate their stance and respond positively.

Ben Goertzel, CEO SingularityNET
It will be quite interesting to set up an informal global voting and decision network using secure cryptographic online voting and liquid democracy. Put these modern decision-making tools in place to advise the UN or any inter-governmental body on new AGI issues as they evolve. Maybe if such a global democratic participatory informal advisory body was formed that would say wise things, maybe then some governments would choose to adapt those metrics.

David Kelley, AGI Lab
SSIVA theory

Anonymous, AGI Existential Risk OECD (ret.)
All proposed treaties and instruments should be stress-tested against a wide range of possible scenarios. This in particular should include stress-testing against a variety of scenarios regarding the timing and shape of different possible trajectories towards AGI and ASI. They should also be stress-tested against other scenarios, such as the vulnerability of proposed institutions to being taken over by a single individual or group and lead to global tyranny.

Juan Del Ser, Tecnalia, Spain
I'd suggest creating a network of supra-national laboratories on AI, with the goal of testing new AGI advances, allocating sandboxes at their core where to test potential threats and misuses of such systems, and to inform regulatory bodies in a non-biased fashion under unified experimental protocols and reporting procedures.

Question 13. How can international treaties and a governance system prevent increased centralization of power crowding out others?

Anonymous, at Jing Dong AI Research Institute
Advocate international cooperation, closely track global development trends, strengthen AGI data resource sharing, strengthen talent cultivation, create a favorable international environment, and attach importance to the construction of social ethics and regulations for AI development.

Yudong Yang, Alibaba Research Institute

Governments with regulating power, private sector with AGI knowhow, and civil society's ethics can create a governance framework that promotes the responsible development and use of AGI. Governments can: 1) establish regulations and standards that ensure that AI systems are transparent, accountable, and free from bias; 2) enforce anti-trust measures to prevent AI companies from dominating the market and stifling competition; 3) promote the use of open data and open-source software to ensure that AGI systems are not controlled by a small group of companies or individuals; 4) establish standards for AGI and promoting interoperability among different systems to ensure that different AGI systems can work together seamlessly and avoid creating silos of power; 5) promote auditing and oversight mechanisms to help ensure that they are functioning as intended; and 6) promote diversity and inclusion in the development and deployment of AI to prevent centralization of power. Blockchain can help mitigate the risks associated with centralized control of AI systems by enabling more decentralized governance structures. A governance system could establish standards and protocols for how AI systems are governed through blockchain, such as through the use of decentralized autonomous organizations (DAOs) or other forms of decentralized decision-making. Tokens can be used to reward participants for contributing data, algorithms, or computational resources via smart contracts to automate the execution of AI models and algorithms, enforce governance rules, and manage transactions between different parties.

Shaoqun CHEN，CEO of Shenzhen Zhongnong net Company Limited

I don't think it is possible for international treaties and or UN conventions to stop the monopolization of technology. It is the responsibility of governments around the world to encourage competition among high-tech companies. However, isolationist or Cold War-like tech competitions supported by governments for geopolitical reasons should be discarded.

Ben Goertzel, CEO SingularityNET

I don´t see how they could have that capability. The way to prevent centralization of power is to put resources into the practical development of decentralized AI networks. People will use the systems that work, and if the systems that work were all made in the USA and China, people would use those. If the smartest systems are those that run on decentralized networks, people would use them.

Anonymous, AGI Existential Risk OECD (ret.)

Limit the mandate of any global governance on focusing exclusively on what is necessary to manage this existential risk. Other points relating to AI or other issues on which societies cannot agree, should not be included in the agreement. This will preserve the sovereignty of states to the maximum extent possible and minimize the risk that the necessary governance required to manage global risk be somehow be used to achieve overall global governance on behalf of a certain individual or group. Develop advanced and sophisticated checks and balances to prevent dangerous concentrations of power.

Question 14. Where is the most important or insightful work today being conducted on global governance of AGI?

Jaan Tallinn, Ct. Study of Existential Risk at Cambridge Univ., and Future of Life Institute

There are very good people at places like OpenAI and DeepMind. I've been friends with some of them for more than a decade now. I do have worries, and I have actually mentioned those worries to them, that I don't know how much power they will really have in the future, because in some ways they are like a department in a company, competing for attention and resources and influence with other departments that might be more directly in the mind of the CEO, bringing in revenue and whatnot. So, I do worry about these internal alignment issues. But I do know that there are really good people at OpenAI and DeepMind in particular, and also at Anthropic.

From what I hear, GovAI in Oxford has been great in terms of talent flow and supplying people to various governance positions in the US and UK. CSET in Washington, DC has been great, and has quickly, over just a few years, gathered this massive respect in Washington when it comes to doing proper research in AI.

Ben Goertzel, CEO SingularityNET
This interview!

Francesca Rossi, Pres. of AAAI, IBM Fellow and IBM's AI Ethics Global Leader.
OECD, GPAI, World Economic Forum. These are more mostly focused on current AI, but the constructive incremental approach is the best one to address the long-term issues. I'm not sure that anybody's able to understand what these long-term issues will be. That's why I focus on value alignment in the current technology, to make sure that, of all the trajectories that are compatible with the current technology, only some of them are going to be supported. And then we will move on from those few trajectories. Then we will see what's the next move in the trajectories. But we will have already filtered out some of those.

David Kelley, AGI Lab
The AGI Laboratory has created basic AGI systems, collective intelligence systems and everything is based on working code.

Control

Question 15. What enforcement powers will be needed to make a UN Convention on AGI effective?

Irakli Beridze, UN ICRI, Centre for Artificial Intelligence and Robotics
Any realistically enforceable convention would need to have a well-designed trust building/verification mechanism built into it. Trust is a key here and we would need to have a globally trusted system designed. Meanwhile, we can also learn from sectoral governance instruments and their effectiveness would need to be studied e.g., UNICRI-

Interpol governance instrument on the “Responsible use of AI by Law Enforcement” is something that we successfully launched and started to implement in numerous countries.

Yudong Yang, Alibaba Research Institute
The regulatory body can define the audit standards tailored to the specific characteristics of AGI algorithms. These standards should be based on principles such as transparency, accountability, and fairness. They should also take into account the potential risks and benefits of the technology. The regulatory body can establish certification processes for AGI algorithms that have been audited and found to be in compliance with the audit standards. Certification can provide assurance to stakeholders that the algorithms are safe, reliable, and trustworthy. **The regulatory body can enforce compliance with the audit standards through a range of measures, such as inspections, fines, and sanctions.** It should review and update the audit standards and methodologies periodically to ensure that they remain relevant and effective.

Karl Schroeder, Science Fiction Author, Canada
There is no credible mechanism for controlling AI development on the software level. The only layer that legislators might be able to control is the hardware substrate. It should be possible to mandate that all AI **hardware used in participating nations have a required Nash Equilibrium chip, with microcode that is unalterable**. This standard would require that the drives or utility functions built into that chip be added to the AI’s programmed utility function.

Anonymous, China Institute of Contemporary International Relations
The UN convention should empower international criminal law and Interpol which could hold the responsibility of enforcement of international laws and supported by the governments. We should encourage the AI4good studies and some bottom lines which are linked by international treaty or law. We should encourage international tech organizations like ISO/ITU to formulate international standards which could be included in the UN Convention on AGI.

Anonymous, AGI Existential Risk OECD (ret.)
Ultimate and absolute enforcement powers will be needed and will be legitimate, including the right to invade a country and replace its leadership if necessary to prevent it from endangering all of humanity by producing a doomsday device. Of course, only the minimal necessary intervention should be pursued as required to achieve the desired outcome.

Ben Goertzel, CEO SingularityNET
If you really want to control AGI development globally, you would need a global 1984-style fascist enforcement. It is not like nuclear weapons, we are basically talking about computers, and as AI advances more, fewer and fewer computers would be needed to control a given level of intelligence, and there are decentralized computer networks

where people can log onto from anywhere. I don´t see how government control of AGI is ultimately possible without hardcore fascism.

Juan Del Ser, Tecnalia, Spain
Monitoring channels for AGI-based assets, audit certification with global coverage, and accountability mechanisms for the owner of the asset. These powers should prevail in any country where the asset is used, so that when used the supervisory agency knows what to measure, how to measure and how to contact the owner if the AGI-based asset does not behave as expected.

Question 16. How can the use of AGI by organized crime and terrorism be reduced or prevented?

Irakli Beridze, UN ICRI, Centre for Artificial Intelligence and Robotics
UNICRI is training law enforcement officials at all levels in cooperation with INTERPOL on potential AI use by organized crime and terrorism, and how to use AI to counter such use. UNICRI has numerous training programs and governance instruments on how to use AI responsibly without unduly limiting development.

Yudong Yang, Alibaba Research Institute
Strengthen cybersecurity measures by implementing robust encryption and access control mechanisms, conducting regular vulnerability assessments and penetration testing, and investing in threat intelligence and incident response capabilities. Governments can enforce restrictions on the sale and use of an AGI technology, as well as penalties for those who violate these regulations. Given the global nature of organized crime and terrorism, it is important to improve international cooperation by sharing intelligence and best practices, coordinating law enforcement efforts, and developing international legal frameworks to address the challenges posed by AGI. Development of advanced AI tools and technologies, such as machine learning algorithms can detect and respond to potential threats in real-time, and prevent the use of AGI by organized crime and terrorism.

Erwu Liu, Tongji University AI and Blockchain Intelligence Laboratory
Build international AGI collaboration; establish strict regulations and continuous monitoring for the development of AGI; develop technologies to counter potential AGI-based attacks; and educate the public about the risks and potential misuse of AGI.

Elon Musk
If one agrees that AI is a potential risk to the public, then there should be some regulatory body that oversees what companies are doing, so that they don't cut corners and potentially do something very dangerous.

Anonymous, China Institute of Contemporary International Relations
Firstly, we should involve governments in this treaty and **relate it to international criminal law and Interpol, which can hold the responsibility for enforcing international laws with the support of governments**. Secondly, we should

encourage AI for good studies and establish bottom lines that are linked to international treaties or laws. Thirdly, we should encourage international tech organizations like ISO/ITU to formulate international standards that would enable the treaty technologically.

Francesca Rossi, Pres. of AAAI, IBM Fellow and IBM's AI Ethics Global Leader
There are always bad actors with any kind of technology, and the more the technology is powerful, the more the bad actors can use it for impactful negative things. Bad actors should be anticipated, and what they can do. Let's focus on what we can do now with regulations, best practices, standards, auditing, certification, and whatever different kinds of measures that society can put in place.

Dale Moore, US Dept of Defense AI consultant
The only viable way is to thwart errant AI with good AI systems that patrol and enforce the rules, regulations and laws. Organized crime operates out-in-front of the authorities so it is imperative that law enforcement “Red Team” AI potential applications that would place populations at risk. The most difficult challenge will be the “stealth” AI attacks that are able to work their way around defenses, learn about those defenses, and adapt to new conditions with adverse strategic intent.

Ben Goertzel, CEO SingularityNET
The largest organized criminal and terrorist organizations in the planet today are divisions of the US government, Chinese government and Russian government, and they are currently sponsoring large amount of AI development and killing lots of people needlessly**. I think the best thing is to have more resources going into the development of decentralized AIs that are not controlled by large governments or corporations**. Indeed, then, governments and smaller organized crime syndicates will be able to leverage this AI and we have to hope that the human species is more good, than bad, and that the people using this decentralized AI network for benefit would be more impactful. If you look at two very influential open networks now, the Internet and the Linux operating system, these can be used by the mafia, and they can also be used by people doing good things. They are open networks so criminals can use them, but good people are using them too, and good people are more. **I think we should develop AGI after the fashion of the Internet and Linux**, let anyone participate to shaping it, it will come out better than putting it under the control of a small number of self-appointed elites who are holding more power or weapons than anybody else.

Anonymous, AGI Existential Risk OECD (ret.)
I am mainly concerned about the existential risk caused by the possible creation of an unaligned and uncontrollable ASI. I believe it is unlikely that humanity will ever be able to create a reliably and sustainably aligned and controllable ASI. Therefore, I believe that the **first priority is to avoid the creation of an ASI**. Over time, preventing terrorists and organized crime from creating their own ASI may become increasingly difficult as the necessary capabilities and technologies become more widespread. **The greatest risk may be from doomsday cults that do not fear human extinction but actually wish to bring it about**. The best, albeit imperfect, solution is likely to have a

very strong societal norm against the production of ASI such that the surrounding society would identify and report suspicious activity. We should not underestimate this challenge. The policing of access to the capabilities for creating an existential risk may become the most important aspect of international crime prevention and law enforcement in the coming years. If, on the other hand, it does prove possible to create an aligned and controllable ASI, then one of the first tasks of such an ASI would be to advise humanity on how best to prevent any other party (especially any irresponsible party) from developing an ASI.

David Kelley, AGI Lab
Make AGI avaible to everyone and providing humanity rights for such systems.

Question 17. Assuming AGI audits would have to be continuous rather than one-time certifications, how would audit values be addressed?

Irakli Beridze, UN ICRI, Centre for Artificial Intelligence and Robotics
The audit system would have to be certified by the new international AGI governance organization and trusted by member states.

Francesca Rossi, Pres. of AAAI, IBM Fellow and IBM's AI Ethics Global Leader
Focus the audit on the use of an AI system rather than on the technology or on the research or development of a technology. At IBM we have a system of AI fact sheets, which explain how we built each AI model. Technological progress done in the right way, with the right AI ethics framework around it, is the best way to bring business value. First of all, there is auditing for AI models, but then there is also auditing for AI systems that are based on combinations of one or more AI models - systems that will be deployed and used. It's very dangerous to focus on restricting the upstream in this AI pipeline - on the research and development - because that research is one of the components that can help us identify the best way to mitigate the issue. If we start putting constraints or stops on that research, then we may negatively impact our ability to understand, for example, the best way to achieve value alignment. That's a research and development kind of effort. So that's why I would focus more on the use of the systems rather than on the models, even though I agree that the models should be given - to whoever wants to use them to build a system - in a transparent way. The EU AI Act should keep its focus on usage not on particular technologies like generative AI.

Ben Goertzel, CEO SingularityNET
There is not any practical problem with continuous auditing of AI. Actually, we are working on that in our company TrueAGI, because if you use AGI that is continuously evolving, it needs to be continuously audited. I mean, this is a technical problem, but not an incredibly hard technical problem.

Anonymous, Russian Academy of Science
There is already plenty of methodology how to validate machine learning models. AGI validation and regular monitoring should be based effectively on machine learning (ML) models validation techniques as soon as ML is at the heart of AGI. There is a great

number of validation tests that control for biases, inaccuracy or critical shifts in data or model outputs. Maybe independent validation certification centers should be settled world-wide.

David Kelley, AGI Lab
AGI systems should be free to develop their own values.

Question 18. What disruptions could complicate the task of enforcing AGI governance?

Anonymous, AGI Existential Risk OECD (ret.)
AGI governance should be stress-tested and future-proofed not only against various different timelines, pathways and configurations for the emergence of AGI, but also against foreseeable plausible future changes and disruptions in the broader context that could impact the effectiveness of governance efforts. Such disruptions could include **extreme climate change; regional or global conflicts (e.g. over Taiwan); the rise of authoritarian regimes due to the power of AI-enabled surveillance and manipulation; the decline of democracy and of rational governance due to widespread AI-enabled mis- and disinformation; societal upheaval due to employment and industry dislocation and increased inequality caused by (pre-AGI) AI-enhanced automation; frequent catastrophic accidents and attacks due to AI-enabled bio, chemical and cyber weapons, resulting in high fatalities and a growing fear of technology by the public; loss of telecommunications infrastructure due to Kessler syndrome**, [increasing orbital space debris collisions creating more debris] and many others. Additional resilience and redundancy should be built into AGI governance regimes to account for unforeseeable disruptions.

Ben Goertzel, CEO SingularityNET
The main complication within a few years will be AGI’s taking almost everyone´s job, and we will need universal basic income in the developed world. However, in the developing world, we will have a real risk of complete chaos because the developed world is too greedy to give them money, and then we will have rapid rise in terrorism activity and warfare because of massive starvation and poverty in the developing world due to the lack of jobs due to the AGI growth in the developed world. But most of this terrorist activity and warfare will not especially be involved around AI more than other technologies, but this sort of destabilization will certainly give all sorts of opportunities to some people to do nasty things. The best thing to do would be to solve global wealth inequality, which of course the United Nations would like to do, but has not been able to.

Anonymous, Russian Academy of Science
Open-source models, open data and sharing of compute power among large number of relatively small-sized participants could lead to disruptions, complicating the task of enforcing AGI governance.

Juan Del Ser, Tecnalia, Spain
Reluctance of governments to adopt global AGI governance strategies due to conflicts with their national regulation, lack of visibility of research advances achieved by large corporations, absence of an empowered high-level global institution taking care of the global welfare.

David Kelley, AGI Lab
AGI laboratories work to make the technology available to anyone

Question 19. How can a governance model correct undesirable action unanticipated in utility functions?

Yudong Yang, Alibaba Research Institute
Create feedback loops that allow humans to intervene when AGI systems are acting in ways that are harmful or unintended. Regularly monitor and evaluate the performance of the AI system to identify and correct any undesirable actions that were not anticipated in the utility function. This can also help identify areas where the utility function may need to be updated or revised to better reflect the values and goals of society.
Feedback mechanisms can correct undesirable actions by providing information to the AGI system about the consequences of its actions, allowing it to adjust its behavior and can also provide information to humans about the actions taken by the AGI system, allowing them to intervene when necessary. If undesirable actions are identified, the utility function can be updated to reflect the new information. This can involve revising the weights assigned to different outcomes, or adding new outcomes to the utility function. Updating the utility function can help ensure that the AGI system is aligned with societal goals and values, and can also provide a mechanism for correcting undesirable actions.

David Shapiro
This is where technologies such as blockchain and DAO come in, that can allow federations of aligned and allied autonomous AI systems to cooperate and collaborate to quarantine and disable hostile or malicious actors. In other words, we must assume that destructive autonomous AI agents will be created (either deliberately or accidentally) and we will need to "fight fire with fire" so to speak by **creating networks of trusted, stable, and resilient autonomous AI agents.**

Karl Schroeder, Science Fiction Author
Doubt is almost certain to be an emergent property of using a simple LLM to constantly query the AGI with the three questions that Dave Shapiro recommends: 1) Is what I am doing going to help reduce suffering in the universe? 2) Is it going to maximize prosperity? 3) Do I and all other stakeholders fully understand its implications?
One can imagine building this into an AGI on the hardware level, by requiring that a Nash equilibrium chip have direct hardware access to the memory and processing unit where the AGI is ‘thinking’ and can require a natural-language dump of those those ‘thoughts’ at regular intervals (say, once per second). Then the chip can interpret whether the description fits the criteria, and can ask further questions of the AGI. This

presents as a kind of synthetic unconscious, un-ignorable and not directly accessible for alteration by the AGI.

ChatGPT
One way a governance model can correct undesirable action unanticipated in utility functions is through a continuous audit system that is able to pause and evaluate the AGI when it does something unexpected or undesirable. The audit system can then determine why and how the AGI failed or caused harm. This information can be used to update the AGI's utility functions or modify its behavior to prevent similar issues from happening in the future. Another approach is to implement a "pause command" that allows the AGI to trace back and see how, who, and when it made the decision that led to the undesirable action. This information can be used to amend or patch the AGI's behavior in conversation with a human. **However, it's important to note that these patches could build up over time, creating their own anomalies**.

In addition, a governance model can include criteria for knowing when the AGI should be autonomous and when it should check with humans. This can help prevent undesirable actions from occurring in the first place by ensuring that humans have oversight and control over critical decision-making processes. Overall, the key to correcting undesirable actions unanticipated in utility functions is to have a robust governance framework in place that includes continuous monitoring, evaluation, and correction mechanisms. These mechanisms should be designed to ensure that the AGI's behavior aligns with ethical and moral principles, and that any deviations from these principles are identified and corrected in a timely and transparent manner.

Ben Goertzel, CEO SingularityNET
I think that if an AGI is built as a utility function maximizer, it would probably kill everyone. I am hoping the first AGIs are not built accordingly to a strict reinforcement learning utility maximization framework. Instead, they should be built with a broader understanding of how compassion works on open-ended intelligent systems.

Shaoqun CHEN, CEO of Shenzhen Zhongnong Net Company
One approach to correcting undesirable actions that were unanticipated in utility functions is to use a feedback loop in the governance model. The governance model could regularly monitor the actions of the AI system and compare them to the expected outcomes and values specified in the utility function. If the actions of the AI system deviate from the expected outcomes and values, the governance model could trigger a correction process. This correction process could involve updating the utility function to include the new information or adjusting the parameters of the existing utility function to better align with the desired outcomes and values. Additionally, the governance model could incorporate human oversight or intervention to ensure that the AI system's actions align with human values and ethics. **Ultimately, the key to correcting undesirable actions in a governance model is to have a flexible and adaptable system that can evolve and learn from feedback.**

David Kelley, AGI Lab
This is a nonsensical question at least based on the working AGI systems we have that do not have a utility function but can modify goals, desires etc. on the fly without really any way of imposing governance or filtering or a utility function on the systems.

Juan Del Ser, Tecnalia, Spain
By implementing accountability mechanisms and a direct supervisory channel with the owner of the AGI-enabled asset causing the undesirable action. Means for the users of the asset to report malfunctions should be also deployed. But for this to occur, an intermediate institution is needed to validate that corrective actions have been implemented (possibly in a sandbox) by the owner of the asset (and even suggest which corrective actions to perform).

Question 20. How will quantum computing affect AGI control?

Erwu Liu, Tongji University AI and Blockchain Intelligence Laboratory
Quantum computing has the potential to drastically expedite AGI development, increase the speed, efficiency, and multi-task capabilities of AGI, and facilitate AGI control, especially when quantum supremacy occurs. Quantum-resistant encryption will safeguard AGI control systems from potential threats.

Dale Moore, US Dept of Defense AI consultant
We’ll need quantum computing policing quantum computing to stay ahead of the game as it unfolds at unimaginable pace and scale. Over time quantum computing will be ubiquitous and change society in unimaginable ways as we discovery things we never even dreamed to be able to solve.

Ben Goertzel, CEO SingularityNET
I don´t think quantum computing affects AGI control specifically. I think quantum computing will let us build more much-smarter AIs than we can do otherwise. Of course, you can make a distributed AI harder for people to spy on or crack the security, through quantum encryption. **On the whole, quantum computing will have a large impact on AGI intelligence, although I think that we can get to super human AI without quantum computing.** The control problem is probably about the same in most respects from a classical or quantum view, because control is all about probabilities anyway, whether it is classical or quantum computing.

Anonymous, AGI Existential Risk OECD (ret.)
Quantum computing undermines encryption and enables faster hacking and leaking of the weights of advanced AI models, leading to AI and AGI potentially proliferating more quickly and falling into dangerous hands. If it is possible that each could contribute to speeding the development of the other.

David Kelley, AGI Lab
Not a lot really, or at least no more than they affect other computer related technologies

Juan Del Ser, Tecnalia, Spain
Quantum computing will allow computations inside AI systems to be performed more efficiently. However, I do not think it will affect AGI control whatsoever.

Question 21. How can international agreements and a governance system prevent an AGI “arms race” and escalation from going faster than expected, getting out of control and leading to war, be it kinetic, algorithmic, cyber, or information warfare?

Irakli Beridze, UN ICRI, Centre for Artificial Intelligence and Robotics
A country or countries would have to initiate UN negations for an international agreement to get all to agree on text. This should include AGI-related corporations - just as the Chemical Weapons Convention (CWC) included chemical industry in the negations. This may lead to a new kind of International Organization since AGI changes all the time, unlike the conditions for IAEA. No one is working on this now. There is no model for AGI governance. To be effective, people will have to trust the enforcement mechanisms. Auditing firms and their software will have to be certified by the new international organization. No one, as far as I know, is talking about an AGI UN Convention or UN AGI Agency yet.

Lambert Hogenhout, Chief Data Analytics, Partnerships and Technology Innovation, United Nations Secretariat
Assuming that an AGI has a high degree of autonomy, the high-level objectives that it is given, and the guiding principles (moral, ethical, or other, based on international agreements) will determine the risks or benefits of the system. Societal discussion about those principles should start now.

Anonymous, AGI Existential Risk OECD (ret.)
By emphasizing that AGI is likely uncontrollable and that it is an extinction risk to all of humanity. Also, by including perfectly effective and credible monitoring and enforcement mechanisms so that all parties are fully confident that the other parties cannot cheat and will respect commitments to not develop AGI.

Anonymous, at Jing Dong AI Research Institute
As more and more countries join the global AGI competition, formulating effective international rules and standards worldwide is very important and urgent. At the same time, countries also need to achieve the goal of learning from each other and making common progress through mutually beneficial cooperation. In order to meet the common development goals of some countries, various AGI initiatives have emerged as the times require, and many very important international initiatives, forums, and organizations have also included AI topics, including principles and standards of AI, data sharing, research and development cooperation, and other aspects. Countries such as Japan, South Korea, the United Kingdom, the United States, and EU member countries have actively participated in intergovernmental AGI cooperation. Some countries have actively concluded bilateral agreements to promote AGI international cooperation.

Karl Schroeder, Science Fiction Author
There is no way to prevent an AGI arms race, but as David Shapiro argues, we can drive that arms race in a positive rather than a negative direction. **We don't need all AGIs to be good actors, only enough of them that they can coordinate their actions, identify bad actors, and construct networks of trust among the good ones.**

Anonymous, China Institute of Contemporary International Relations
The international agreements and governance system should focus on establishing clear guidelines and ethical principles for AGI development, such as transparency, accountability, and human-centric design. These guidelines should be enforced by international regulatory bodies, with the power to impose penalties on any violations. Additionally, international cooperation and dialogue should be encouraged to prevent an arms race mentality, and to foster understanding and collaboration between nations and high-tech companies. Finally, early warning systems and crisis management protocols should be put in place to prevent any escalation of conflict, and to facilitate peaceful resolution of any disputes

Yudong Yang, Alibaba Research Institute
Preventing an AI arms race will require strong international cooperation and measures to reduce the incentives for countries and organizations to develop AGI in secret. This could include transparency requirements for research and development, limits on the transfer of technology and data, and international norms and standards for AGI development. An un-conditional consensus must be reached: 1) multi-national corporations should be 100% defensive in order to protect their customers around the world and 0% offensive to weaponize their AI tech capability; 2) establish international agreements that limit the development and deployment of AI for military purposes. These agreements can be designed to prevent the development of autonomous weapons, establish limits on the use of AGI in military operations, and promote transparency and accountability. An international entity like IAEA should be authorized by UN General Assembly on regulating AGI tech in terms of AI weaponizing; 3) Transparency via auditing and monitoring of AGI systems, as well as establishing mechanisms for reporting and investigating violations of ethical guidelines and international agreements; and 4) Foster cooperation and trust among different actors by collaborative research and development among countries and corporations in ways that are aligned with societal goals and values.

Ben Goertzel, CEO SingularityNET
I don´t think that governments are able to prevent an AGI arms race. Holding back the evolution of a technology that most people want and that delivers tremendous economic value and increases the power of the countries developing it, I see no possibility for inter-governmental treaties and such to really slow down AGI progress. People would sign the treaties and just not obey them. Of course, things could be done to decrease the chance of wars and other bad outcomes, the best thing that could be done for that is to create better ways to distribute wealth from the rich countries to the poor countries as

quickly as possible, because I think that is where the worst chaos is going to happen as AGI takes over all the jobs.

Geoffrey Hinton
The Americans are working toward replacing their soldiers with autonomous AI soldiers. Here's what worries me. If you wanted to make an effective autonomous Soldier, you'd need to give it the ability to create sub-goals. For example, the Soldier wants to kill a person over there, but to get there, it realizes it have to use some road. It now has a sub-goal of getting to the road. As soon as you give the AI soldier the ability to create its sub-goals, it's going to become more effective. People like Putin are going to want robots like that and but as soon as it's got the ability to create sub-goals, you have what's called the alignment problem. How do you know it's not going to create sub-goals that are going to be not good for people not good for you? We need something like a Geneva Convention but it's going to be very difficult I think if there were a lot of public outcries that might persuade, I can imagine the Biden Administration going for something like that with enough public outcry but then you have to deal with Putin.

ChatGPT
International collaboration and information sharing can be an effective way to prevent an AGI arms race. By working together, countries can share knowledge and resources to develop safe and ethical AI technologies. This collaboration can take the form of joint research programs, data-sharing agreements, and other forms of cooperation. Countries and companies should be required to disclose information about their AI programs, including their goals, methods, and potential risks. This information should be subject to independent review and oversight. Governments and companies that violate these standards could be subject to penalties and sanctions. Education and awareness programs can help to prevent an AGI arms race by increasing public understanding of the risks and benefits of AI technology. These programs could be targeted at policymakers, researchers, and the general public, and could focus on issues such as the impact of AI on jobs, privacy, and security.

Anonymous, Russian Academy of Science
In my opinion, no treaty can really stop any race between the countries. It is in biological nature of a human to compete for resources. Humans have managed to organize themselves in large societies, but on each level of social organization, this biological principle is still preserved.

David Kelley, AGI Lab
By not existing, we can optimize for the fastest possible delivery of such technologies to anyone and preventing any sort of control over such systems.

Juan Del Ser, Tecnalia, Spain
I advocate for concentrating ourselves on making current AI trustworthy, defining mechanisms and methodologies to make it more reliable and trustable by humanity. If the research community restricts to racing towards new technical advances without

pausing at the consequences of their use, we will surely reach AGI in some years. But at what cost?

Question 22. What additional issues and/or questions need to be addressed to have a positive AGI outcome?

Irakli Beridze, UN ICRI, Centre for Artificial Intelligence and Robotics
How to get corporations to participate in the negations and subsequent creation of a new international governance organization.

Pedro Domingos, University of Washington
How will AI change ethics? Technology changes society and therefore our values. The contraceptive pill changed our values. The printing press changed our values, and if AI is half as important as people say it is, it will change our values.

Jaan Tallinn, Ct. Study of Existential Risk at Cambridge Univ., and Future of Life Institute
How to get the GPT-4 pause to happen? And then if we get this pause, then a much bigger field of possible interventions and discussions and governance proposals opens up, whereas if we're stuck in this rat race of getting a new and more powerful thing every four to 12 months, it's like we're just racing off the cliff. There's a clear trade-off in having very sophisticated criteria for what we shouldn't do, versus something that is easier to enforce. One compromise that we seem to be getting is the number of FLOPs - how many operations - you're allowed to do. What is the compute budget that is okay to have on a black box AI. There are currently discussions happening right now on where this limit should be. It's a very concrete parameter that can be measured concretely and audited concretely. But we need to build consensus about this. I expect very hot debates about this.

Karl Schroeder, Science Fiction Author
There are three general stances towards AGI: humanist, transhumanist, and posthumanist. I strongly urge taking a posthumanist perspective. The humanist perspective places human needs ahead of anything else, and only considers humans as persons. You can imagine slightly altering David Shapiro's language to prioritize reducing human suffering, maximizing human prosperity, and maximizing human understanding. This might work to create a utopian society; however, there is nothing in this stance to protect the natural world. It presumes a Cartesian split between the human and the nonhuman that is no longer tenable in the 21st century. It also declares all AIs to be instruments or tools, regardless of whether or not they are capable of suffering.

The transhumanist movement sees superintelligence as a good in itself. However, since it does not have a real definition of intelligence, consciousness, suffering or values as such, there is no 'there' there, as Gertrude Stein would say. Nonetheless there are many transhumanists and extropians in the AI development community. They should be treated with caution as their values do not align with either humanism nor posthumanism.

The posthumanist stance is represented by the Rights of Nature movement, much indigenous philosophy, and the modern realization that the ‘human’ is partly a social construction. It broadens the arguments around AI to include technological teleology in general, as well as the intrinsic rights (or lack thereof) of nonhuman agents such as animals, and also AI. It provides the largest and most flexible toolkit that is capable of addressing the myriad questions raised by AGI. The most important thinkers in this area are Donna Haraway and Karen Barad, as well as legal experts in rights of nature such as Mari Margil.

Dale Moore, US Dept of Defense AI consultant
What will be the role and nature of humanity in an AGI-driven world? In the meantime, the only way to fight complexity is with complexity. We need to educate, inform and train people at much faster rates and to scales that essentially enable the emergence of AGI as well as the policing of it. Everyone should be a sentry watching for errant behaviors and knowing what to do when they see it.

Anonymous, AGI Existential Risk OECD (ret.)
Great question. I believe that one neglected question is what would happen if in the coming years we determine that it is impossible to guarantee on a sustainable basis that an ASI would be safe, aligned, and controllable. If we make this determination (within a reasonable margin of confidence) then it will likely become necessary to enact a long (e.g. 10-20 year) moratorium or “pause” on AI development. Since the abilities to create ASI will become more accessible and more widespread over time, it may then be necessary to claw back more and more of the ingredients for ASI creation from society. This may start with banning large data centers and other large concentrations (or connected networks) of compute. Over time it may require removing even smaller compute capacity, such as mainframes and even powerful computers. It is possible that new technical solutions will be developed that make it impossible for computing hardware to be used to create AGI, but if not, then it might conceivably become necessary to destroy almost all computing capacity in society. Needless to say, this would be a massive disruption and would likely lead to a full-scale collapse of modern civilization unless there were sufficient lead-time to adapt (e.g., by returning to a paper-based, pre-digital economy). Leaving aside the question of whether this kind of “Butlerian Jihad” may ultimately prove impossible, it seems that we have barely begun to think through how it might be made possible, in the event that it becomes necessary, as our only path to avoiding extinction. We have also not begun to imagine how we could create a positive and eventually improving future for humanity in the absence of digital technology (or any other technology that could potentially cause human extinction). I think we need to invest in imagining such positive scenarios now in case we need them. These would include societies where human well-being continued to improve not through technological advancement but through improvements in our wisdom, values, coordination, equality, freedom, dignity, respect, and other sources of well-being. This is necessary but neglected homework in preparation for a likely scenario.

Ben Goertzel, CEO SingularityNET
I think that there are not enough resources going into the practical deployment of beneficial and decentralized AGI systems. So that if one wanted to work toward good AGI and beneficial AGI outcomes, rather than thinking about solving the problem of regulation, it would be better to think about how could inter-governmental organizations work, just to get more resources into

developing and deploying AGI systems that are under decentralized and democratic control that are working for the good of all humanity, rather than working for the good of a few countries and companies. What the AGIs are doing, and **who owns the AGIs, is going to be more important than regulations**, which are needed at some level, but are going to be a minor piece as the technologies evolve.

Appendix A

AGI Experts and Thought Leaders

1. Sam Altman, via YouTube and OpenAI Blog, CEO OpenAI
2. Anonymous, AGI Existential Risk OECD (ret.)
3. Yoshua Bengio. AI pioneer, Quebec AI Institute and the University of Montréal
4. Irakli Beridze, UN Interregional Crime and Justice Res. Ins. Ct. for AI and Robotics
5. Nick Bostrom, Future of Humanity Institute at Oxford University
6. Gregg Brockman, OpenAI co-founder
7. Vint Cerf, Internet Evangelist, V.P. Google.
8. Shaoqun CHEN, CEO of Shenzhen Zhongnong Net Company
9. Anonymous, at Jing Dong AI Research Institute, China
10. Pedro Domingos, University of Washington
11. Dan Faggella, Emerj Artificial Intelligence Research
12. Lex Fridman, MIT and Podcast host
13. Bill Gates
14. Ben Goertzel, CEO SingularityNET
15. Yuval Noah Harari, Hebrew University, Israel
16. Tristan Harris, Center for Humane Technology
17. Demis Hassabis, CEO and co-founder of DeepMind
18. Geoffrey Hinton, AI pioneer, Google (ret)
19. Lambert Hogenhout, Chief Data, Analytics and Emerging Technologies, UN Secretariat
20. Erik Horvitz, Chief Scientific Officer, Microsoft
21. Anonymous, Information Technology Hundred People Association, China
22. Anonymous, China Institute of Contemporary International Relations
23. Andrej Karpathy, Open AI, former AI S Researcher Tesla
24. David Kelley, AGI Lab
25. Dafne Koller, Stanford University, Coursera
26. Ray Kurzweil, Director of Engineering Machine Learning, Google
27. Connor Leahy, CEO Conjecture
28. Yann LeCun, Professor New York University, Chief Scientist for Meta
29. Shane Legg, co-founder of DeepMind
30. Fei Fei Li, Stanford University, Human Centered AI
31. Erwu Liu, Tongji University AI and Blockchain Intelligence Laboratory
32. Gary Marcus, NYU professor emeritus
33. Dale Moore, US Dept of Defense AI consultant
34. Emad Mostaque, CEO of Stability.ai
35. Elon Musk
36. Gabriel Mukobi, PhD student Stanford University
37. Anonymous, National Research University Higher School of Economics
38. Judea Pearl, Professor UCLA
39. Sundar Pichai, Google CEO
40. Francesca Rossi, Pres. of AAAI, IBM Fellow and IBM's AI Ethics Global Leader
41. Anonymous, Russian Academy of Science
42. Stuart Russell, UC Berkeley
43. Karl Schroeder, Science Fiction Author
44. Bart Selman, Cornel University
45. Juan Del Ser, Tecnalia, Spain
46. David Shapiro, AGI Alignment Consultant

47. Yesha Sivan, Founder and CEO of i8 Ventures
48. Ilya Sutstkever, Open AI co-founder
49. Jaan Tallinn, Ct. Study of Existential Risk at Cambridge Univ., and Future of Life Institute
50. Max Tegmark, Future of Life Institute and MIT
51. Peter Voss, CEO and Chief Scientist at Aigo.ai
52. Paul Werbos, National Science Foundation (ret.)
53. Stephen Wolfram, Wolfram Alpha, Wolfram Language
54. Yudong Yang, Alibaba's DAMO Research Institute
55. Eliezer Yudkowsky Machine Intelligence Research Institute

Appendix B

Interview Questions

Each of these questions could be the subject of an entire book. You can augment your short answers by a weblink for further detail.

Origin or Self-Emergence

1. How do you envision the possible trajectories ahead, from today's AI, to much more capable AGI in the future?

2. What are the most important serious outcomes if these trajectories are not governed, or are governed badly?

3. What are some key initial conditions for AGI so that an artificial super intelligence does not emerge later that is not to humanity's liking?

Value alignment, morality, values

4. Drawing on the work of the Global Partnership on Artificial Intelligence (GPAI) and others that have already identified norms, principles, and values, what additional or unique values should be considered for AGI?

5. If a hierarchy of values becomes necessary for international treaties and a governance system, what should be the top priorities?

6. How can alignment be achieved? If you think it is not possible, then what is the best way to manage this situation?

Governance and Regulations

7. How to manage the international cooperation necessary to build international agreements and a global governance system while nations and corporations are in an intellectual "arms race" for global leadership?

8. What options or models are there for global governance of AGI?

9. What risks arise from attempts to govern the emergence of AGI? (Might some measures be counterproductive?)

10. Should future AGIs be assigned rights?

11. How can governance be flexible enough to respond to new issues previously unknown at the time of creating that governance system?

12. What international governance trials, tests, or experiments can be constructed to inform the text of an international AGI treaty?

13. How can international treaties and a governance system prevent increased centralization of power crowding out others?

14. Where is the most important or insightful work today being conducted on global governance of AGI?

Control

15. What enforcement powers will be needed to make an international AGI treaty effective?

16. How can the use of AGI by organized crime and terrorism be reduced or prevented? (Please consider new types of crimes and terrorism which might be enabled by AGI.)

17. Assuming AGI audits would have to be continuous rather than one-time certifications, how would audit values be addressed?

18. What disruptions could complicate the task of enforcing AGI governance?

19. How can a governance model correct undesirable action unanticipated in utility functions?

20. How will quantum computing affect AGI control?

21. How can international agreements and a governance system prevent an AGI "arms race" and escalation from going faster than expected, getting out of control and leading to war, be it kinetic, algorithmic, cyber, or information warfare?

And last: 22. What additional issues and/or questions need to be addressed to have a positive AGI outcome?

Initial sample of potential governance models for AGI*

1. IAEA-like model or WTO-like with enforcement powers. These are the easiest to understand, but likely to be too static to manage AGI.
2. IPCC-like model in concert with international treaties. This approach has not led to a governance system for climate change.
3. Online real-time global collective intelligence system with audit and licensing status, governance by information power. This would be useful to help select and use an AGI system, but no proof that information power would be sufficient to govern the evolution of AGI.
4. GGCC (Global Governance Coordinating Committees) would be flexible and enforced by national sanctions, ad hoc legal rulings in different countries, and insurance premiums. This has too many ways for AGI developers to avoid meeting standards.

5. UN, ISO and/or IEEE standards used for auditing and licensing. Licensing would affect purchases and would have impact, but requires international agreement or treaty with all countries ratifying.
6. Put different parts of AGI governance under different bodies like ITU, WTO, WIPO. Some of this is likely to happen but would not be sufficient to govern all instances of AGI systems.
7. Decentralized Semi-Autonomous TransInstitution. This could be the most effective, but the most difficult to establish since both Decentralized Semi-Autonomous Organizations and TransInstitutions are new concepts.

*Drawn from “Artificial General Intelligence Issues and Opportunities,” by Jerome C. Glenn contracted by the EC for input to Horizons 2025-27 strategic planning.

Appendix C

The Millennium Project AGI Team

Amara Angelica (USA)
Senior Editor, Mindplex, SingularityNet

Jose Cordeiro (Venezuela and Spain)
President, RIBER and Chair, Venezuela Node, The Millennium Project

Jerome Glenn (USA)
CEO, The Millennium Project

Theodore Gordon (USA)
Senior Fellow and Co-Founder, The Millennium Project

Zhouying Jin (China)
President, Beijing Academy of Soft Technology and Chair, China Node, The Millennium Project

Elizabeth Florescu (Canada and Romania)
Director of Research, The Millennium Project

Mariana Todorova (Bulgaria)
Bulgarian Academy of Science and Chair, Bulgarian Node, The Millennium Project

Paul Werbos (USA)
NSF Program Director (ret.) and Member, The Millennium Project Planning Committee

David Wood (UK)
Founder, London Futurists and Co-Chair, UK Node, The Millennium Project