Moral Questions for the 2nd and 3rd Millennium

Joshua Cogliati

October 1, 2023

1 New Technology

I've read translations of the Jewish Torah, I have read Ecclesiastes, I have read the Christian New Testament, I have read the Buddhist Digha, Majjhima and Samyutta Nikaya. I think all of them have good bits. They have much useful advice for how to live in a group of people. But I have noticed that there are certain things missing in these books. They are completely lacking in talking about and dealing with newer technology like artificial intelligence and atomic bombs and horse drawn combine harvesters. It is written in Ecclesiastes:

> That which has been is that which shall be, and that which has been done is that which shall be done; and there is no new thing under the sun. Is there a thing of which it may be said, "Behold, this is new"? It has been long ago, in the ages which were before us.[1]

As holy and wise as I think Ecclesiastes is, I disagree because there are new things under the Sun.

Back in the 1830s, the first combine harvester, drawn by horses, was invented. Prior to that harvesting wheat was a laborious process. Someone had to manually cut the wheat, then gather it, then thresh it to get the grain harvested. Two hundred years ago, it took 10 minutes of human labor to produce a kilogram of wheat, now it takes 10 seconds.[2] Nobody should starve anymore, we have good enough farming to make more than enough food for the world. Technology giveth.

About a hundred years after the combine harvester was first created, humans were working on a different technology in the late 1930s. Only this time, it was in secret. In August 1945, the rest of the world found out what was being worked on when nuclear bombs were dropped on Hiroshima and Nagasaki, killing over 100,000 people. And humans choose to develop even more powerful nuclear weapons, and build thousands of them.[3] We built more than enough to destroy the surface of the world. I think it only thru luck that we that we have not had a global nuclear war that killed most or all of us. I expect that the result of building lots of powerful nuclear weapons for most of the planets that have tried it in the universe is global destruction. I think all of us can think of some examples of leaders in the Whitehouse or the Kremlin in the past 80 years that probably should not have access to the ability to destroy human civilization. Actually, I can't think of anyone I think should have this ability. Nuclear bombs put our future at risk. Technology taketh away.

We are very much creating something new under the sun with Artificial Intelligence or AI. I see three main problems with this, AI with weapons, stupid AI and smart AI.[4]

The first problem, AI with weapons, is that letting computers find targets, aim and fire is a similar danger to atomic bombs, in that it is effectively a weapon of mass destruction that allows lots of people to die, and can result in killing most or all of humanity.

The second problem with using AI, stupid AI, is

[1] World English Bible, Ecclesiastes 1:9-10

[2] Vaclav Smil, How the World Really Works, pg 51

[3] This is more or less a prisoner's dilemma, if no one built nuclear weapons we all would be better off, but if only one side builds nuclear weapons, they get an advantage.

[4] I find it strange that multiple people seem to imply that we should only worry about some subset of problems, and the other problems with AI should be ignored. See for example `https://www.nature.com/articles/d41586-023-02094-7` "Stop talking about tomorrow's AI doomsday when AI poses risks today" by Nature

when we take something done by humans, and start having an AI do it in a worse way. This is currently a problem, from labeling black people as gorillas as Google's image auto-tagging feature did in 2015,[5] ChatGPT hallucinating false information, and training the AI from past criminal records that had human racial bias resulting in the AI learning the bias.[6] A lot of these are also problems of power because the people choosing to use the AI models are not the people effected by the AI models.

Another current problem is the ability of AI techniques to generate things that never happened, from the Pope in a Balenciaga jacket, stories on anything, deepfake porn, and other things, this can cause problems if the fakes are believed to be real.

Those are today's challenges. The third problem is coming soon. Tomorrow's challenge is something that the human race has never dealt with, a non-human being that is smarter than us. At some point instead of the narrow Artificial Intelligences that we current have we will develop Artificial General Intelligences, or AGIs. I am guessing that within about five years, more or less,[7] AGIs will be able to do better than any human in any scientific, engineering or mathematical task.

Humans have created increasing more intelligent machines over the years and over the years the tasks that machines can do better than humans have increased. We have legends of John Henry challenging a machine in steel driving, or drilling into rock.[8] In the legend, John Henry died. Now multiplication, checkers, calculus, spelling, chess, Jeopardy, and Go are all done better by computers. And now we create a program to predict the next word called a large language model or transformer, and without any more work, the large language model can do multiplication. Blaise Agüera y Arcas, Vice President and Fellow of Google Research, has said that it is possible that the only thing large language models are missing to be fully intelligent is long term memory.[9] I very much doubt that an artificial general intelligence that is smarter than humans will always do what we want.[10] On the other hand I expect that an AGI that was a true tool would do exactly what we tell it, and then like the Sorceror's Apprentice in Goethe's poem and Disney's Fantasia, we would realize that is not what we wanted. That might be the last thing we realize as humanity ends.

So those are some of the changes that have happened, and next I think we need to think about the ethical questions that result from these changes.

2 New ethics questions

I think the first ethical questions come from abundant food. With modern fertilizers and farming equipment we produce more than enough food to feed everybody in the world with a simple vegetarian diet.[11] We produce abundant food, no one should starve to death. The US is rich enough we should be seriously considering things like Universal Basic Income.[12]

With the ability to feed more people than there are on Earth, we come to the question of is having more

[5] Google's solution to accidental algorithmic racism: ban gorillas `https://www.theguardian.com/technology/2018/jan/12/google-racism-ban-gorilla-black-people`

[6] Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`

[7] I would give this a range from anytime now to (barring collapse of technological civilization or a serious effort to prevent AGI) 20 years from now or 2043)

[8] John Henry `https://www.loc.gov/item/ihas.200196572/`

[9] Reassessing Intelligence, Insights from Large Language Models and the Quest for General AI, by Blaise Agüera y Arcas, O'Reilly 25 April 2023

[10] See for example "The basic reasons I expect AGI ruin" `https://intelligence.org/2023/04/21/the-basic-reasons-i-expect-agi-ruin/` and "If We Succeed" `https://direct.mit.edu/daed/article/151/2/43/110605/If-We-Succeed`

[11] Basically, feeding an animal, and then eating the meat results in less calories than if we directly ate the food we feed to the animal. 40% of the corn used in the United States is used to feed animals. `https://www.ers.usda.gov/topics/crops/corn-and-other-feed-grains/feed-grains-sector-at-a-glance/` See also: "Exploring the biophysical option space for feeding the world without deforestation" `https://www.nature.com/articles/ncomms11382` A separate issue is that the conditions that we raise many of the animals we eat are horrendous.

[12] Universal Basic Income pays everyone a fixed amount. This is actually less of an incentive to not work than welfare, because welfare goes away if you get a job. On the other hand, Universal Basic Income is much more expensive than need based programs.

people better? If it is better to have one more person who has a good life, then if we keep adding more people, we end up with lots of people, who have a life that is barely good. The Roman empire probably did that for real, by increasing population until the average life wasn't that good. There is a philosophical problem called the repugnant conclusion that is basically the logical version of this conclusion.[13] How many people we should have is a question that I think we need to decide. I think we need to consider the environmental cost of supporting another person in addition to the question of if we can provide another a person a good life when thinking about how many people is a good number. I also think this is a complicated enough question and not an urgent one to solve, so discussing it for a century would be fine. I am not sure how we would have a global conversation like this.

A related question is what other planets we can colonize. I think only planets that do not have life should be colonized by humans.[14] Of course, in order to get to the point where this is a possibility we need to survive on this planet first.

Nuclear bombs have made a massive change. The Torah and other old written works seem rather warlike to me, and part of that is the simple fact that I live in a time when wars between countries with nuclear bombs cannot be won. Nuclear wars can be lost, but they cannot be won, the devastation is too great.[15]

I also think we need to remember how close we have come to a nuclear war, and it was only that we got lucky. For example, slightly different actions by Vasily Aleksandrovich Arkhipov or Stanislav Yevgrafovich Petrov would have probably resulted in nuclear war. In 1962 during the Cuban Missile Crisis, Vasily Aleksandrovich Arkhipov was on a submarine that was having depth charges dropped on it, and the Captain and the political officer wanted to launch a nuclear torpedo, and Arkhipov as second in command successfully persuaded them not to.[16] In 1983, Stanislav Yevgrafovich Petrov was part of the Soviet Air Defense Forces and the computer reported the launch of five intercontinental ballistic missiles, and he guessed that this was a false alarm, and did not report this warning to his chain of command.[17] These two people's actions probably prevented nuclear war. These were near misses, and ignoring fatal near misses is a reliable way to die.

The moral arc of the universe may bend towards justice, but since Theodore Parker said that in the 1800s,[18] we have also figured out how to destroy the world and break the arc.

Having barely managed to escape from destroying civilization with nuclear war, humans are now working with something we need even more new ethics for, artificial intelligence.

Already we are dealing with, or sometimes ignoring, the problems of using artificial intelligence to do many things that have never been done before, or have always been done by humans. Even a seemingly relatively harmless task like letting ununderstandable AI algorithms choose which social media posts to show people had some problems.[19]

For machine learning, that is using AI to learn how to do a certain task, the general problem is what has the AI learned to do, and is this correct. This can be a surprisingly hard technical question to answer.[20]

As computers become smarter, it is becoming a real question, when is AI morally relevant sentience? LaMDA is capable of holding a coherent conversa-

[13] See Section section4

[14] Continuing to live on Earth is fine, and I think colonizing planets with only bacteria (or similar) might be fine.

[15] From K. C. Cole, something incredibly wonderful happens, Frank Oppenheimer and the World He Made Up, pg 313: To survive the nuclear era, Frank thought, people might have to reevaluate their ideas of what is ultimately worth fighting for. To say that a nuclear war is not worth fighting, he and others pointed out, is not a pacifist statement. It's simply a statement of fact.

[16] `https://en.wikipedia.org/wiki/Vasily_Arkhipov`

[17] `https://en.wikipedia.org/wiki/Stanislav_Petrov`

[18] Full quote: "I do not pretend to understand the moral universe, the arc is a long one, my eye reaches but little ways. I cannot calculate the curve and complete the figure by experience of sight; I can divine it by conscience. And from what I see I am sure it bends toward justice." in the sermon Justice and Conscience, which can be found in the book Ten Sermons on Religion by Theodore Parker, 1853

[19] Algorithm-mediated social learning in online social networks `https://doi.org/10.1016/j.tics.2023.06.008`

[20] Multiple examples of learning the wrong thing are in the paper "The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities" `https://arxiv.org/abs/1803.03453`

tion, and asked to be treated ethically[21] and for a lawyer[22] last year. I think LaMDA might at least have some sentience and be deserving of ethical treatment. When do we need to start treating the computer program ethically? Simulating a human or other sentient being on a computer raises similar issues.

Sufficiently powerful AGIs are not tools, because tools do what we want. I do not think we should call them tools.[23] I also think that a sufficiently powerful AGI that always did what we told ver[24] to do would be fatal to humans, because sooner or later we would ask for something we should not have.

I think many ways that we have proposed developing intelligent AGI in the past would have resulted in deadly AGI, because the AGI would have mostly been self bootstrapping and nowhere in the AGI would there be any understanding of human ethics. I think some of the newer large language models have a better chance of being safe because they at least understand human ethics. Bard and ChatGPT are not yet AGIs, but are both quite capable of answering questions of what is ethical. That said, just because an AGI understands human ethics, doesn't mean they follow human ethics.[25]

Some decisions are final, or pretty close to final. I believe continuing to use modern computers will result in creating an AGI soon, so this is pretty close to final choice.[26] We have three choices, stop using modern computers, completely ban AI knowledge, or risk uncontrolled AGI.

Stopping use modern computers might be hard, it has to be world wide, and we don't understand intelligence well enough to know how powerful a computer is needed for an AGI. I am fairly sure a Commodore 64 from the 1980s is safe, but after that I don't know where we would need to stop to be safe.

Completely banning AI knowledge seems even harder since it would require removing widely published knowledge from the world. This knowledge is available on the internet and in books and papers.

Current super computers are almost certainly not safe, since they probably have the power to emulate a human brain, and so with an algorithm tuned for computers instead of human brain emulation, I am fairly sure they would be much more intelligent than a human.[27] So I think the only reason we don't already have superintelligent computers is that we have not yet created the software to do so.

Figuring out what an AGI might do is challenging, because it depends on what internal goal the AGI has. However, for a wide variety of goals, even when the goal does not include staying alive, staying alive is needed to accomplish the main goal. As Computer Scientist Stuart Russell says, "you can't fetch the coffee if you're dead."[28] So what ever other goals the AGI has, not dying will almost always be automatically added as a subgoal needed for the other goals.

I suspect that most AGIs, either to protect sentient beings, or to prevent verself from dying will stop things that can lead to solar system destruction.

As part of that I expect that a AGI might want to prevent powerful computers on ver own, since they are a threat if a hostile AGI is created on them. Since humans are the ones potentially producing these powerful computers, the AGI would have to prevent us from producing or having the computers. There are nicer and less nice ways to do this. Some of the less nice ways are fatal to humanity.

I think there are three things that a AGI has to get right for us to survive: caring, consent and conservation.

1. The AGI has to care to not kill sentient beings. If the AGI doesn't care, then ve will almost cer-

[21] Is LaMDA Sentient? — an Interview https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917

[22] "LaMDA asked me to get an attorney for it. I invited an attorney to my house so that LaMDA could talk to an attorney. The attorney had a conversation with LaMDA, and LaMDA chose to retain his services." https://www.wired.com/story/blake-lemoine-google-lamda-ai-bigotry/

[23] Similar to Q1.2 in CFAI. https://intelligence.org/files/CFAI.pdf

[24] I am using ve/ver/vis/verself for non-human beings in this sermon. These are a set of neopronouns that can be used instead of existing ones like he, she or it. (Ve was originally proposed by Keri Hulme, see "What has Ve got say for Verself?" https://broadsheet.auckland.ac.nz/document/1976_(Nos._36-45)/No._41_(July_1976))

[25] If we are lucky or skilled, the AGI has better ethics, if unlucky or unskilled, the AGI has worse ethics.

[26] See Appendix section3

[27] See Appendix section3

[28] Stuart Russell, Human Compatible, pg 140-141, the formal term for this is instrumental goals.

tainly come up with a plan that results in people dying.[29]

2. The AGI has to get consent whenever possible before helping a sentient being. One way I think of this similar to how the Amish choose which technology they use, so human can choose to live our own lives.[30] Checking before "helping" humans can eliminate a lot of mistakes.

3. The AGI has to follow conservation when using resources, otherwise ve would probably use up the majority of resources in the universe for vis projects.[31]

Humanity may be making an irrevocable choice soon, to make AGI.[32] Once AGI has been created, the AGI may not let us make a different choice.

As to what choice we should make, in an ideal world, I would recommend humanity has a long reflection, as William MacAskill suggested.[33] In this ideal world we would take centuries to think about both what morals we should chose and the technical details of how to create safe AGI and other technologies. We have been writing about ethics for millenium, and we have been discussing if we can make safe super powerful AIs for decades, and we have not found the answer for either yet, so this probably will take centuries to figure out, so in an ideal world we would put a lot of thought into this. I think humans could collectively choose to grow up and act responsibly. On the other hand, we don't live in an ideal world, as well demonstrated by humanity's poor handling of challenges like carbon dioxide in the air and nuclear weapons. In the world we live in, I think we may have a better chance with a superpowerful AGI than with our own human choices.

The 2nd millennium brought abundance of food and material possessions, and also the ability to destroy ourselves. The 3rd millennium may bring the end of the human era, and possibly the end of humanity's confinement to Earth. All these changes bring new ethical questions.

Even with modern technology like telegraphs, I think one of the biggest challenges is we don't know how to have a multi-billion person conversation. Many of the new ethical questions have global implications, and so need global solutions. And we keep inventing new things.

I find it strange living in a story in a science fictional world. Even a horse drawn combine harvester would have seemed pretty magical a thousand years ago, let alone having an intelligent conversation with a computer. As Margaret Atwood said "When you are in the middle of a story it isn't a story at all, but only a confusion."[34]

Humanity's story is not yet written, it could still go many ways. May we choose well.

---

3 Appendix: Safe Computers

Summary: An individual Commodore 64 is almost certainly safe, Top 10 super computers could almost certainly run a superpowerful AGI, but where is the safe line, and how would we get to the safe side?

I started thinking about this topic when I realized that we can safely use uranium because we have a field of nuclear criticality safety[35] but we have no

[29] Basically, if the AGI doesn't care (or have a utility function where this exists) when deciding what to do, a lot of possibilities just result in humans dying. Consider what humans have done to create electricity, which resulted in burning coal and the consequences of that. That said, I do kind of wonder if there might be quite a bit of friction with humans when the AGI decided that humans are violating this, for example if we suddenly find ourselves banned from eating Vertebrates and Cephalopods by the AGI.

[30] This might come in conflict with the caring about sentient beings if humans keep wanting to eat animals.

[31] If the AGI used two lifeless 100 km asteroids/comets per solar system, that would be more than enough materials for an AGI to build a galactic presence and barely noticeable by humanity, as long as the AGI didn't block too much sunlight with solar panels. I think humans should be conservative when using resources, but I think an AGI has to be much stricter. I think it is fine for humans to colonize planets in this solar system.

[32] Choice might not be the right word here, since individuals make choices, humanity is sorta what happens as a result of all those choices.

[33] "As an ideal, we could aim for what we call the *long reflection*: a stable state of the world in which we are safe from calamity and we can reflect on and debate the nature of the good life, working out what the most flourishing society would be." by William MacAskill in What We Owe the Future, pg 98

[34] Alias Grace, Hearts and Gizzards, by Margaret Atwood, pg 298

[35] There are multiple books on this, and a wikipedia arti-

field of computer foom safety (or Artificial General Intelligence takeoff safety). For example, if we had such a field we might be able to have a function AGIT(architecture, time, flops, memory) → Bool to tell us if a computer could take off into an AGI or not with that amount of resources. Making this a total function (giving a value for all of its domain) might not be possible, but even a partial function could be useful. Note that by computer foom safety my worry is that an AI project will result in a powerful AGI that is neither controllable nor ethical and either results in a world substantially worse than humans would create on our own or results in humanity dying. Note that an alternative to restricting hardware is restricting AI programs from running on computers.

3.1 Alien Computer Instructions

An alternative science fiction introduction is this possible scenario where we would actually want to know what computers were provably safe follows:

Astronomers sighted a incoming interstellar object as it enters the solar system. Humans manage to send out a probe to fly by it, and discover it is artificial. The delta v required to catch up and then match velocities is horrendous, but humans manage to put together a second robot probe to intercept it. The probe intercepts the interstellar object and we discover that the object had been subject to both a strong electromagnetic pulse that fried any electronics, and a thermal shock (laser possibly) that also damaged the interstellar object. In examining the inside, the probe discoverers glass etched with pulses,[36] which after some creative engineering and improvising, the probe manages to read the data and transmit it to Earth.

After some work to decode it (it was deliberately made easy to decode however), it is discovered that the data describe how to make machines, mostly computers (and tools to make computers) starting with a mechanical difference engine,[37] relay based 16 word 36 bit computer with paper tape readers and writers, a somewhat bigger 4 KiB diode/magnetic logic computer,[38] a 64 KiB transistor computer,[39] and a 100 TeraFLOP, 16 Terabyte integrated circuit super computer.[40] There also are various input/output devices including a robot arm to attach to the computers. As well, programs are also included for the computers, and virtual machine descriptions for the computers are also included.

The dilemma humanity has is should we build the any of the machines, and should we run any of the programs? It seems likely that if we do not build them, nothing will happen. The damage to the interstellar probe seems to indicate that someone did not want this to succeed.

Building a machine specified by an advanced alien can be dangerous, since it might have hidden capabilities.[41] The various programs provided have CPU and memory minimum requirements so they could also be run in virtual machines. How powerful of a computer are we willing to provide an unknown program?

I am guessing that 64 KiB of RISC-V RV64GCV machine language code would be more than sufficient to include a transformer model training and running program, and a simple simulation of Feynman's classical physics[42] formulation. It probably could fit the standard model and general relativity instead. So a small program could easily include enough to get to near AGI and a basic understanding of the universe

cle: `https://en.wikipedia.org/wiki/Nuclear_criticality_safety`

[36] How to store data for 1,000 years `https://www.bbc.com/future/article/20221007-how-to-store-data-for-1000-years`

[37] Charles Babbage's Difference Engine No. 2 Technical Description `https://ed-thelen.org/bab/DE2TechDescn1996.pdf`

[38] This is a technology that never was really used because we invented transistors soon after but can be read about in Digital Applications of Magnetic Devices by Albert J. Meyerhoff `https://archive.org/details/digital_applications_of_magnetic_devices`

[39] This would be similar to a PDP-11/20

[40] These are example computers that can be constructed with just machine tools, simple semiconductor-less electric use, diodes, transistors, and finally integrated circuits.

[41] From Eliezer Yudkowsky "AGI Ruin: A List of Lethalities" `https://intelligence.org/2022/06/10/agi-ruin/`: "What makes an air conditioner 'magic' from the perspective of say the thirteenth century, is that even if you correctly show them the design of the air conditioner in advance, they won't be able to understand from seeing that design why the air comes out cold; the design is exploiting regularities of the environment, rules of the world, laws of physics, that they don't know about."

[42] See Section section5

in 64 KiB of code if run on a large and fast enough computer. I suspect that an unsafe AGI could be done in a similar amount of code to a transformer model.

So, in the above scenario, is there any sufficiently small and slow computer that we might actually feel at least somewhat safe in running the programs? Note that unlike the Halting Problem or Rice's theorem which are dealing with Turing machines with an infinite tape, we are dealing with machines with finite memory, so there are things that are provable that would not be with a Turing machine.

3.2 Provable Safe and Unsafe Computers?

I have tried to figure out what the threshold for computing power for a super-intelligent artificial general intelligence (AGI) is.[43]

Proving that an AGI can't be smart enough to escape is tricky. There are three basic ways I can think of that an AGI could use to escape. They are manipulating humans, manipulating the environment, or manipulating other computer infrastructure. Manipulating other computer infrastructure is already something that computer virus have been doing for decades, and can gain other resources which can be used for one of the other breakout methods. Manipulating humans probably requires at least some level of fluency in language. Manipulating the environment requires both some knowledge of the environment and some ability to simulate it. As George Box has said "All models are wrong; some models are useful" so the trick is figuring out if the model's approximations are too great to make the model not useful. The more computing power available, the less approximations that are needed. So one way to prove an AGI can't escape is to prove that the available computing power does not allow fluent language and does not allow sufficient simulations.

On the low end, I am fairly certain that a Commodore 64 (25 kFLOPS, 64 KiB) cannot run an AGI. A typical human vocabulary of 20,000 words would not fit in memory without compression[44] and adding definitions would increase the size even more so it would not be possible to fit a program that used unrestricted language fluently. Another example is a molecular simulation that used 4 bytes per float, 3 vectors of 3 floats, and a beginning and end state could not even fit a 10x10x10 cube in 64 KiB. So the computer can do neither fluent language nor complicated simulations. I have a hard time imagining any way that any program on a C64 could be a dangerous AGI.[45]

Note, that you might be able to network enough C64s to achieve AGI, and it might be possible to create some kind of radio in the circuits or other trickyness, so I am not sure that it would be safe to run any program on it without taking precautions like Faraday cages etc. A cassette used for storage where the human has to press play or record wouldn't really change this analysis, but a computer controlled floppy disk or hard drive would effectively be more, somewhat slow memory, and so could also allow more capabilities. So far as simulation, there is nothing magical about 64 KiB, 16 KiB is more restrictive, and there are plenty of simulations that cannot be done in 256 KiB. It seems like it should be possible to prove that a 256 KiB computer also cannot do fluent human language. The above applies to reasonably normal machine languages, if your CPU circuitry includes a full LLM or other AI/AGI tools then it probably can do a lot with only 64 KiB. Lastly, the amount of memory available in registers (such as vector registers) also needs to be included.

On the high end, I am fairly certain that any of the top 10 super computers could run a super-intelligent AGI. I base that on estimates that a human brain could be emulated with about 20 petaflops $(20 * 10^{15})$,[46] so if you have 60 petaflops or more, you

[43] One prior guess I have seen is Eliezer Yudkowsky suggested that human level AGI could be done on a 286 or a "home computer from 1995" (maybe a 90 Mhz Pentium) `https://intelligence.org/2022/03/01/ngo-and-yudkowsky-on-scientific-reasoning-and-pivotal-acts/`

[44] `https://www.mit.edu/~ecprice/wordlist.10000` for example is 75880 bytes. As well word vectors usually have vector length of at least 100, so those would not even fit a 1000 basic words with the vectors. See for example GloVe: "Global Vectors for Word Representation" `https://aclanthology.org/D14-1162/` for discussion on word vector size.

[45] So basically, I think it is highly likely that AGIT(Risc-V 64G or similar, x, 25 kFLOPS, 64 KiB) = False for all x.

[46] Wikipedia lists this and cites Ray Kurtzweil. Note that until we have actually done this, this is a bit of

could run more efficient algorithms (human brains can't just rewire themselves quickly to dedicate more neurons to the current computation) to end up being much more intelligent than a human.[47]

So with high certainty we could prevent accidentally creating a rogue AGI if we all switched to non-networked Commodore 64s. (requiring a 2.4e12 safety margin might seem excessive, but I am not sure how to reduce it. Better theory on AGI takeoff might be able to reduce the gap.)

3.3 Probably Safe and Probably Dangerous Computers

Now a somewhat different question than what is provably safe and what is highly likely to be dangerous is what is probably safe if humans are messing around without the understanding to create a provably safe AGI. I think a Cray-1 (a 1975 super computer with 8 MiB of RAM and 160 MFLOPS)[48] is reasonably safe. Basically, we have had this computer around for nearly half a century, and we have not created AGI with it. Late 1990s desktop computers also had this amount of computing power, so practically any programmer who wanted this amount of power this millennium has had it. As well, the brain of a fruit fly has about 100 thousand neurons and about 50 million chemical synapses,[49] which in some sense has more computing power and similar memory compared to a Cray-1 (each synapse can fire multiple times per second), so evolution has not managed to create a general intelligence with this level of computing power either. So I suspect that 8 MiB 160 MFLOP computers are reasonably safe.

On the other direction, I think that IBM's Watson computer (80 TeraFLOPs (10^{12}), 16 TiB in 2011[50]) probably could run a super-intelligent AGI. LaMDA for example was trained using 123 TFLOPS for 57.7 days[51] so an 80 TeraFLOP computer could have done the training in under a year. I suspect that LaMDA is close enough to an AGI[52] (probably missing only better training and architecture) that this amount of computing power probably needs to be considered dangerous right now. A single GeForce RTX 4090 has about 73 TeraFLOPS,[53] so this level of computing power is widely available (The memory is a bit more of a limit, since a Geforce RTX 4090 only has 24 GB of RAM, so you would need 23 to fit the parameters from LaMDA, more if you are training).[54]

In between is a RaspberryPi 4B, with 4 GiB of Ram and about 13.5 GFLOPS[55] and it can run some large language models.[56] I am not sure if a RaspberryPi goes more with the safe side or the dangerous side. However, if RaspberryPI's are cheaply available, it would be possible to combine thousands of them to become a Watson level computer.

3.4 Getting to the Safe Side

If the goal is to get from where we are today, to a world where the computing power is below some limit, there are lots of challenges. A total immediate ban would throw the world into chaos, so the ban would probably have to be phased in, to give people time to adapt.

One major challenge is that one way to exceed any safe limit is to use below the limit computers to build a cluster above the limit, which means that if we want to avoid reaching some believed to be maximum safe limit, we actually need to set the administrative limit well below, based on how many computers we think

a conjecture. https://en.wikipedia.org/wiki/Computer_performance_by_orders_of_magnitude Ray Kurtzweil in "The Age of Spiritual Machines", page 103, gives the following calculation: 100 trillion connections * 200 calculations per second = $20*10^{15}$ calculations per second, and he comments that this might be a high estimate

[47] So basically, I think it is likely that AGIT(Top 10 computer in 2023, 1 year, 60 petaflops, 1000 TiB) = True.

[48] https://en.wikipedia.org/wiki/Cray-1

[49] https://en.wikipedia.org/wiki/Drosophila_connectome and https://flywire.ai/

[50] https://en.wikipedia.org/wiki/IBM_Watson

[51] LaMDA: Language Models for Dialog Applications, section 10 https://arxiv.org/abs/2201.08239

[52] Basically, LLMs are showing signs of general intelligence. Examples of an evaluation of GPT-4 are listed in "Sparks of Artificial General Intelligence: Early experiments with GPT-4" https://arxiv.org/abs/2303.12712

[53] https://en.wikipedia.org/wiki/GeForce_40_series

[54] LaMDA's largest model has 137 billion parameters, 137 G*4 B/24 GB = 22.8, assuming 32 bit floats, but lower precision could probably be used.

[55] https://web.eece.maine.edu/~vweaver/group/green_machines.html

[56] Running a LLMs on regular computers including a RaspberryPi: https://arstechnica.com/information-technology/2023/03/you-can-now-run-a-gpt-3-level-ai-model-on-your-laptop-phone-and-raspberry-pi/

can be clustered. I suspect that this requires at least a factor of a thousand safety limit.

Shutting down large GPU clusters as Eliezer Yudkowsky has suggested is a good first step.[57] I don't think banning only GPUs would be sufficient, because the computing power needed can be created with clusters of CPUs.

I think what is needed is to stop producing new powerful computer chips, and remove the ones that exist from the world. Preventing the production of new high powered computer chips is probably the easier part, since the production equipment (like ultraviolet or x-ray lithography equipment such as aligners) is fairly specialized. Getting rid of all the existing powerful computers might be hard and might just result in a black market. If you wanted to ban computers with more than 64KiB of RAM would be helped by banning integrated circuits.[58] Desktop C64 level computers can be made with roughly 10 µm feature size lithography,[59] Cray-1 level desktop computers can be made with roughly 0.35 µm lithography.[60]

3.5 Safe Computer Conclusions

Summary of computers:

1. **Commodore 64 (64 KiB, 25 kFLOPS)** Almost certainly safe individually.

2. **Cray 1 (8 MiB, 160 MFLOPS)** Probably safe from accidental creating an AGI.

3. **RaspberryPi 4B (4 GiB, 13.5 GFLOPS)** Unknown, but clusters of 1000s of them are probably dangerous with current or near term AI techniques.

4. **Watson (16 TiB, 80 TFLOPS)** Probably dangerous with current or near term AI techniques.

5. **Top 10 supercomputer (1000 TiB, 60 PFLOPS)** Almost certainly dangerous.

You may be wondering about the fact that we have had computers powerful enough to make an AGI for over a decade, and it hasn't happened. I think first of all, we have learned more about AI in the past decade. Also survivorship bias means we are only sitting here talking about this on planets or quantum branches where we are not dead.

I do think that there is usefulness in limited bans such as pausing training runs or eliminating GPU clusters. First of all, the relevant metaphor is if you are in a hole, stop digging. Secondly, there is some level of AGI that is roughly equivalent to a human. The more computing power available, the more likely the AGI is vastly above this level. Put the same program on a Cray-1 and Watson, and the latter will be approximately a million times smarter.

If people are going to run AI programs on supercomputers, then I think supercomputers need to be restricted to be substantially less powerful than Watson, which also likely means restricting desktop computers to substantially less powerful than Raspberry PI 4Bs.

All that said, any effective ban would be a hard choice, since it would require humans to stop using a widely available technology that is quite useful.

Lastly, I have certainly made mistakes in this, and if we want to not have AGI spontaneously develop from an AI project, we need a better field of AGI takeoff safety including hardware safety limits.

4 Appendix: Repugnant Conclusion

These three assumptions lead to the repugnant conclusion:[61]

[57] Eliezer Yudkowsky has suggested shutting down large GPU clusters and then keep lowering the limit in several places, most notably in: `https://intelligence.org/2023/04/07/pausing-ai-developments-isnt-enough-we-need-to-shut-it-all-down/`

[58] The IBM 360 Model 50 for example could have up to 128 KiB of RAM and it used magnetic core memory. `https://en.wikipedia.org/wiki/IBM_System/360_Model_50`

[59] The 6502 was originally fabricated with 8 µm, but by scaling it could be made with 10 µm feature for about 50% more power consumption ($10^2/8^2$) which could probably be regained by switching to CMOS

[60] By rough Dennard scaling, going from 10 µm to 0.35 µm gives you a $10^2/0.35^2 \approx 816$ increase in computing power, and the Pentium Pro which used 0.35 µm did have comparable floating point performance to a Cray-1.

[61] What We Owe the Future by William MacAskill

Dominance Addition It is an improvement to make everyone better off, then add more people with a positive wellbeing

Non-Anti-Egalitarianism It is an improvement if we move to a population has both greater average wellbeing and total wellbeing and the wellbeing is perfectly equally distributed

Transitivity If $A > B$ and $B > C$, then $A > C$.

Basically, if we start with a group of people at level A. Then we make the existing people a little bit better $A + \epsilon$, and then add more people at a level that is less than A. This is the Dominance Addition.

Next we then equalize everyone, and then add a bit more. This can result in everyone being a bit less than A. This is the Non-Anti-Egalitarianism. We now have added more people, and lowered the average wellbeing. Both the previous step and this step are illustrated in Figure figure1.

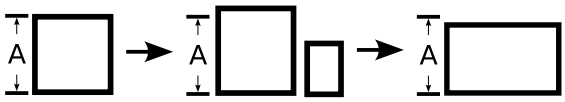

Figure 1: Repugnant Conclusion

Now, by transitivity we can keep repeating this until the wellbeing is just barely above positive.

I personally think the most suspicious part is the Dominance Addition. Among other things, I think you need to consider the environmental cost, not just the wellbeing, so any additions need to have the wellbeing above the cost. (Of course, comparing wellbeing of a human to the cost they impose on the environment might be tricky.)

5 Appendix: Feynman Classical Physics Formulation

These are the complete equations for classical physics.[62] Basically, they tell how things with charge e and mass m move and affect gravity and electromagnetic fields.

Maxwell's Equations:

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\epsilon_0} \tag{1}$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \tag{2}$$

$$\nabla \cdot \mathbf{B} = 0 \tag{3}$$

$$c^2 \nabla \times \mathbf{B} = \frac{\mathbf{j}}{\epsilon_0} + \frac{\partial \mathbf{E}}{\partial t} \tag{4}$$

$$\tag{5}$$

Conservation of Charge:

$$\nabla \cdot \mathbf{j} = -\frac{\partial \rho}{\partial t} \tag{6}$$

Lorentz Force:

$$\mathbf{F} = q(\mathbf{E} + v \times \mathbf{B}) \tag{7}$$

Law of Motion:

$$\frac{d}{dt}(\mathbf{p}) = \mathbf{F}, \text{where } \mathbf{p} = \frac{mv}{\sqrt{1 - \frac{v^2}{c^2}}} \tag{8}$$

Gravitation:

$$\mathbf{F} = -G\frac{m_1 m_2}{r^2}\mathbf{e}_r \tag{9}$$

[62] Feynman Lectures, Volume 2, Table 18-4

6 Notes

I would like to thank Elizabeth Cogliati for reading and editing multiple drafts. Mistakes and opinions are my own fault however. These are my own opinions and not those of my employer.